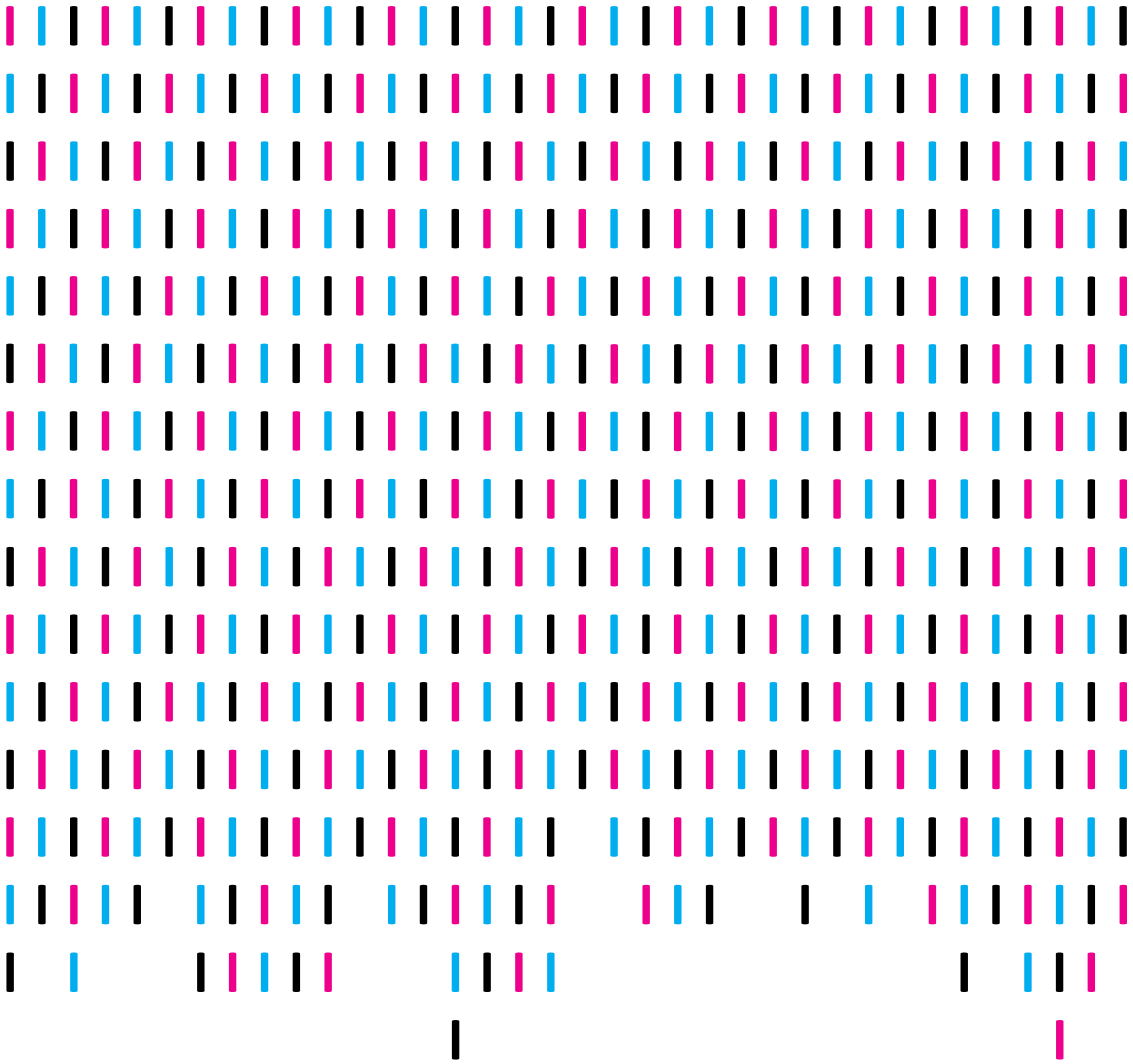

NEPS WORKING PAPERS

Sabine Zinn and Ariane Würbach

A STATISTICAL APPROACH TO ACCOUNT FOR HEAPING PATTERNS: AN APPLICATION TO SELF-REPORTED INCOME DATA

NEPS Working Paper No. 40
Bamberg, April 2014

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Working Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at
https://www.neps-data.de/projektübersicht/publikationen/nepsworkingpapers

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

A Statistical Approach to Account for Heaping Patterns: An Application to Self-Reported Income Data

Sabine Zinn
Leibniz Institute for Educational Trajectories, Bamberg, Germany
Max Planck Institute for Demographic Research, Rostock, Germany

Ariane Würbach
Leibniz Institute for Educational Trajectories, Bamberg, Germany
University of Bamberg, Germany

E-mail address of lead author:
sabine.zinn@lifbi.de

Bibliographic data:

Zinn, S. and Würbach, A. (2014). *A statistical approach to account for heaping patterns: An application to self-reported income data* (NEPS Working Paper No. 40). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

A Statistical Approach to Account for Heaping Patterns: An Application to Self-Reported Income Data

Abstract

Self-reported income information particularly suffers from misreporting due to the sensitivity of the issue and the error-proneness of the memory. This leads to an intentional coarsening of the data, which is called heaping or rounding. If it does not occur completely at random–which is usually not the case–heaping and rounding has detrimental effects on the results of statistical analysis. For instance, it has an effect on empirical statistics (e.g., percentiles) as well as on inferences from multivariate analyses. Conventional statistical methods do not consider this kind of reporting bias, and thus might produce invalid inference. In this paper, we describe a novel statistical modeling approach that allows us to deal with self-reported heaped income data in an adequate way. We suggest modeling heaping mechanisms and the true underlying model in combination. This way we are able to simultaneously estimate the parameters of the true distribution and to determine the heaping pattern present in the data. To describe the true net income distribution, we use the 3-parametric Dagum distribution. Heaping points are identified from the data by applying a heuristic procedure comparing a hypothetical income distribution and the empirical one. To determine heaping behavior, we employ two distinct models: On the one hand, we assume piecewise constant heaping probabilities, and on the other hand, heaping probabilities are considered to increase steadily with proximity to a heaping point. We validate our novel approach by a range of simulation studies. To illustrate the capacity of the novel approach, we conduct a case study using income data from the adult cohort of the German National Educational Panel Study.

Keywords

heaping, self-reported income data, piecewise constant heaping probabilities, piecewise bell-shaped heaping probabilities, Dagum distribution, German National Educational Panel Study

1 Introduction

This paper introduces a novel statistical modeling approach to adequately deal with general heaping patterns prevalent in self-reported numerical data such as income data. We suggest modeling heaping mechanisms and the true–but unobserved–distribution of the numerical variable of interest in combination. This way we are able to simultaneously estimate the parameters of the true distribution and to determine the heaping pattern present in the data.

As a motivating example, we concentrate on individual net income. In various fields, information on income is essential for analysis. For instance, it is crucial to determine the relative income poverty in a country. Likewise, the amount of income is considered as one of the driving forces behind many decision-making processes such as the decision to have a child. Besides the data available from pension insurance systems, large-scale surveys are the only source of income information. One such survey providing accordant data is, for example, the adult cohort sample of the National Educational Panel Study (NEPS Starting Cohort 6).[1] However, especially income information often suffers from misreporting due to the sensitivity of this issue (Miller & Paley, 1958; Hanisch, 2005). The heightened sensitivity of the corresponding question stimulus induces an intentional coarsening of the data. Retrospective data collection and time restrictions during the interview additionally impair the memory process, and hence, the precision of the

[1] SC6 version D-3.0.0, see Blossfeld et al. (2011) for a general discussion of the study design and Leopold et al. (2011) for a general documentation of the Scientific Use File.

responses. Respondents may then rely on heuristics that lead to varying degrees of great differences between reported and true values. Misreporting data thus also occurs in a variety of other applications, for example, when reporting cigarette counts (Wang et al., 2012), age of death, and weight data (Camarda et al., 2007), ultrasound images in foetal medicine (Wright & Bray, 2003), unemployment duration (Torelli & Trivellato, 1993), and gestational age (Pickering, 1992).

The phenomenon of coarsening data by rounding true values to even multiples of reported units is called heaping. In self-reported income data the strong preference, for instance, for multiples of 100, 500, and 1,000, is striking. Figure 1 illustrates this pattern on the basis of the frequency distribution of the self-reported individual net income in the NEPS Adult Cohort. Heaping values has detrimental effects on the results of statistical analyses when occurring not at random (Heitjan & Rubin, 1991; Gill et al., 1997). For instance, it has an effect on distribution parameters (e.g., mean and variance), empirical statistics (e.g., percentiles), as well as inferences from multivariate analyses (Wang et al., 2012; Hanisch, 2005). Conventional statistical methods do not consider this kind of reporting bias, and thus might produce invalid inference. Research on evidence and quantifying the extent of heaping is abundantly available, and tests are also provided, see for example, Beaman & Grenier (1998) or Roberts & Brewer (2001).

Though substantial research exists on the statistical modeling of heaping, models are mainly defined for specific applications, for example, for dealing with misreported cigarette counts (Wang et al., 2012). The basic idea of recent modeling approaches is to define a model for the latent exact values and a set of heaping rules to work on these values. Roughly speaking, the approaches only differ in the parametrization of models and in the manner of how the rules are set. Depending on the area of application, the variable of interest is specified, for example, to be normally distributed (age of children recorded by month, see Heitjan & Rubin (1991); log-transformed individual nuchal translucency, see Wright & Bray (2003)), to follow a Poisson distribution (yearly death counts, see Camarda et al. (2007); number of cigarettes smoked per day, see Wang et al. (2012)), to have a log-normal distribution (net income per month, see Drechsler & Kiesl (2012)) or to be piecewise exponentially distributed (unemployment spells, see van der Laan & Kuijvenhoven (2011)). The set of rules describing the latent heaping mechanism might be defined either explicitly or implicitly. An explicit definition of heaping rules means to constitute a detached deterministic or stochastic model according to which values are heaped. Such processing results in mixture model approaches, see for example, Heitjan & Rubin (1991); Torelli & Trivellato (1993); Wright & Bray (2003); van der Laan & Kuijvenhoven (2011); Wang et al. (2012); Bar & Lillard (2012). Implicit heaping rules can be found in, for example, Pickering (1992) and Camarda et al. (2007), who both model the heaping mechanism through a composite link function superimposed onto the latent distribution of the variable of interest. For statistical inference, in the different fields of application both Bayesian and frequentist methods have proven their value, see for example, Heitjan & Rubin (1991) and Camarda et al. (2007). Analysis based on heaped data can immensely be improved if heaped data are replaced by imputed ones. Once the true latent distribution of the variable of interest is known, this can easily be conducted. A technique that has proven to be successful in this context is the method of multiple imputation, see for example, Heitjan & Rubin (1991); van der Laan & Kuijvenhoven (2011); Drechsler & Kiesl (2012); Wang et al. (2012). Applying multiple imputation to heaped data means to create several data sets by replacing heaped values several times with values drawn from the true latent distribution. The resulting data sets can then be analyzed separately using standard complete-data methods. A final statistical result is obtained by subsequently combining the outcomes of the distinct analyses applying the combination rules of Rubin (Rubin, 1987).

