Making Use of Multiple Imputation to Analyze Heaped Data

F. el Messlaki[1], L. Kuijvenhoven[2], and M. Moerbeek[1]

[1]Department of Methodology and Statistics, Utrecht University, The Netherlands
[2]Statistics Netherlands, The Hague, The Netherlands

Abstract

This paper examined the use of multiple imputation to analyze heaped data. When people are asked to recall certain durations such as unemployment spells, they tend to round their answers off to the nearest year or half year causing abnormal concentrations of response at these durations. In order to model these heaped data, a method is developed which specifies the heaping mechanism and the underlying true model referred to as the estimated underlying model. This model is used to create a new data set using multiple imputation so that new durations are generated for the persons who have rounded off their duration. The recent paper examined whether it is more favourable to obtain the estimates from the estimated underlying model directly or from the method in which multiple imputation is used. A simulation study is performed in which misspecification of the model and misspecification of the heaping mechanism is introduced so that the estimates using the different methods can be compared. The results show that multiple imputation leads to more precise estimates and is more robust for model misspecification than estimates based directly on the estimated true underlying distribution. Both methods seemed to be robust to misspecification of the heaping intervals.

keywords: unemployment duration model, geometric model, misspecification, rounded data, incomplete data.

1 Introduction

When data are analyzed, one common problem that researchers have to deal with is the incompleteness of the data. There are some situations in which the data are neither entirely missing nor perfectly known. Instead, only a subset of the complete data sample space in which the true unobservable data lie, is observed (Heijtjan & Rubin, 1990). This kind of incomplete data is referred to as coarse data, which is a generalization of the missing data mechanism (Rubin, 1976). Censored, grouped, heaped and missing data are all special cases of coarse data and are discussed in Heitjan & Rubin (1990). The current study concerns coarseness in the form of heaped data.

In duration analysis, heaping refers to a phenomenon in which persons tend to round certain durations off to the nearest year, half year or month causing heaps in the data distribution (e.g. Heitjan & Rubin, 1990). Heaping is a form of recall errors caused by memory effects and can affect retrospectively collected data. When respondents are asked when a certain event took place, they often do not remember accurately when this event occurred. This phenomenon, in which the respondent may report events as being more recent or remote than they actually are, is referred to as the forward and backward telescoping effect, respectively (Lynn, 2009).

The Dutch Labour Force Survey (LFS)[1], is subject to a certain heaping pattern in which concentrations of spell lengths at multiples of six months are observed when respondents are asked to recall for how long they have been unemployed. Research from other countries has also been hampered by a certain heaping pattern in the LFS (e.g. Kraus & Steiner, 1998; Torreli & Trivellato, 1993). For example, Torelli and Trivellato (1993) showed that inaccurately reported durations cause clear spikes at multiples of six months and even more pronounced spikes at multiples of twelve months in the Italian LFS. It is therefore very likely that some of the respondents have rounded their durations off to the nearest year or half year.

Using the unemployment durations from the Dutch LFS, Statistics Netherlands makes statistics on long-term and short-term unemployment. As a form of publication, a table containing a number of categories of unemployment durations is used. Measurement errors arise, when the heaped unemployment distribution is used to estimate the number of persons that fall within a certain category. As a consequence, certain number of persons are assigned to one of the categories while they belong to another. Since the heaping pattern leads to a biased distribution of the unemployment durations, a questionable representation of the number of persons within each category is obtained.

In general, the described heaping pattern emerges from episode-based questionnaires, in which the respondents are asked to recall certain episodes or durations but report them inaccurately (e.g. Kraus & Steiner, 1998; Torreli & Trivellato, 1993). However, some heaping may be caused by true behavior. For instance, economic circumstances may lead to a high unemployment rate on a certain time point and cause heaps in the data.

Unfortunately, there is a lack of data that would make it possible to distinguish heaping due to the recall process from heaping due to the true behavior, and a lack of data that would allow inference of the true behavior underlying the heaped responses. In the absence of such data, previous analyses have come up with different models

[1]The LFS is referred to as Enquête Beroepsbevolking (EBB) in Dutch and is collected by Statistics Netherlands

and strong assumptions to distinguish between the true and heaped unemployment data (e.g. Petoussis, Gill & Zeelenberg, (2004); Wolff & Augustin, (2000); Torreli & Trivellato, 1993). For example, Torreli & Trivellato (1993) investigated the effect of the heaps and concluded that it is necessary to take the heaping mechanism into account when the unemployment durations are estimated.