In order to develop a model that allows us to properly analyze heaped income data, we decided to adapt the procedure of van der Laan & Kuijvenhoven (2011). For our purposes this procedure seems to be the most appropriate one as it does not state one specific distribution for the variable of interest. Thus, it can be extended in a straightforward way to cope with our research. We

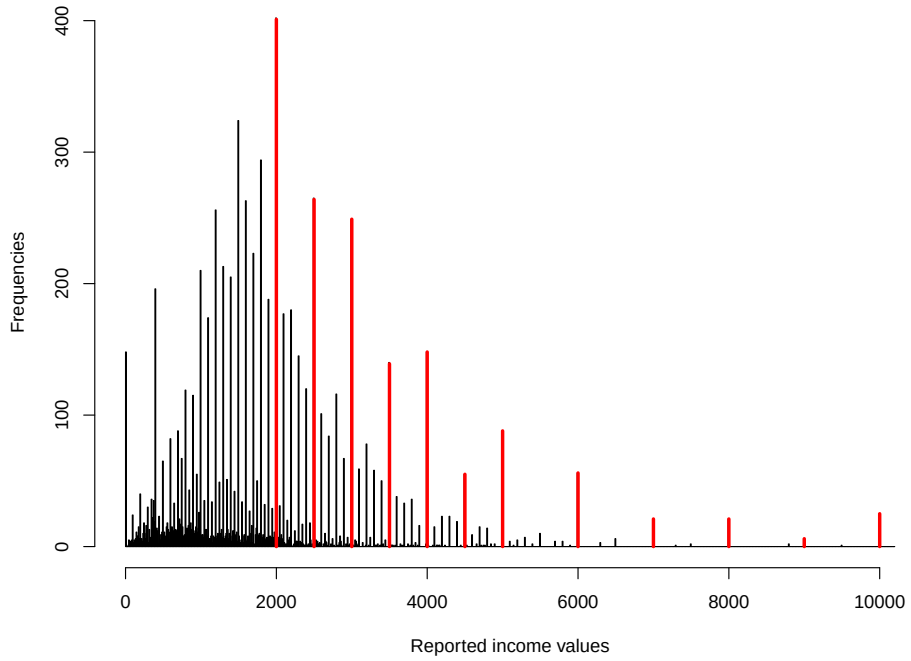

Figure 1. Frequency distribution of self-reported individual net income in the NEPS Adult Cohort, Wave 2009/2010.

designed an approach that permits modeling the heaping mechanism present in the data and the true underlying model in combination. The true underlying model describes the distribution function of the income variable, if the data are reported correctly. The heaping mechanism works on top of this by inducing shifts of true values to heaping points. Formalizing the heaping mechanism demands a specification of a set of heaping points and a function quantifying the probabilities to heap. As true underlying model we choose the 3-parametric Dagum distribution. Bandourian et al. (2002) show that this distribution is particularly well suited to describe the (net) income distribution in various countries. Heaping points are identified from the data by applying a heuristic procedure comparing a hypothetical income distribution and the empirical one. Heaping probabilities are specified using two alternative settings: On the one hand, we assume piecewise constant heaping probabilities, and on the other hand, heaping probabilities are considered to increase steadily with proximity to a heaping point. Simulation studies are used to test the validity of the novel approach. To illustrate its capacity, we apply it to the individual net income information collected in the adult cohort sample of NEPS. The remainder of the paper is as follows: In section 2, we present a formal modeling strategy for heaping. The general method will be described in detail as well as the different procedures for modeling the heaping mechanism. This section is followed by the description of the model estimation in section 3. Section 4 presents the simulation studies conducted, and discusses their results. In section 5 we apply our approach to real data. The paper concludes with remarks and further possible extensions of the general method.

2 Heaping model

A formal model for describing heaping behavior demands that three issues to be addressed: First, the model for the true underlying distribution of the variable of interest–in our case, individual

net income–has to be defined. Then, we need to specify the heaping mechanism operating on the data. This requires that the set of numbers preferred when reporting values, the so-called heaping points, be identified. On the other hand, the probabilities of heaping have to be determined.

2.1 Latent distribution of true values

In order to specify a formal heaping model, we need to introduce some notation. The vector $\boldsymbol{z} = (z_1, \ldots, z_n)$ comprises the values reported for the variable of interest. Because we study net income, it is $z_i \in \mathbb{R}_0^+, i = 1, \ldots, n$. The vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ gives the true values corresponding to $\boldsymbol{z}$, $y_i \in \mathbb{R}_0^+, i = 1, \ldots, n$. These values are not directly observed. The set of heaping points is described by $H = \{h_1, \ldots h_S\}$, $S \in \mathbb{N}$. For reasons of convenience, we assume that all heaping points h_b lie in the value range of the income variable, $b = 1, \ldots, S$. It is implausible to state that all heaping points attract all possible (true) income values to the same extent. Therefore, we define that each heaping point h_b has a certain catchment area from which values can be heaped to h_b, that is, a heaping point cannot pull values from outside its catchment area. We denote these catchment areas by $I_b = [s_b, t_b]$ where s_b and t_b describe the lower and upper bounds of the respective intervals. The function $v_b(y)$ describes the probability to round value y to heaping point h_b. We denote the true underlying probability distribution function of the variable of interest by $f(y)$ and by $F(y)$ the corresponding cumulative distribution function. In fact, for describing the (true) net income distribution we use the 3-parametric Dagum distribution, which was found to be ideally suited for this purpose (Kleiber & Kotz, 2003; Bandourian et al., 2002). The accordant density is

$$f(y \mid \psi) = b^{-aq} aq y^{aq-1} \left[1 + \left(\frac{y}{b}\right)^a\right]^{-q-1}, \tag{1}$$

where $y \in \mathbb{R}^+$, ψ comprises the unknown parameters a,b, and q, with $a, b, q \in \mathbb{R}^+$. As the Dagum distribution is only defined for values greater than zero, without loss of generality, we model zero income by very small values, for example, by 10^{-6}.

2.2 Identification of heaping points

The set of heaping points can be determined by either defining them ex ante or by identifying them from the data at hand. We rely on the second approach by employing a heuristical procedure capable of catching the heaping points from the given data. The basic idea is to define a set of hypothetical heaping points, which are then checked for being–according to the data–real heaping points. For this purpose, we compare the empirical cumulative distribution function (cdf) $\hat{F}(y)$ (estimated from $\boldsymbol{z}$) with the cumulative distribution function of a hypothetical income distribution. This hypothetical income distribution is designed such that its cumulative distribution function $F^h(y)$ roughly resembles the degree of smoothness of the real (unobserved) cumulative income function $F(y)$. The degree of smoothness of $F^h(y)$ can then be used as a prototypical gauge of the one of the underlying true income distribution $F(y)$. Because smoothness is a prerequisite of the Dagum distribution, $F^h(y)$ features the intended shape when being defined as a Dagum distribution, whose parameter vector ψ is estimated using the reported values $\boldsymbol{z}$. We parameterize $F^h(y)$ accordingly and use it to simulate n hypothetical income values $\boldsymbol{w} = (w_1, \ldots, w_n)$. On the basis of $\boldsymbol{w}$, the empirical cumulative distribution function $\hat{F}^s(y)$ is computed. A hypothetical heaping point h_b^0 is worth being considered a potential heaping point only if the value of the accordant increment of $\hat{F}(h_b^0)$ exceeds the median of all increments of $\hat{F}^s(y)$. If this is the case, h_b^0 is identified as de facto heaping point if this value also exceeds the corresponding value of the increment of $\hat{F}^s(h_b^0)$. We have conducted several simulation studies to validate the feasibility of our heuristic. In section 4.1 and section 4.2, two settings are detailed.

Overall, we find that it has been performing well. Figure 2 illustrates the processing of the heuristic using income data simulated as described in section 4.1. It should be noted that the heuristic presented also allows us to identify heaping points that are not considered as common ones, such as heaping points not being multiples of 100, 500, and 1,000.

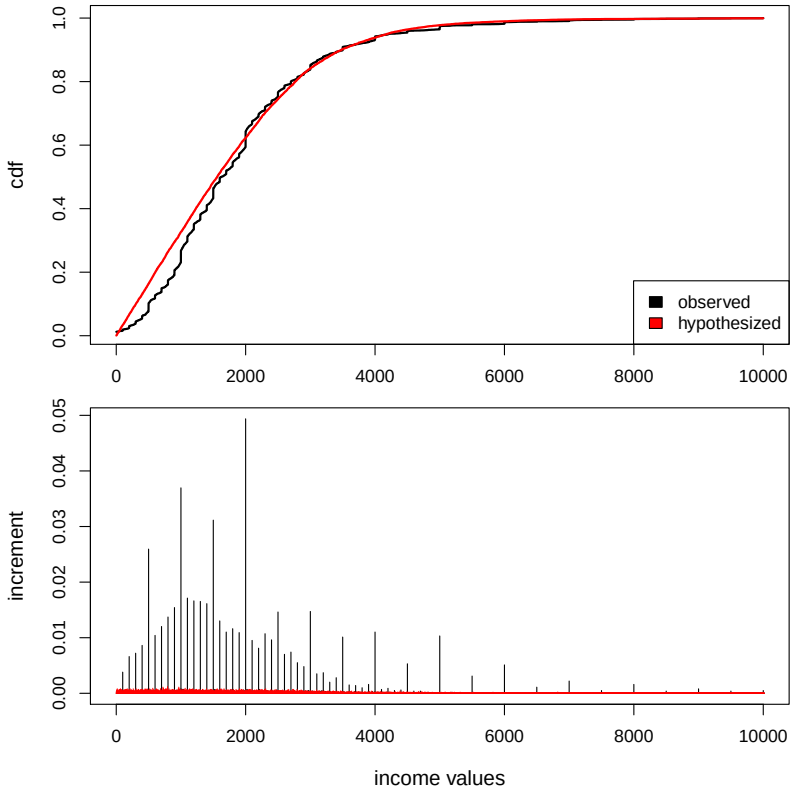

Figure 2. The upper graph shows the empirical cumulative distribution function (cdf) estimated from observed net income (given in red) as opposed to the empirical cdf estimated from values simulated from the respective hypothetical income distribution (given in black). The lower graph gives the increments of the empirical cdf of the observed income values and the corresponding increments of the cdf of the hypothesized income values.

2.3 Heaping Probabilities

A further aspect that we have to address when designing a formal heaping model is the quantification of heaping probabilities. To cope with this issue, we define probability functions $v_b(y)$ capturing an interviewee's propensity to round his true income y to heaping point h_b. Such function might depend on various factors. Certainly, individual characteristics of the interviewed person play a role as well as the conditions of the interview situation. But those aspects are retained for further research. Likewise, the magnitude of true value of the reported item and its proximity to h_b and to other heaping points might affect the propensity to heap. For instance, someone who earns €476 might be more prone to round up this value to €500 than to €1,000. To keep it simple, in this paper, we assume that the propensity to heap depends only on the true level of income and on its proximity to a heaping point. In line with this, we define two distinct patterns of heaping behavior: First, we constitute within all catchment areas I_b a uniform heaping behavior. This results in piecewise constant heaping probabilities of the following form:

$$v_b(y) = \begin{cases} p_b, & \text{if } y \in I_b, \text{ for } y \neq h_b \\ 0, & \textit{otherwise.} \end{cases} \tag{2}$$

Here, p_b denotes the constant heaping probability contributing to heaping point h_b, $b = 1, \ldots, S$. This pattern of heaping behavior especially focuses on the variability of heaping probabilities due to the magnitude of the regarded values. To account for the fact that people's propensity to heap is likely to increase with proximity to a heaping point, alternatively, we define heaping probabilities to steadily increase with the proximity to h_b:

$$v_b(y) = \begin{cases} \eta_b \exp\big\{ -2\xi_b^{-2}(y - h_b)^2 \big\}, & \text{if } y \in I_b, \text{ for } y \neq h_b \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

with $\eta_b \in [0, 1]$ and $\xi_b = 0.5(t_b - s_b)$, $b = 1, \ldots, S$. Figure 3 illustrates both heaping probability functions described. Please note that the probability of heaping a value y located on heaping point h_b to precisely that heaping point h_b is zero (as there is nothing to heap). The given definitions allow for overlapping catchment areas, that is, values can be heaped to different heaping points. For example, value $1,598$ might be heaped to $1,500$, $1,600$, or $2,000$.

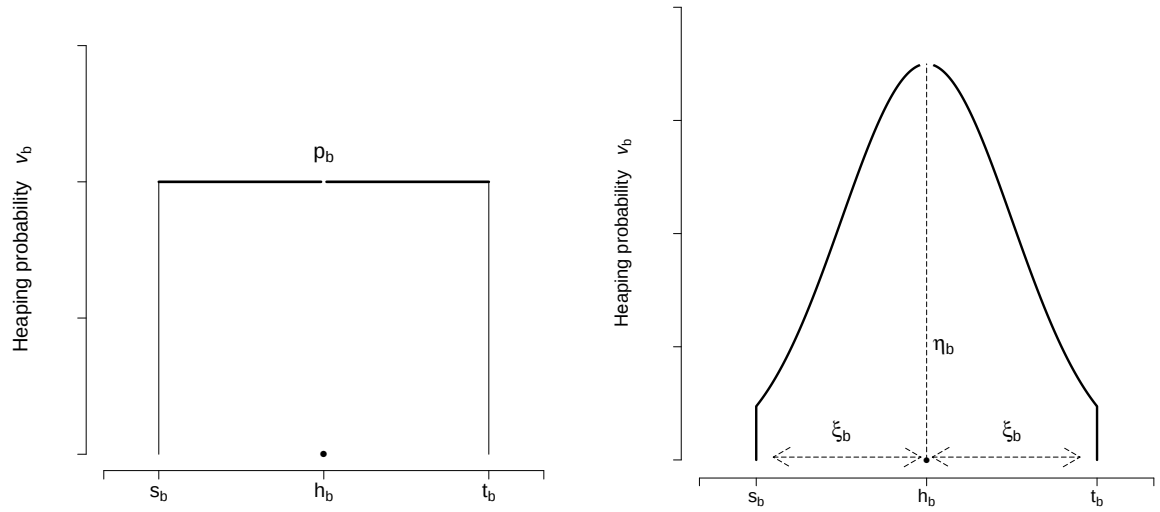

Figure 3. Piecewise constant heaping pattern (left graph) and heaping pattern based on probabilities that steadily increase with proximity to h_b (right graph).

Both heaping probability functions thus defined resemble a multinomial distribution with $S + 1$ categories, that is, with probability $v(y) = \sum_{b=1}^{S} v_b(y)$ income value y will be heaped to one of the S heaping points $h_1, \ldots, h_S$. Because the probability of heaping y to h_b is zero outside the related catchment area I_b, alternatively, $v(y)$ can be written as $v(y) = \sum_{b:y \in I_b} v_b(y)$. The probability for y not being heaped is $1 - v(y)$.

Subsequently, the vector ϕ is defined to comprise all parameters necessary to fully determine $v_b(y)$ and $v(y)$. That is, in the case of piecewise constant heaping probabilities $\phi = (p_b)_{b \in \{1,\ldots,S\}}$ and in the case of steadily increasing/decreasing heaping probabilities $\phi = (\eta_b)_{b \in \{1,\ldots,S\}}$.

3 Model estimation

To fully capture the heaping mechanism interfering when income values are reported, we have to determine the unknown parameter vectors ψ and ϕ of the true income distribution and of the heaping model. For this reason, in a first step, we construct the likelihood function of observing z_i. For that, we ally the latent income distribution and the heaping model as follows: If the observed value z_i is not heaped, that is, $z_i = y_i$, the probability of observing z_i is

$$g_1(z_i \mid \psi, \phi) = \big(1 - v(z_i \mid \phi)\big) f(z_i \mid \psi). \tag{4}$$

Please note that this definition also accounts for the fact that values located at heaping points might be reported correctly. Otherwise, if z_i is heaped to a heaping point h_b, that is, $z_i \neq y_i$, the corresponding probability is

$$g_2(z_i \mid \psi, \phi) = v_b(y_i)\big(F(t_b \mid \psi) - F(s_b \mid \psi)\big). \tag{5}$$

In words, the probability of observing a value z_i, which is heaped to h_b, is determined by the difference between the cdf at the upper bound and the cdf at the lower bound multiplied by the probability of heaping its unobserved correspondent y_i to h_b ($y_i \in I_b \setminus h_b$). Clearly, in the case of constant heaping probabilities, $v_b(y_i)$ is p_b. However, in the case of steadily increasing/decreasing heaping probabilities, $v_b(y_i)$ cannot be derived so easily. Nevertheless, one way to determine g_2 is using the heaping probability $v_b(E(y_i))$ of the expected value of y_i, for $y_i \in I_b \setminus h_b$ instead of $v_b(y_i)$. If the width of I_b is chosen to be reasonably small, $E(y_i) \approx \lim_{\epsilon \to 0} h_b \pm \epsilon$, and thus $v_b(E(y_i)) \approx \eta_b$. Hence, for the function g_2 we yield the following representation

$$g_2^*(z_i \mid \psi, \phi) = \begin{cases} p_b\big(F(t_b \mid \psi) - F(s_b \mid \psi)\big), & \text{for heaping probability function (2)}, \\ \eta_b\big(F(t_b \mid \psi) - F(s_b \mid \psi)\big), & \text{for heaping probability function (3)}. \end{cases}$$

Combining the functions g_1 and g_2^* yields the likelihood function g of observing z_i:

$$g(z_i \mid \psi, \phi) = g_1(z_i \mid \psi, \phi) + \mathbf{1}(z_i \in H) g_2^*(z_i \mid \psi, \phi),$$

where

$$\mathbf{1}(z_i \in H) = \begin{cases} 1, & \text{if } z_i \in H, \\ 0, & \text{otherwise.} \end{cases}$$

indicates whether z_i is on a heaping point or not. In a second step, we define the log-likelihood function l of observing the income data at hand:

$$l(\theta \mid \boldsymbol{z}) = \sum_{i=1}^{n} \ln g(z_i \mid \theta) \tag{6}$$

with $\theta = (\psi, \phi)$. Maximizing l yields estimates $\hat{\theta} = (\hat{\psi}, \hat{\phi})$ for the parameter vector $\theta = (\psi, \phi)$:

$$\hat{\theta} = \text{argmax}_\theta \; l(z \mid \theta).$$

In the optimization process, we have to account for the constraints imposed on the parameter vectors ψ and ϕ. That is, we have to ensure that the parameters a, b, and q of the Dagum distribution (1) are bigger than zero and that the heaping probabilities p_b or η_b range between zero and one. The following constraint system summarizes these requisites:

(i) $a > 0, b > 0,$ and $q > 0$,
(ii) $p_b \in [0, 1]$ and $\eta_b \in [0, 1]$, respectively, for all $b = 1, \ldots, s$,
(iii) $v(y_i) \in [0, 1]$ for all $y_1, \ldots, y_n$.