Accordingly, Kuijvenhoven & Van der Laan (2009) tried to model the Dutch LFS by correcting it from the distortion caused by the heaps. They estimated a model that specifies the heaping mechanism and the underlying true duration distribution (which cannot be observed because of the heaping mechanism). For some of the respondents in a heap, multiple imputation is used to generate new durations from the estimated underlying distribution and heaping mechanism.

Imputation is used because it provides more precise estimates. Heijtjan & Rubin (1990) showed this by applying multiple imputation to a demographic data set with coarse age measurements. The heaped ages are imputed using multiple imputation with ages that are plausible using a simple naive model and a new more complex model that relates true age to the observed heaped age values and other background variables. They showed that simple and easily programmed multiple imputation methods improved inferences from this kind of data.

Another reason for using multiple imputation is that it is thought to be less sensitive for model misspecification. Since the prominent elements of the method used by Kuijvenhoven & Van der Laan (2009) are the underlying model and the heaping intervals, misspecification of the model parameters or heaping intervals will presumably result in biased estimates. The applicability of the model should therefore be tested. This is done by adding misspecification in these two elements. Misspecification of the model parameters will indicate whether the distinctive shape of the data is discerned by the model or not. Furthermore, the used method should be robust to misspecification of the heaping intervals. It is still unclear whether assuming that persons make use of forward or backward telescoping affects the estimated durations. Therefore, the misspecification in the heaping intervals will show whether the estimates are robust for violation of these assumptions.

A disadvantage of multiple imputation is that it is time consuming. Using multiple imputation in the method takes three times as long to obtain the estimates. Therefore, a substantial amount of time can be saved if multiple imputation is not necessary to obtain the true unemployment durations. In this paper, the necessity of making use of multiple imputation is tested by introducing the two forms of misspecification into the method with and without imputation.

In brief, misspecification of the model parameters and misspecification of the heaping intervals are introduced to 1) the current method used by Kuijvenhoven & Van der Laan (2009), referred to as imputation method and 2) to the same method without making use of imputation, referred to as the estimated underlying model. These methods will be compared. Based on the results, conclusions can be made about whether it is more favourable to base the estimates directly on the estimated underlying model or on the imputation method.

This paper attempts to answer the following questions:

- Is it more favourable to base the estimates of the durations on the imputed data or directly on the estimated underlying model?
- Is the imputation method more robust to misspecification of the model parameters than the estimated underlying model?

- Is the imputation method more robust to misspecification of the heaping mechanism than the estimated underlying model?

The organization of the paper is as follows: Section 2 presents the method developed by Kuijvenhoven & Van der Laan (2009). In section 3 the implementation of the current study is described. First, the simulation study used to examine misspecification is described. Second, a bootstrap method for the real data is outlined and used to make a comparison between the duration obtained from the different methods. Section 4 provides the results of the analysis of the research questions. The last section sets out the conclusion and the discussion points.

2 Model Specification

The method developed by Kuijvenhoven & Van der Laan (2009) is used to make statistics on long-term and short-term unemployment. The durations obtained from this method will be published in a table containing the unemployment categories as is mentioned in the introduction. Kuijvenhoven & Van der Laan (2009) aim to produce this table with the true unemployment durations corrected from the heaping mechanism. Since the distribution of the true durations x cannot be observed because it is coarsened by the heaping mechanism, the method takes the heaping mechanism into account. This is done by specifying a model for the underlying true distribution and a model for the heaping mechanism. The parameters of these models are estimated after modelling both the heaping mechanism and the underlying true distribution simultaneously. The proportion of persons in a certain heap who have rounded their durations off can then be derived from the estimated models. For these persons, multiple imputation is used to generate new durations from the estimated underlying distribution and heaping mechanism. This process will be described in more details in the following subsections:

- Specify a model for the heaping mechanism (subsection 2.1)
- Specify a model for the underlying true distribution (subsection 2.2)
- Obtain the parameters for both models by modelling the models simultaneously (subsection 2.3)
- Generate new durations from the estimated underlying true distribution for the people who rounded their duration off by making use of multiple imputation (subsection 2.4).

The data used in this study is the Dutch LFS[2] for the years 2002-2007. An illustration of this data for the year 2005 is shown in Figure 1. It seems that the respondents tend to round off to the nearest year or half year and as a result very pronounced heaps arise every multiple of six and twelve months.

The heaping pattern as shown in Figure 1 is also found for the years 2002 until 2007. It is clearly visible that a heaping pattern is present that distorts the data. The next section will describe a model for this heaping pattern.

[2]This data has a rotating panel design. The respondents were contacted for five times. For the first wave, respondents are visited at home by an interviewer from Statistics Netherlands for a Computer Assisted Personal Interviewing (CAPI). The next four waves were obtained using Computer Assisted Telephone Interviewing (CATI).

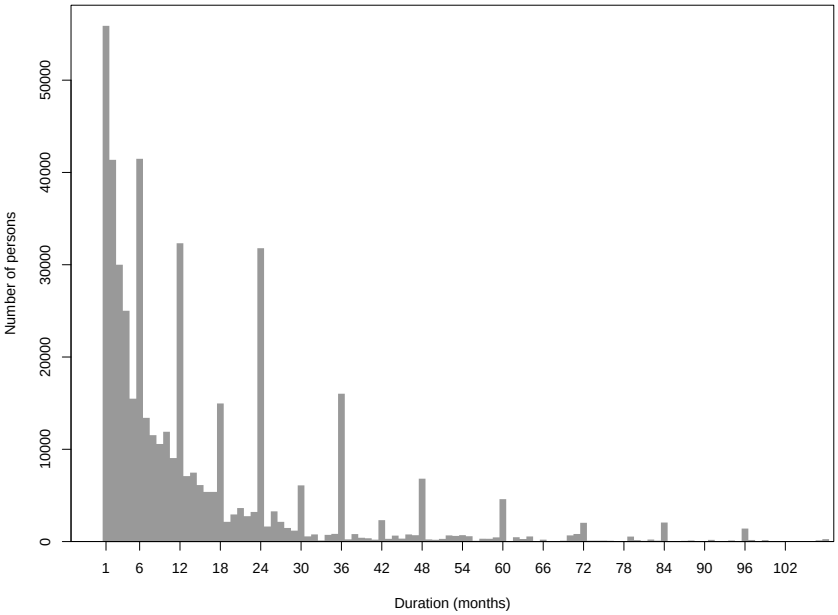

Figure 1: Frequency distribution of durations of unemployment for the year 2005 (Source: Statistics Netherlands)

2.1 The heaping mechanism

The heaping mechanism is used to determine which durations in the heaps are rounded. This section describes a statistical model that defines this heaping mechanism.

It is assumed that every person i in the data has a true duration x_i for unemployment. However, some durations are not reported precisely but are rounded. For those persons who rounded their value off, duration y_i is observed. This means that y_i is the observed value of the true duration x_i. The interest of the study is to find the distribution of x_i.

For this reason, the locations of the heaps should be specified. The locations of the heaps to which people round up or down are given by h_j for which $j = 1, \ldots, k$. The probability that a person i rounds off to a certain heap h_j depends on the true duration x_i and is given by $Pr(h_j|x_i)$.

The probability of rounding off is modelled by a multinomial distribution with $k + 1$ categories. For every duration x_i, there is a probability of rounding off to one of the k heaps $\sum_{j=1}^{k} Pr(h_j|x_i)$ and a probability of not rounding $1 - \sum_{j=1}^{k} Pr(h_j|x_i)$. The sum of these probabilities should equal 1 and therefore $\sum_{j=1}^{k} Pr(h_j|x_i) \leq 1$ for all x_j.

If the observed value y_i is not a duration that is rounded and therefore it is a true duration, then the probability to observe y_i equals the probability that the true

duration x_i equals y_i multiplied with the probability of not rounding which is given by:

$$\Pr(y_i) = f(y_i) - \sum_{j=1}^{k} f(y_i) \Pr(h_j|y_i) = f(y_i) \left(1 - \sum_{j=1}^{k} \Pr(h_j|y_i)\right), \tag{1}$$

in which $f(y)$ refers to the distribution of the observed durations y.