This system can be specified in the form of inequality equations which are linear in the components of ψ and ϕ. More precisely, the constraints (i) corresponding to the parameters of the latent distribution are linear inequality equations as such. The same applies to the constraints (ii) corresponding to the parameters of the heaping probability functions (2) and (3). The constraints (iii) being due to function g_1 (i.e., $y_i = z_i$) are made up by sums of terms of the form 'component of ϕ multiplied by a known factor'. More precisely, for heaping probability function (2) the constrains (iii) are

$$\sum_{b:z_i \in I_b} p_b \in [0,1] \text{ for all } z_i, i = 1, \ldots, n.$$

Hence, they can be described by sums of p_b (multiplied by factor 1). The corresponding constraint system being up to heaping probability function (3) is

$$\sum_{b:z_i \in I_b} \eta_b C_b(z_i) \in [0,1] \text{ for all } z_i, i = 1, \ldots, n$$

with $C_b(z_i)$ is

$$C_b(z_i) = \begin{cases} \exp\{-2\xi_b^{-2}(z_i - h_b)^2\}, & \text{if } z_i \in I_b, \text{ for } z_i \neq h_b, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the constraint systems corresponding to the heaping probability functions (2) and (3) have a linear representation in the components of ϕ.

The optimization problem at hand is a classical non-linear optimization problem with a linear constraint system. Such a problem can be solved, for example, using the "BFGS-B" algorithm (Byrd et al., 1995) which is a variant of the Broyden-Fletcher-Goldfarb-Shanno algorithm allowing for box constraints (i.e., to each variable a lower and a upper bound is assigned). Accordant functionality is provided, for example, by the *constrOptim* function, which is part of the basic configuration of the statistical software R.

A large number of parameters might hamper the success and the efficiency of the maximization procedure. For instance, let all multiples of 50 be heaping points in the interval from zero to 5,000. Then, 101 parameters would have been necessary to estimate to solely determine the heaping probabilities. A way to counteract this problem is to further restrain the parameter space, for example, by assuming that some components of ϕ are equal. In the concrete case this means to assume, for example, congenial heaping behavior associated with multiples of 100 up to 3,000 and congenial heaping behavior regarding multiples of 1,000 equal to or more than 3,000, see section 4 for illustration (cf. Table 1 or Table 4).

4 Simulation

To underpin the feasibility of the heuristic suggested to identify heaping points from given data (cf. section 2.2) and to test the validity of the heaping model proposed, we perform three simulation studies. In all three studies, we draw $N = 10,000$ values from a Dagum distribution parameterized with $a = 3.6$, $b = 2,416$, and $q = 0.43$. In the first simulation study we assume uniform heaping patterns, that is, we rely on heaping probability function (2). The second simulation study builds on steadily increasing/decreasing heaping probabilities, that is, is based on heaping probability function (3). In the third study, we test whether our method is also capable of dealing with very high proportions of heaped data. All chosen settings constitute the following values as heaping points: zero[2], multiples of 100 up to 5,000, multiples of 500 up to 10,000, and

[2]For convenience, zero corresponds here de facto to a very small positive value, such as 10^{-6}, because the Dagum distribution is only defined for values greater than zero; see section 2.1.

multiples of $1,000$ up to $10,000$. In sum, this yields 61 heaping points. The catchment areas of all heaping points are determined to be symmetrical around the respective heaping point. Here, zero poses an exception because (in our consideration) negative income values are meaningless. The widths of the distinct catchment areas depend on the magnitude of the heaping point itself. We determine catchment areas associated with heaping points, which are multiples of 100 and not of 500 (Mod100), to have width 100. Likewise, heaping points that are multiples of 500 and not of $1,000$ (Mod500) as well as heaping points that are multiples of $1,000$ (Mod1000) feature width 500 and width $1,000$, respectively. To zero a catchment area from zero to 250 is assigned. As already mentioned in the previous section, a large number of parameters to estimate decreases the chance to find an optimal solution. The setting described so far calls for an estimation of 64 parameters in total: 61 parameters for the heaping probability function (one parameter for each heaping point) and three parameters for the underlying true distribution. The number of parameters can be reduced remarkably by assuming similar heaping behavior in each of the eight intervals $[0;500]$, $(500;1,000]$, $(1,000;1,500]$, $(1,500;2,000]$, $(2,000;3,000]$, $(3,000;4,000]$, $(4,000;5,000]$, and $(5,000;10,000]$. That is, in each of these intervals, the probabilities of heaping to a multiple of 100, 500, and $1,000$, respectively, are assumed to be identical. This results in the following grouping of heaping probabilities:

Set 1: Probability of heaping to zero

Set 2, ..., Set 8: Probabilities of heaping to a multiple of 100 and not of 500 (Mod100) in the intervals $[0;500]$, $(500;1,000]$, $(1,000;1,500]$, $(1,500;2,000]$, $(2,000;3,000]$, $(3,000;4,000]$, and $(4,000;5,000]$

Set 9, ..., Set 14: Probabilities of heaping to a multiple of 500 and not of $1,000$ (Mod500) in the intervals $[0;500]$, $(1,000;1,500]$, $(2,000;3,000]$, $(3,000;4,000]$, $(4,000;5,000]$, and $(5,000;10,000]$

Set 15, ..., Set 20: Probabilities of heaping to a multiple of $1,000$ (Mod1000) in the intervals $(500;1,000]$, $(1,500;2,000]$, $(2,000;3,000]$, $(3,000;4,000]$, $(4,000;5,000]$, and $(5,000;10,000]$.

These sets are formed by taking the nature of income values reported in the NEPS Adult Cohort, Wave 2009/2010 (cf. section 5), into consideration. Please note that in some intervals for some multiples no heaping points exist by definition. For example, in the interval $(500;1,000]$ no multiple of 500 exists that is not also a multiple of $1,000$. By classifying the heaping points as described, the number of parameters to estimate is reduced to 23: Three parameters have to be estimated to determine the Dagum distribution and 20 to determine the heaping probabilities.

Table 1
Heaping probabilities p_b, for $b = 1,\ldots,S$, applied in Simulation 1 "piecewise constant heaping probabilities".

	Zero		Mod100		Mod500		Mod1000	
Interval	Set	Value	Set	Value	Set	Value	Set	Value
$[0;500]$	Set 1	0.40	Set 2	0.35	Set 9	0.20	–	–
$(500;1,000]$	–	–	Set 3	0.40	–	–	Set 15	0.10
$(1,000;1,500]$	–	–	Set 4	0.40	Set 10	0.15	–	–
$(1,500;2,000]$	–	–	Set 5	0.35	–	–	Set 16	0.15
$(2,000;3,000]$	–	–	Set 6	0.35	Set 11	0.15	Set 17	0.10
$(3,000;4,000]$	–	–	Set 7	0.25	Set 12	0.25	Set 18	0.20
$(4,000;5,000]$	–	–	Set 8	0.15	Set 13	0.35	Set 19	0.45
$(5,000;10,000]$	–	–	–	–	Set 14	0.40	Set 20	0.50

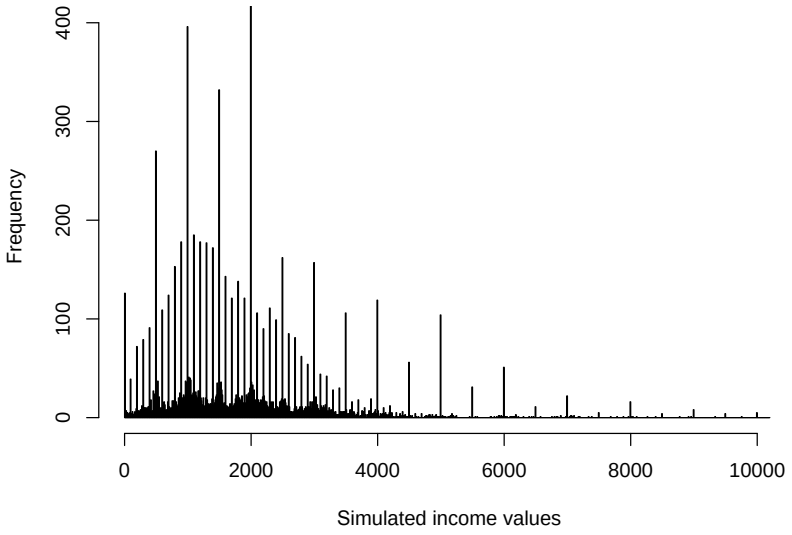

Figure 4. Histogram of simulated income distribution according to Simulation 1 "piecewise constant heaping probabilities".

4.1 Simulation 1: Piecewise constant heaping probabilities

In order to obtain a data set which roughly resembles the heaped income data as reported in the NEPS Adult Cohort, Wave 2009/2010 (cf. Figure 1 on page 4), we shift values drawn from the Dagum distribution according to the heaping probabilities p_b, for $b = 1, \ldots, S$, given in Table 1. The values p_b were chosen quite arbitrarily. Figure 4 depicts the simulated income distribution. In sum, 50.1% of the values are heaped. Table 2 shows how many of the values have been heaped to zero, to multiples of 100, 500, and 1,000.

Table 2
Percentages of values heaped in Simulation 1 "piecewise constant heaping probabilities".

Interval	Zero	Mod100	Mod500	Mod1000	Total
$[0; 500]$	1.24	2.59	2.59	0.00	6.42
$(500; 1,000]$	0.00	5.12	0.00	3.68	8.80
$(1,000; 1,500]$	0.00	6.56	3.07	0.00	9.63
$(1,500; 2,000]$	0.00	4.62	0.00	4.89	9.51
$(2,000; 3,000]$	0.00	6.22	1.44	1.46	9.12
$(3,000; 4,000]$	0.00	1.74	1.01	1.08	3.83
$(4,000; 5,000]$	0.00	0.36	0.53	1.03	1.92
$(5,000; 10,000]$	0.00	0.00	0.55	1.02	1.57
Total	1.24	27.21	9.19	13.16	50.08

As a first step, we apply the heuristical procedure described in section 2.2 to identify heaping points from the data. For this purpose, we first assume a set of hypothetical heaping points, which we then test for being de facto heaping points. To keep it simple, we construct the set of hypothetical heaping points such that it comprises all values of the simulated data set. Applying our heuristic to this set, we find it capable of identifying all 61 heaping points set previously. All detected heaping points are now assigned to one of the sets defined above. Accordant heaping probabilities and the parameters of the underlying Dagum distribution are estimated via the maximum likelihood approach described in section 3 using heaping probability

function (2). To solve the actual optimization problem, we use the Broyden-Fletcher-Goldfarb-Shanno algorithm implemented in the R function *constrOptim*. We derive initial values for the parameters of the Dagum distribution by fitting the observed data to a Dagum distribution–disregarding any heaping. For each of the 20 groups of heaping probabilities considered, we find initial values by computing relative frequencies. These are simply calculated as the quotient of half the number of values at the respective heaping points and the number of all values in the accordant catchment areas. On a desktop workstation equipped with Intel(R) Core(TM) i7, CPU 2.80GHz, 8GB RAM, under Windows 7, using a 64bit system, model fitting takes approximately 110 min. Table 3 gives the estimates $\hat{\theta}$ of the model parameters θ, their standard errors, and the respective 95% confidence intervals. Standard errors and confidence intervals are derived by basic bootstrapping. In sum, 100 bootstrap samples are taken. Overall, we

Table 3
Parameter estimates and measures of uncertainty (standard errors and 95% *confidence intervals CI) corresponding to Simulation 1 "piecewise constant heaping probabilities".*

Parameter	True Value	Estimated	Standard Error	CI lower	CI upper
Dagum distribution					
a	3.60	4.48	0.15	4.23	4.72
b	2,416.00	2,919.09	38.47	2,842.38	2,994.98
q	0.43	0.29	0.01	0.26	0.31
Heaping probabilities					
Set 1	0.40	0.36	0.03	0.30	0.40
Set 2	0.35	0.37	0.02	0.33	0.41
Set 3	0.40	0.40	0.01	0.37	0.42
Set 4	0.40	0.41	0.01	0.39	0.44
Set 5	0.35	0.35	0.01	0.33	0.37
Set 6	0.35	0.35	0.01	0.33	0.37
Set 7	0.25	0.24	0.02	0.22	0.27
Set 8	0.15	0.13	0.03	0.08	0.17
Set 9	0.20	0.21	0.01	0.19	0.23
Set 10	0.15	0.16	0.01	0.15	0.18
Set 11	0.15	0.13	0.01	0.11	0.15
Set 12	0.25	0.24	0.02	0.18	0.27
Set 13	0.35	0.36	0.04	0.32	0.46
Set 14	0.40	0.46	0.04	0.40	0.55
Set 15	0.10	0.10	0.01	0.09	0.11
Set 16	0.15	0.16	0.01	0.14	0.17
Set 17	0.10	0.11	0.01	0.10	0.13
Set 18	0.20	0.22	0.02	0.18	0.25
Set 19	0.45	0.45	0.03	0.38	0.50
Set 20	0.50	0.44	0.03	0.36	0.50

find for all parameters reasonable estimates, standard errors, and confidence intervals. The estimates of the heaping probabilities are very precise. Only in the interval $(5,000; 10,000]$, we observe notable discrepancies between estimated and true values of heaping probabilities (this concerns Set 14 and Set 20). The probability of heaping to a multiple of 1,000 (Set 20) is clearly underestimated, while the probability of heaping to a multiple of 500 (Set 14) is overestimated, though not significantly different from the true value. However, this is clearly due to only few observations being made in the accordant income range. With respect to magnitude, also the estimates for the parameters a, b, and q of the underlying distribution are close to the true ones. However, here the estimated values differ significantly–but not enormously–from the true ones. To get an idea about the overall discrepancy of the estimated and true Dagum distribution, we compare their density functions. The left graph of Figure 5 displays the respective curves. In fact, the shapes of both curves are pretty similar. The estimated curve is just a little bit flatter

at its mode. Overall, the maximal difference between both curves is smaller than $6 \cdot 10^{-5}$, which we deem acceptable.

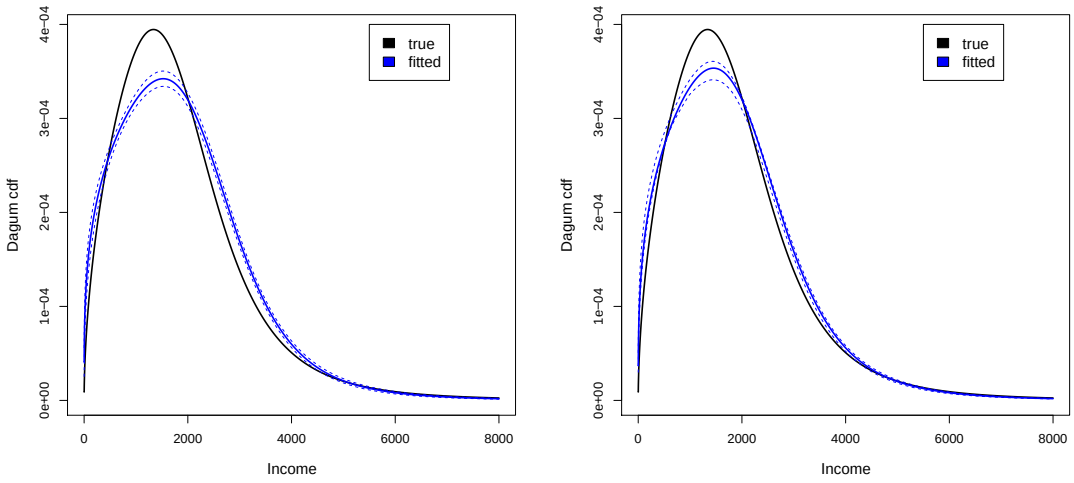

Figure 5. True and estimated density functions of the Dagum distribution. (The dashed lines mark the prediction intervals.). The black line marks the true density function and the blue line indicates the estimated one. The left graph shows the outcome of the piecewise constant model (Simulation 1) and the right graph depicts the outcome of the model assuming steadily increasing/descreasing heaping probabilities (Simulation 2).

4.2 Simulation 2: Steadily increasing/decreasing heaping probabilities

In the second simulation study we heap our data using heaping probability function (3), that is, heaping probabilities are assumed to increase with their proximity to heaping points. Table 4 shows the corresponding parameters η_b, for $b = 1, \ldots, S$. The parameters are chosen such that the shape of the resulting simulated income distribution (cf. Figure 6) resembles more or less the individual net income distribution as observed in the NEPS Adult Cohort, Wave 2009/2010 (cf. Figure 1 on page 4). Apart from that, the parameters η_b were picked at random. In sum, 48.5% of the values in the simulated data were heaped. Table 5 documents the percentages of values being heaped according to the nature of the heaping points. Either they are heaped to zero, to a multiple of 100, of 500, or of 1,000.

Table 4
Heaping probabilities applied in Simulation 2 "steadily increasing/descreasing heaping probabilities".

	Zero		Mod100		Mod500		Mod1000	
Interval	Set	Value	Set	Value	Set	Value	Set	Value
[0; 500]	Set 1	0.55	Set 2	0.55	Set 9	0.25	–	–
(500; 1,000]	–	–	Set 3	0.55	–	–	Set 15	0.20
(1,000; 1,500]	–	–	Set 4	0.55	Set 10	0.25	–	–
(1,500; 2,000]	–	–	Set 5	0.50	–	–	Set 16	0.25
(2,000; 3,000]	–	–	Set 6	0.55	Set 11	0.30	Set 17	0.30
(3,000; 4,000]	–	–	Set 7	0.55	Set 12	0.35	Set 18	0.40
(4,000; 5,000]	–	–	Set 8	0.50	Set 13	0.40	Set 19	0.50
(5,000; 10,000]	–	–	–	–	Set 14	0.55	Set 20	0.50

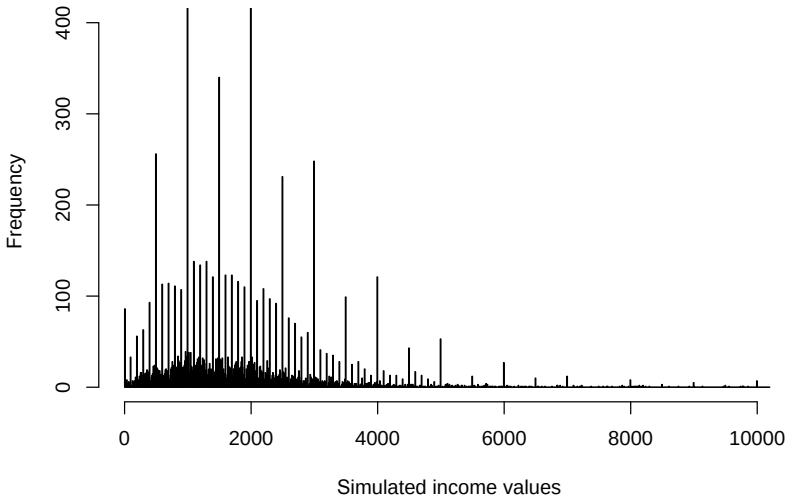

Figure 6. Histogram of simulated income distribution according to Simulation 2 "steadily increasing/descreasing heaping probabilities".

Table 5
Percentages of values heaped in Simulation 2 "steadily increasing/descreasing heaping probabilities".

Interval	Zero	Mod100	Mod500	Mod1000	Total
$[0; 500]$	0.86	2.20	2.36	0.00	5.42
$(500; 1,000]$	0.00	4.02	0.00	4.60	8.62
$(1,000; 1,500]$	0.00	4.90	3.09	0.00	7.99
$(1,500; 2,000]$	0.00	4.22	0.00	4.59	8.81
$(2,000; 3,000]$	0.00	6.09	2.16	2.39	10.64
$(3,000; 4,000]$	0.00	2.15	0.95	1.16	4.26
$(4,000; 5,000]$	0.00	0.95	0.42	0.52	1.89
$(5,000; 10,000]$	0.00	0.00	0.28	0.59	0.87
Total	0.86	24.53	9.26	13.85	48.50

To identify heaping points from the simulated data, we use the heuristical procedure presented in section 2.2. We determine all simulated data points as being potential heaping points. The heuristic performs well and delivers the same set of heaping points as in Simulation 1, that is, it is able to identify all heaping points previously set. After having determined the set of heaping points, we use the maximum likelihood approach described in section 3 to estimate the parameters of our heaping model. For this purpose, we again use the Broyden-Fletcher-Goldfarb-Shanno algorithm. Initial values for the parameters of the Dagum distribution are the result of fitting the observed data to a Dagum distribution, and initial values for the parameters of the heaping probability function are relative frequencies. The accordant frequencies are obtained by dividing the number of values at the respective heaping points by the number of values in the accordant catchment areas. With approximately 80 min, the procedure needs considerably less time to fit the model to the data as under simulation Setting 1. Table 6 shows the estimates $\hat{\theta}$ of the model parameters θ, their standard errors, and the respective 95% confidence intervals. Again, 100 bootstraps have been taken to yield the latter ones. In sum, our approach reproduces the parameters η_b, for $b = 1, \ldots, 20$ of heaping probability function (3) with high accuracy. However, we observe notable–though, not highly significant–deviations. In the interval $[0; 1,000]$, the probabilities of heaping to a multiple of 100 different from 500 are underestimated (concerning

Table 6
Parameter estimates and measures of uncertainty (standard errors and 95% confidence intervals CI) according to Simulation 2 "steadily increasing/descreasing heaping probabilities".