In the case that y_i is a rounded duration and is therefore equal to a h_j, then the probability of rounding off to a heap from a specific duration should be taken into account. This probability is obtained by summing up all y_i that are located within heap h_j and is defined as:

$$\sum_{j:y_i=h_j} \int_{-\infty}^{\infty} Pr(h_j|x) f(x) \mathrm{d}x, \tag{2}$$

The integral in equation 2 can be replaced by a sum when the durations are discrete.

The discussed probabilities all depend on the parameter vector θ. The probabilities are summed up to obtain the total probability of observing y_i. This probability is defined as:

$$Pr(y_i|\theta) = f(y_i|\theta) \left(1 - \sum_{j=1}^{k} Pr(h_j|y_i, \theta)\right) + \sum_{j:y_i=h_j} \int_{-\infty}^{\infty} Pr(h_j|x, \theta) f(x) dx. \tag{3}$$

Suppose the distribution function of the observed durations is defined as $g(y_i|\theta) = Pr(y_i|\theta)$, the log likelihood function can then be defined as:

$$l(\theta|\mathbf{y}) = \sum_{i=1}^{n} \ln g(y_i|\theta). \tag{4}$$

The LFS has a complex survey design used by Statistics Netherlands. Using the information that is collected through the sample, estimations can be made for certain characteristics of the population from which the sample is taken. Hereby, a weight is assigned to every observation in the sample. These weights are a result of a weighting technique which is used by Statistics Netherlands[3]. Those weights should be determined in such a way that the new estimator has better characteristics (in terms of precision and bias) compared to the initial estimator. The weighted distribution of the response variable should be the same as the distribution of the population. This can be obtained by giving the underrepresented groups a smaller weight and the overrepresented groups a larger weight. These weights are included in the loglikelihood function (e.g. (Chambers & Skinner, 2003)). The weighted loglikelihood function can now be defined as:

$$l_w(\theta|\mathbf{y}) = \sum_{i=1}^{n} w_i \ln g(y_i|\theta). \tag{5}$$

[3]The observations are weighted in two steps. In the first step inclusion weights are assigned to the observations, calculated in such a way that they can correct for uneven inclusion probabilities that follow from the applied sampling method. In the second step final weights are determined; this step reduces the bias caused by non-response, using information on sex, age, country of origin, official place of residence and some other regional classifications. In addition, information from registrations on country of origin, registration at employment office and income is used (Statistics Netherlands). The rotating panel design is also explicitly used and all waves are weighted together in one step.

As is mentioned in the introduction, people tend to round off to half years and years. In some heaps, persons are located there because they have rounded their value up to a half year and others to a year. Rounding off to a year is therefore stronger than rounding off to a half year. A person within the sample has therefore more duration to which he or she may round his value to. For example, persons that have a true duration 2.6 may round their durations off to 3 years but also to 2.5 years. Thus, the half year and year intervals are assumed to overlap each other.

Suppose there are k intervals $I_j = [t_{j-1}, t_j)$ with $j = 1, \ldots, k$. The probability to round off is given by p_j and within the interval I_j, persons round off to h_j. Within every interval it is then assumed that the probabilities $\Pr(h_j|y_i)$ are constant.

It should be noted that the parameter vector θ is divided in a parameter vector $\mathbf{p}$ that describes the heaping mechanism and a parameter ϕ that describes the underlying true model which will be described in the next subsection. Formula 3 can now be rewritten as:

$$g(y_i|\phi, \mathbf{p}) = f(y_i|\phi) \left(1 - \sum_{j:y_i \in I_j} p_j\right) + \sum_{j:y_i = h_j} p_j \left(F(t_j|\phi) - F(t_{j-1}|\phi)\right). \quad (6)$$

For all y $\sum_{j:y \in I_j} p_j \leq 1$ should apply.

2.2 Underlying true model

Now that a model is specified for the heaping mechanism, another model should be specified for the underlying true distribution $f(x|\phi)$. Because the unemployment durations are measured in intervals, a discrete time model will be used for the underlying distribution, namely the geometric model. The geometric distribution is a discrete memoryless random distribution. In probability theory, memorylessness is a property of certain probability distributions wherein any derived probability from a set of random samples is distinct and has no information (i.e. "memory") of earlier samples. The probability that someone is unemployed for a certain duration is an event that can be described as the outcome of a sequence of simpler, conditional events. For example, conditional on being unemployed through last month, the probability of finding a job this month is given by λ. The probability of being unemployed for a certain time interval can be given by the following probability function:

$$f(t) = \lambda(1-\lambda)^{t-1} \text{ for } t = 1, 2, 3, \ldots$$

Duration distributions are conveniently specified using the notion of a hazard function. Suppose the distribution of a duration T is specified by its density $f(t)$ or its distribution function $F(t) = \Pr(T < t)$. Two alternative specifications are the survival function $S(t) = 1 - F(t) = \Pr(T \geq t)$ and the hazard function:

$$\lambda(t) = f(t)/S(t) = -dlnS(t)/dt$$

The hazard function at the point t gives, roughly, the conditional probability that a duration will end at t given that it lasts until t or longer. For the geometric distribution, the hazard is assumed to be constant $\lambda(t) = \lambda$ for every interval. The distribution function of the geometric model is given by:

$$F(t) = P(T \leq t) = \sum_{x=1}^{t} \lambda(1-\lambda)^{x-1} = 1 - (1-\lambda)^t \quad (7)$$

The expected value is $\mathrm{E}(T) = 1/\lambda$ and the variance is $\mathrm{var}(T) = (1-\lambda)/\lambda^2$.

The geometric distribution is required to analyze the hazard within every interval. The time axis will be divided in a number of intervals $a_0 < a_1 < \ldots < a_m$ with $a_1 = 0$ and $a_m = \infty$. These intervals are used to apply the geometric distribution for every interval. Since this process takes place step by step, a model is obtained that is called the 'stepwise geometric model'. The cumulative density function of the stepwise geometric model is:

$$F(t) = 1 - \left(\prod_{j=1}^{m(t)-1} (1-\lambda_j)^{(a_j - a_{j-1})} \right) (1-\lambda_{m(t)})^{(t - a_{m(t)-1})} \tag{8}$$

where $m(t)$ is defined as $a_{m(t)-1} \leq t < a_{m(t)}$.

The hazard can now be estimated and then be used to determine the durations of unemployment. It is important to choose an adequate number of intervals and their boundaries should be specified properly by the user in order to describe the true duration distribution in a proper way. The bounds of these intervals are: 1, 2, 3, 4, 5, 10, 15, 20, 50, 100, 200, ∞. As is illustrated in Figure 1, the distribution of the durations is positive skewed which means that relatively more short-term unemployment durations are observed. Therefore the intervals are smaller for these short durations. Another reason for using these small intervals is that the hazard shows strong variations here.

2.3 Implementation of the model

Given that the heaping mechanism and the stepwise geometric model are specified, these models are implemented to obtain one model for the observed distribution. This is done by modelling the stepwise geometric model and the model that defines the heaping mechanism simultaneously to obtain the parameters. The parameter ϕ describes the underlying true model and the parameter vector $\mathbf{p}$ describes the heaping mechanism.

For this model it is assumed that persons round off to the heaps: 6, 12, 24, 30, 36, 60, 72,.... It is assumed that the probability of rounding off from a duration of 72 months and larger is constant. Another assumption is that persons within the interval $[x-2, x+4)$ round off to the duration x. For the first heap (6 months), this means that persons who have a duration of 4, 5, 7, 8 or 9 months tend to round of to duration 6.

Now that the estimated underlying model is specified, the imputation method that is used to generate new duration for the persons who have rounded their durations off will be described in the following paragraph.

2.4 Imputation Method

In this section the imputation method will be described. After the heaping pattern is modelled, the durations that are located within the heaps because of rounding should be imputed. This is done by estimating the proportion $\hat{f}_j$ of durations that are located within a heap as a consequence of rounding. The proportion of durations that are located in a certain heap due to rounding can be given by:

$$\hat{f}_j = \frac{\hat{p}_j \left(F(t_j|\hat{\phi}) - F(t_{j-1}|\hat{\phi}) \right)}{g(h_j|\hat{\phi}, \hat{\mathbf{p}})}. \tag{9}$$

where ϕ describes the underlying true model and $\mathbf{p}$ describes the heaping mechanism.

Subsequently, for these durations a new duration is imputed stochastically using the estimated underlying true distribution and the heaping mechanism. The proportion of recall errors within every duration is estimated using the underlying true distribution and the heaping mechanism. Then, for every rounded duration a new duration is generated from the corresponding interval using the estimated underlying true distribution. For every imputation, a set of new estimates of the durations is obtained. This set of estimates is then averaged to obtain one estimate for every duration. The imputation is repeated 25 times to obtain more precise estimates.