Parameter	True Value	Estimated	Standard Error	CI lower	CI upper
Dagum Distribution					
a	3.60	4.28	0.09	4.11	4.45
b	2,416.00	2,801.25	41.37	2,737.57	2,900.03
q	0.43	0.31	0.01	0.29	0.34
Heaping Probabilities					
Set 1	0.55	0.53	0.05	0.46	0.64
Set 2	0.55	0.51	0.02	0.47	0.56
Set 3	0.55	0.51	0.02	0.46	0.54
Set 4	0.55	0.52	0.02	0.48	0.55
Set 5	0.50	0.48	0.02	0.45	0.51
Set 6	0.55	0.56	0.02	0.53	0.59
Set 7	0.55	0.53	0.03	0.48	0.58
Set 8	0.50	0.54	0.04	0.44	0.61
Set 9	0.25	0.30	0.02	0.26	0.33
Set 10	0.25	0.28	0.01	0.25	0.31
Set 11	0.30	0.32	0.02	0.29	0.36
Set 12	0.35	0.37	0.03	0.31	0.42
Set 13	0.40	0.35	0.03	0.29	0.43
Set 14	0.55	0.53	0.07	0.42	0.70
Set 15	0.20	0.21	0.01	0.19	0.23
Set 16	0.25	0.24	0.01	0.22	0.26
Set 17	0.30	0.31	0.01	0.28	0.33
Set 18	0.40	0.39	0.03	0.34	0.44
Set 19	0.50	0.46	0.05	0.39	0.56
Set 20	0.50	0.52	0.04	0.43	0.59

Sets 2 and 3), while the probability of heaping to 500 is overestimated (Set 9). Similarly, in the interval $(4,000; 5,000]$, the parameter determining the heaping probabilities corresponding to multiples of 100 that are not also multiples of 500 is overestimated (Set 8). Compared with this, the parameters corresponding to heaping probabilities related to multiples of 500 and 1,000 are underestimated (Sets 13 and 19). For values greater than 4,000, we find that our parameter estimates are less accurate (Sets 8, 13, 19, 14 and 20). However, this is clearly caused by the small number of values in this interval. To sum up, for the design chosen, Simulation 1 reproduces the parameters specifying the true heaping probability function more accurately than Simulation 2. Opposing the true and estimated parameters of the Dagum distribution, we find that Simulation 2 produces better results than Simulation 1: The parameters estimated within Simulation 2 are closer to the true ones, even though they still differ significantly. However, putting the estimated and the true density function of the Dagum distribution on top of each other shows only slight discrepancies between both curves, as can be seen in the graph on the right in Figure 5. In total, the maximal discrepancy between the two points of the density functions is smaller than $5 \cdot 10^{-5}$.

Considering the complexity of the simulation design presented, we rate the results of both simulation studies as being pretty good, though, impediments are obvious. Generally, our approach has to tackle three challenges: First, in both simulation settings considered, catchment areas for heaping points are nested. Nearly all possible income values lie in the catchment areas of two or three heaping points. For instance, the value 123 is attracted by zero and 100, and the value 4,966 lies in the catchment area of 4,500, 4,700, and 5,000. Thus, we suspect that our approach faces difficulties in always assigning heaped values to the correct catchment areas. Second, in our model we assume heaping behavior to depend on the magnitude of the true income value as well as on the types of accessible heaping points–resulting in a large number of parameters

to estimate (in sum, 23). This has clearly a substantial effect on the significance and efficiency of parameter estimates. Finally, both simulated data sets feature a high percentage of heaped values. This complicates the discrimination of heaped and non heaped values. That is, we expect our approach to perform better if (i) catchment areas are less nested or even disjoint, (ii) less parameters are used to specify the heaping probability function, and/or (iii) a smaller proportion of data is heaped. Alternative simulation studies carried out by us substantiate this suspicion.[3] Nonetheless, we have designed our simulation studies with the objective of resembling the (individual) net income data reported in the NEPS Adult Cohort. Hence, concerning grouping of heaping points and the structure of catchment areas we adhere to the setting presented here. Simplifying our setting in this direction would be contradicting to what we observe in the data and thus would make it inapplicable for our purposes.

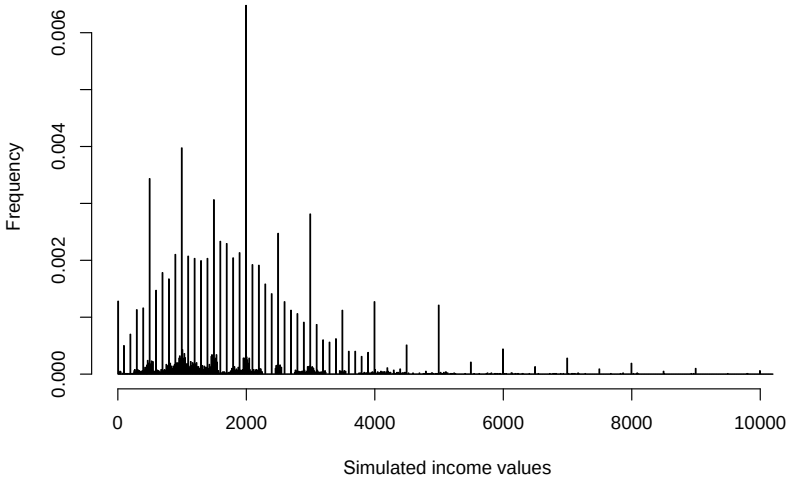

Figure 7. Histogram of simulated income distribution according to Simulation 3 "high proportion of heaped data".

4.3 Simulation 3: High Proportion of Heaped Data

To comprehensively map the NEPS income data, we have to go a step further. The NEPS income data features approximately 70 percent of heaped values (cf. section 5), which significantly exceeds the proportion of values heaped in the simulation studies considered. To test whether our approach is also capable of dealing with such a high proportion of heaped data, we conduct a further simulation study. For this purpose, we increase the parameters determining the heaping probabilities, while otherwise relying on simulation Scheme 1. This way, we obtain a data set with circa 70% of heaped values (precisely, 69.7%). Figure 7 depicts the resulting income distribution. The respective parameters are given in the second column of Table 7. The third column of the table contains the parameter estimates that we find. All results underline the feasibility of our method. However, its impediments are obvious: First, the estimates of the parameters of the Dagum distribution only roughly resemble the true parameters. Second, due to the very high percentage of heaped values, some of the re-estimated heaping probabilities are less accurate. Specifically, this concerns values in the interval $(3,000; 4,000]$: Here the heaping probability corresponding to multiples of 100 that are not multiples of 500 is underestimated (Set 7), while the accordant probability indicating heaping to multiples of $1,000$ is overestimated (Set 18).

[3] On request, simulation settings, data, and results are available from the lead author.

Furthermore, as already noted above, we find wide confidence intervals for areas with only few observations, that is, mainly for ranges of high income.

Table 7
Parameter estimates and measures of uncertainty (standard errors and 95% confidence intervals CI) of the extended simulation study.

Parameter	True Value	Estimated	Standard Error	CI lower	CI upper
Dagum Distribution					
a	3.60	4.99	0.21	4.47	5.32
b	2,416.00	2,980.70	57.51	2,887.76	3,104.73
q	0.43	0.26	0.01	0.23	0.29
Heaping Probabilities					
Set 1	0.45	0.44	0.03	0.39	0.50
Set 2	0.45	0.48	0.03	0.42	0.53
Set 3	0.50	0.50	0.01	0.47	0.52
Set 4	0.50	0.49	0.01	0.46	0.51
Set 5	0.60	0.57	0.01	0.54	0.59
Set 6	0.60	0.57	0.02	0.53	0.61
Set 7	0.55	0.49	0.03	0.42	0.54
Set 8	0.15	0.14	0.03	0.07	0.18
Set 9	0.25	0.25	0.01	0.22	0.27
Set 10	0.15	0.15	0.01	0.14	0.17
Set 11	0.20	0.20	0.02	0.17	0.24
Set 12	0.25	0.25	0.03	0.21	0.30
Set 13	0.35	0.32	0.04	0.24	0.37
Set 14	0.40	0.39	0.05	0.29	0.46
Set 15	0.10	0.10	0.00	0.09	0.11
Set 16	0.20	0.22	0.01	0.20	0.25
Set 17	0.20	0.22	0.01	0.20	0.25
Set 18	0.20	0.25	0.02	0.22	0.30
Set 19	0.50	0.54	0.04	0.49	0.64
Set 20	0.50	0.51	0.04	0.43	0.58

5 Application

To illustrate the potential and also the impediments of our novel approach, we model the heaping behavior evidently present in the individual net income data reported in the NEPS Adult Cohort, Wave 2009/2010. In this data set, $N = 8,685$ persons gave (usable) information about their individual net income. Figure 1 shows the respective frequency distribution. The mean is located at €1,881 and median at €1,700 indicating a distribution skewed to the right. The empirical standard deviation is €1,303. Within the values €331 and €4,200 90% of the probability mass is distributed, marked by the $5th$ and $95th$ percentile. Abnormal concentrations of reported values can clearly be seen. Above the mode of €2,000, spikes at values ending with 500 (Mod500) and 1,000 (Mod1000) are quite obvious. The proportions of these heaping points are given in Table 8 with regard to the intervals conceived in section 4. The relative frequency of values ending in 500 is about 10% and of values ending in 1,000 is about 14%. Most values are rounded to multiples of 100, circa 45%. In sum, more than 70% of the values are rounded to either zero, a multiple of 100, 500, or 1,000. In the intervals (2,000; 3,000], (1,500; 2,000], and (1,000; 1,500], we find their concentration to be highest–which is not surprising when taking into account that these intervals comprise the majority of income values (63.2%).

To determine 'real' heaping points, we restrain the set of potential heaping points to multiples of 100 and apply the heuristic described in section 2.2. Justification for this restraint is to keep

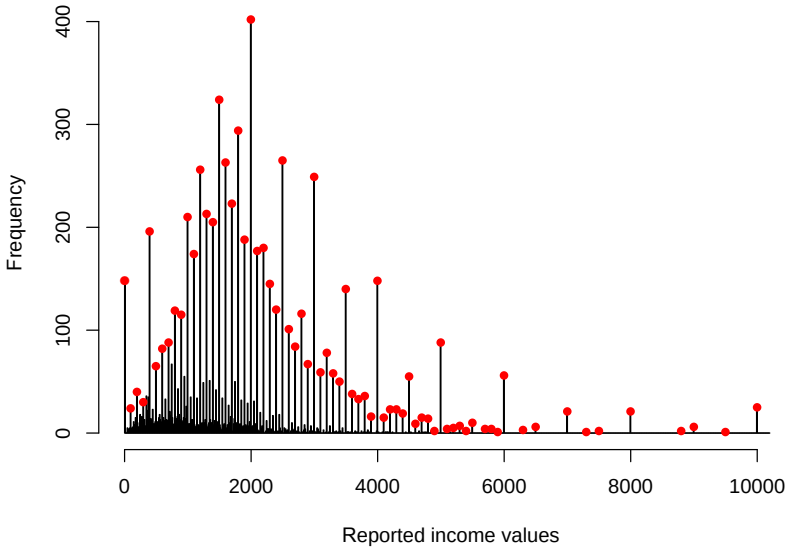

Figure 8. Histogram of individual net income data reported in the NEPS Adult Cohort, Wave 2009/2010 ($N = 8,685$), with heaping points marked by red dots.

the problem to be studied reasonably small. Clearly, such confinement limits the scopes of our analysis: On the one hand, we do not take into account values ending with 5, 10, or 50 for being heaping points. On the other hand, extraordinary heaping points that can be seen in our data, such as 399,[4] are excluded a priori. However, exploiting the data reveals that each of these values occurs less than eight times, which is less than 0.1% of the data size. Therefore, we argue that determining the set of potential heaping points as multiples of 100 is a viable compromise to reduce the complexity of the problem under study. In total, we find 70 heaping points. Figure 8 illustrates them as red points. In this setting, the number of parameters is 73: 70 parameters for the heaping model and three for the underlying distribution. To reduce the number of parameters for estimation, we deem it reasonable to rely on the assumptions presented in section 4. As already argued above, it is a practicable way to facilitate model estimation by imposing equality constraints on the parameters to be estimated. That is, we assume that 20 parameters are sufficient to describe the heaping pattern prevalent in our data (cf. the specification of groups of equal heaping probabilities given on page 10).[5] The definition of catchment areas also borrows from the experiences made in the simulation studies. In accordance with this, we define zero to feature a catchment area from zero to 250. To multiples of 100 that are not multiples of 500 (Mod100) we assign catchment areas of width 100, and to multiples of 500 that are not multiples of 1,000 (Mod500) we assign catchment areas of width 500. Likewise, multiples of 1,000 (Mod1000) are specified to feature catchment areas of width 1,000.

In the following, we focus our analysis more on heaping behavior than on the true income distribution. Our simulation studies show that (in our setting) a heaping model with piecewise constant heaping probabilities yields more accurate estimates of heaping probabilities than a heaping model with steadily increasing/decreasing heaping probabilities. Therefore, relying on these findings, we use a heaping model with piecewise constant heaping probabilities for our

[4] Since April 1, 2003, the threshold value before paying social security contributions into the German system is €400. Thus, there is reason to presume that 399 is a misreported value.

[5] The simulation design does not comprise heaping points greater than 5,000 that are multiples of 100 and not of 500 (Mod100). In the empirical data, however, we identify 5,100, 5,200, 5,300, 5,400, 5,700, 5,800, 5,900, and 6,300 as being heaping points. For convenience, we add these points to Set 8.

Table 8
Percentage of values located at the modulos in the income data of the Adult Cohort in the NEPS, Wave 2009/2010 ($N = 8,685$).

Interval	Zero	Mod100	Mod500	Mod1000	Total
$[0;500]$	1.69	2.99	0.74	0.00	5.42
$(500;1,000]$	0.00	4.38	0.00	2.35	6.73
$(1,000;1,500]$	0.00	9.59	3.68	0.00	13.27
$(1,500;2,000]$	0.00	11.07	0.00	4.61	15.68
$(2,000;3,000]$	0.00	11.36	3.04	2.87	17.27
$(3,000;4,000]$	0.00	4.19	1.59	1.70	7.48
$(4,000;5,000]$	0.00	1.37	0.63	1.01	3.01
$(5,000;10,000]$	0.00	0.38	0.22	1.49	2.09
Total	1.69	45.33	9.90	14.03	70.95

analysis rather than a heaping model with steadily increasing/decreasing heaping probabilities. If modeling were aimed at imputing data, the latter model might be the better choice, because it allows a more precise resemblance of the true Dagum distribution.

We estimate our model by applying the maximum likelihood procedure described in section 3. Table 9 shows the corresponding parameter estimates. In addition, it gives their standard errors and 95% confidence intervals. Both statistics have been derived by basic bootstrapping (number of bootstrap samples: 100).

The parameters for the underlying distribution as well as the heaping probabilities are plausible. The parameters estimated for the Dagum distribution result in a density function with a shape typical of income data, that is, unimodal and positively skewed (cf. Figure 9). We yield an expected individual net income of €1,882 (standard deviation: €1,159)[6] which is more or less the same as the arithmetic mean derived from the heaped data, €1,881 (sample standard deviation: €1,303). As presupposed, we find differences between the percentiles estimated from the heaped data and the percentiles estimated from the true underlying Dagum curve:

	$25th$	$50th$	$75th$
Heaped data	€1,000	€1,700	€2,402
Estimated Dagum curve	€1,007	€1,771	€2,570

In sum, however, the discrepancies are smaller than expected. The only exception being the value of the $75th$ percentile, which differs remarkably.

Concerning the estimated heaping probabilities, five different observations can be pointed out. First, the overall pattern shows that the higher the income of an individual, the more prone he/she is to heap the accordant value. Second, the probability sets for Mod100 (Set 2 up to 8) have the highest values compared to Mod500 (Set 9 up to 14), or Mod1000 (Set 15 up to 20), respectively. Especially when directly comparing competing intervals, the probabilities of heaping to Mod100 remarkably exceed all other probabilities. For example, Sets 9 and 2 cover the interval $[0;500]$, Sets 10 and 4 cover $(1,000;1,500]$, and Sets 11 and 6 cover $(2,000;3,000]$. Here, we find that the probabilities of heaping to Mod500 are substantially smaller than those of heaping to Mod100. We observe a similar pattern when comparing the competing intervals corresponding to Mod1000 and Mod100: Within the intervals $(500;1,000]$ (Sets 15 and 3) and $(1,500;2,000]$ (Sets 16 and 5), the probabilities of heaping to Mod1000 are substantially smaller than the probabilities of heaping to Mod100. The preference for multiples of 100 (compared to

[6] The respective formulas are given in Kleiber & Kotz (2003, p. 214).

multiples of 500 or 1,000) in the range of values up to €3,000 is strengthened by the size of the corresponding standard errors. These are quite small. This means that–although our approach is confronted with highly nested heaping intervals–it assigns values to the respective heaping intervals with high accuracy. In sum, we find evidence of congenial heaping behavior described earlier, which leads people to heap to Mod100 more likely in the range up to €3,000, and to Mod500 or Mod1000 above. This is accompanied by the finding that the intervals of the highest income ranges, namely, the ones above €4,000, feature heaping probabilities that are highest for Mod500 (Sets 13 and 14) and for Mod1000 (Sets 19 and 20). Here, we observe comparatively large standard errors of probability estimates. This is clearly caused by the few observations made within these ranges.

Table 9
Parameter estimates and measures of uncertainty (standard errors and 95% confidence intervals CI).

Parameter	Estimated	Standard Error	CI lower	CI upper
Dagum Distribution				
a	5.68	0.20	5.32	6.12
b	3,062.20	44.66	3,005.72	3,188.79
q	0.22	0.01	0.19	0.23
Heaping Probabilities				
Set 1	0.41	0.03	0.38	0.47
Set 2	0.35	0.02	0.30	0.37
Set 3	0.43	0.02	0.40	0.46
Set 4	0.63	0.01	0.60	0.66
Set 5	0.61	0.01	0.58	0.63
Set 6	0.59	0.01	0.56	0.62
Set 7	0.52	0.02	0.48	0.57
Set 8	0.33	0.05	0.16	0.38
Set 9	0.06	0.01	0.05	0.08
Set 10	0.14	0.01	0.13	0.15
Set 11	0.19	0.01	0.16	0.21
Set 12	0.21	0.02	0.16	0.24
Set 13	0.28	0.04	0.22	0.36
Set 14	0.27	0.05	0.24	0.45
Set 15	0.07	0.00	0.06	0.08
Set 16	0.15	0.01	0.14	0.16
Set 17	0.20	0.01	0.18	0.21
Set 18	0.27	0.02	0.23	0.30
Set 19	0.37	0.04	0.32	0.48
Set 20	0.39	0.05	0.31	0.46

6 Conclusion

This paper presents a statistical approach for modeling heaping patterns arising in self-reported individual net income data. The main idea is to specify a model for the true underlying (unobserved) income distribution while simultaneously modeling the behavior leading to heaped data. Relying on suggestions made in the literature (see Kleiber & Kotz (2003); Bandourian et al. (2002)), we use the 3-parametric Dagum distribution to describe the true net income distribution. To determine heaping behavior, we employ two distinct models: First, we constitute that, within the catchment areas of heaping points, heaping probabilities are constant–resulting in piecewise constant heaping probabilities. Second, we define heaping probabilities to steadily increase with their proximity to a heaping point–specified by piecewise bell-shaped heaping probabilities. The

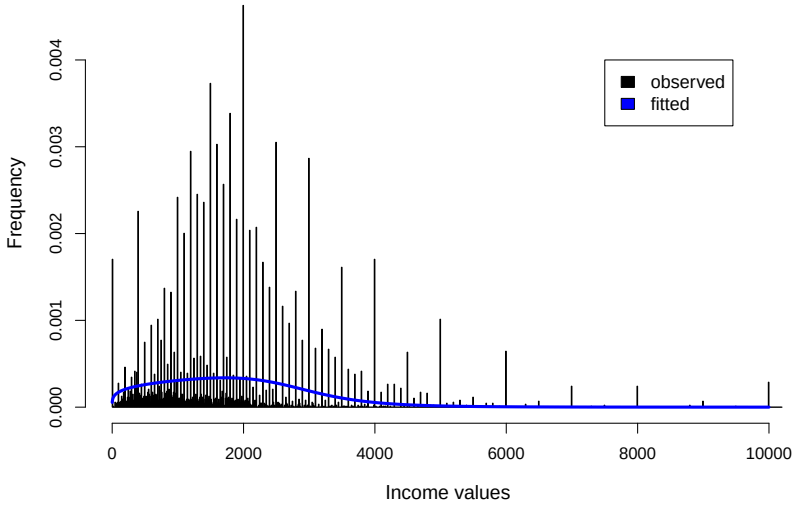

Figure 9. Histogram of individual net income data reported in the NEPS Adult Cohort, Wave 2009/2010, with estimated Dagum density.

first model accounts for the fact that the propensity of an individual to heap depends on his/her true level of income. In contrast, the second model focuses more on the aspect that people's propensity to heap is likely to increase with proximity to a heaping point. The definition of our model demands that all heaping points have to be set, that is known, in advance. To comply with this need, we introduce a heuristic facilitating the identification of heaping points within a given data set. To prove the applicability of this heuristic and the validity of our novel heaping model, we have conducted a set of simulation studies. Irrespective of the presumed heaping mechanism (i.e., piecewise constant or piecewise bell-shaped), the heuristic and the heaping model produce reliable and highly feasible results. However, caveats are obvious: Very high proportions of heaped values (more than 70%) as well as highly nested catchment areas clearly limit the accuracy of parameter estimates. All in all, we find evidence that a heaping model with piecewise constant heaping propensities is better capable of reproducing the true heaping mechanism than a heaping model with steadily increasing/decreasing heaping probabilities. In contrast, the latter heaping model allows us to estimate more accurately the parameters of the Dagum distribution. After having studied the validity of our approach, we applied it to the individual net income data reported in the NEPS Adult Cohort in Wave 2009/2010. We find that for income values of up to €3,000 people have a stronger tendency to heap to multiples of 100 than to multiples of 500 or 1,000. In contrast, in the higher income range (those above €4,000) multiples of 500 and 1,000 are preferred. With respect to the true underlying distribution of the individual net income, contrary to our expectations, we find the median estimated from the heaped data (€1,700) quite close to the one computed from the corrected distribution (€1,771). A similar picture emerges for the $25th$ percentile. Here, we find less than 1% difference. Only the $75th$ percentile differs notably (approximately 7%).

To further improve the modeling of heaping behavior, our method can easily be extended to additionally handle income bracket information. In more detail, most recent surveys dealing with income data ask people who are not willing or not able to give an exact income value to assign themselves into income brackets. This way, data collectors try to counteract item nonresponse and misreporting. In our model, bracket information can be used as a supplement to better figure out the true parameters of the Dagum distribution. De facto, for this purpose, only an accordant term has to be added to the model's likelihood function described in section 3.

When analyzing heaped income data, special points have to be observed that, by nature, feature high concentrations of values, such as €0 indicating not having any income and €399 which is the value directly below the threshold value of €400 for paying social security contributions into the German tax system. So far, our analysis excludes all values that seem to be be 'unusual', that is, this issue has yet not been addressed in our model and presents a subject for future research.

Another issue which still has to be addressed is whether, instead of the 3-parameter Dagum distribution, a 2-parameter distribution such as the log-normal distribution would also suffice to describe the true underlying income distribution. Our simulation studies show that significantly different parameters of the same magnitude result in similar Dagum density functions. This is in contrast to the parameters of, for example, the log-normal distribution. Here, already very small changes in the parameters lead to considerably different density functions. Thus, we suspect that for our purposes the log-normal distribution could suffice as well. However, this has yet to be tested.

We are aware that our approach can also be extended in several other directions. First, up to now we considered symmetric heaping behavior only. This assumption is rather restrictive. For example, it is well known that many people tend to downsize their real income (see, e.g., Maynes (1968)). Our idea of complying with such behavior is to describe heaping probabilities using a logarithmic version of the bell-shaped function, which we suggest would define steadily increasing/descreasing heaping probabilities. Furthermore, we know that heaping propensities are likely to depend on individual characteristics, such as gender and age, and also on external factors, such as interview conditions, interview mode, and interviewer characteristics. Such additional aspects require a consideration of covariates in the heaping model. Likewise, covariates should be considered when describing the true and unobserved income distribution. Here, the ideas recently presented by Drechsler & Kiesl (2014) give insights for further research. They suggest using a log-normal distribution to describe income data and to classify heaping behavior by using an ordered probit model.

As already discussed in the introduction, having identified the true unobserved income distribution allows us to improve any analysis based on heaped income data. To this end, heaped and missing values must simply be replaced by values drawn from the true income distribution. A technique that we regard useful in this context is the method of multivariate imputation by chained equation (mice) (Raghunathan et al., 2001; van Buuren et al., 2006; van Buuren & Groothuis-Oudshoorn, 2011). This technique has been designed to encounter item-nonresponse by replacing missing values with predicted ones on a variable-to-variable basis. That is, once we have extended our heaping approach to allow for the consideration of covariates determining the level of income, we are planning to embed it into the mice framework in order to adequately impute heaped income data.

Generally, our approach can be extended without further ado to also deal with other types of heaped data, for instance, with duration data. For this purpose, the underlying model can simply be replaced by a model consistent with the variable of interest. To describe duration data, for example, a piecewise exponential model is well suited; see, for example, van der Laan & Kuijvenhoven (2011).

Acknowledgements

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6–Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:1.0.0. The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by

the Federal States.

References

Bandourian, R., McDonald, J. B., & Turley, R. S. (2002). *A comparison of parametric models of income distribution across countries and over time.* Brigham Young University: Department of Economics.

Bar, H. Y., & Lillard, D. R. (2012). Accounting for heaping in retrospectively reported event data–a mixture-model approach. *Statistics in Medicine*, *31*(27), 3347–3365.

Beaman, J., & Grenier, M. (1998). Statistical tests and measures for the presence and influence of digit preference. In H. G. Vogelsong (Ed.), *Proceedings of the 1997 Northeastern Recreation Research Symposium* (pp. 44–50). Bolton Landing, NY: Radnor, PA: U.S. Department of Agriculture.

Blossfeld, H.-P., Rossbach, H. G., & Maurice, J. v. (Eds.). (2011). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (Vol. 14 Special Issue). Wiesbaden: VS Verlag für Sozialwissenschaften.

Byrd, E. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190–1208.

Camarda, C. G., Eilers, P. H. C., & Gampe, J. (2007). Modelling general patterns of digit preference. In J. del Castillo, A. Espinal, & P. Puig (Eds.), *Proceedings of the 22nd International Workshop on Statistical Modelling, Barcelona* (pp. 148–153). Barcelona: Institut dÂťEstadistica de Catalunya, IDESCAT.

Drechsler, J., & Kiesl, H. (2012). *MI double feature: Multiple imputation to address nonresponse and rounding errors in income questions simultaneously.* Nuremberg and Regensburg.

Drechsler, J., & Kiesl, H. (2014). *Beat the heap–an imputation strategy for valid inference from rounded income data* (Discussion Paper No. 2/2014). Nuremberg and Regensburg: IAB (Institute for Employment Reseach).

Gill, R. D., van der Laan, M. J., & Robins, J. M. (1997). Coarsening at random: Characterisations, conjectures, counter-examples. In D. Y. Lin & T. R. Fleming (Eds.), *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis* (pp. 255–294). Springer-Verlag.

Hanisch, J. U. (2005). Rounded responses to income questions. *Allgemeines Statistisches Archiv*, *89*(1), 39–48.

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics*, *19*(4), 2244–2253.

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences.* Hoboken, New York: John Wiley & Sons, Inc.

Leopold, T., Raab, M., & Skopek, J. (2011). *Data Manual: Starting Cohort 6 - Adult Education and Lifelong Learning.* Bamberg, National Educational Panel Study.

Maynes, E. S. (1968). Minimizing responses errors in financial data: The possibilities. *Journal of the American Statistical Association*, *63*(321), 214–227.

Miller, H. P., & Paley, L. R. (1958). Income reported in the 1950 Census and on income tax returns. In Conference on Research in Income and Wealth (Ed.), *An Appraisal of the 1950 Census Income Data* (pp. 177–204). Princeton University Press.

Pickering, R. M. (1992). Digit preference in estimated gestational age. *Statistics in Medicine*, *11*(9), 1225–1238.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

Roberts, J. M., & Brewer, D. D. (2001). Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, *28*(7), 887–896.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Torelli, N., & Trivellato, U. (1993). Modelling inaccuracies in job-search duration data. *Journal of Econometrics*, *59*(1-2), 187–211.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

van der Laan, J., & Kuijvenhoven, L. (2011). *Imputation of rounded data* (No. 201108). The Hague/Heerlen.

Wang, H., Shiffman, S., Griffith, S. D., & Heitjan, D. F. (2012). Truth and memory: Linking instantaneous and retrospective self-reported cigarette consumption. *The Annals of Applied Statistics*, *6*(4), 1689–1706.

Wright, D. E., & Bray, I. (2003). A Mixture Model for Rounded Data. *Journal of the Royal Statistical Society, Series D (The Statistican))*, *52*(1), 3–13.