The steps for the multiple imputation are:

- **step 1** For every y_i determine whether it is rounded using the estimated underlying model. If y_i is not rounded, it is equal to x_i and therefore imputation is not needed.
- **step 2** For y_i which is rounded, determine to which heap it rounded using the underlying true model.
- **step 3** For this y_i generate a new duration $\tilde{x}_i$ from:

$$\tilde{x}_i \sim f(y_i|\hat{\phi}, y_i = h^i). \tag{10}$$

- **step 4** Repeat **step 1-3** 25 times.

Step 4 is applied so that each coarse value is replaced with a set of plausible values that represent the uncertainty about the right value to impute. The obtained imputed data sets are then analyzed using standard procedures for complete data. This procedure is referred to as multiple imputation and is first applied by Rubin (1987) who advocated it as a solution for missing data. It results in statistically valid inferences that properly reflect the uncertainty caused by missing values. These steps will result in a new duration for some respondents in the Dutch LFS. For example, for a duration that is rounded to 2 years a new duration will be generated from the corresponding interval which is the interval from 1.5 to 2.5 year with a certain probability. This procedure provides a distribution which is more or less smooth and corrected for the heaps. An illustration of this new distribution is given in Figure 2.

The black line represents the original duration distribution and the gray bars represent the distribution after multiple imputation. This figure shows that the heaps within have disappeared. The durations that were located within a certain heap are now relocated to other duration within the interval to which it belongs.

3 Methods

To evaluate which method (with or without imputation) performs best in providing the estimates and therefore is more robust to a misspecified model, misspecification is added to the model. The next subsections describe the simulation studies used to examine the effect of the misspecification.

3.1 Simulation Study

To examine whether the model with imputation is more robust to model misspecification compared to the model without imputation two simulation studies are performed. In the first simulation study the parameter estimates of the model will be misspecified

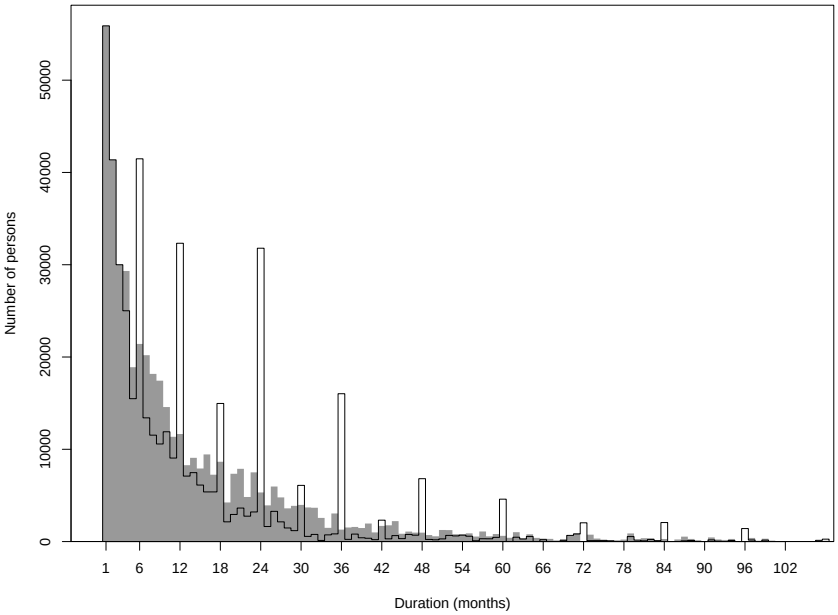

Figure 2: Imputed unemployment duration distribution (gray bars) and the observed distribution (black outlining) of 2005

and in the second simulation study, the intervals of the heaping mechanism will be misspecified.

For these simulation studies, a sample of size $N = 5000$ is taken from a known Weibull distribution with the shape parameter α and scale parameter β. This distribution is often used to describe time to event phenomena (Klein & Moeschberger, 2003). The sample is taken from this specific distribution because its distribution is a good representation of the distribution of the Dutch LFS.

It should be noted though that the unemployment duration in the Dutch LFS is measured as a discrete variable, therefore some modification should be applied to this Weibull distribution to obtain the discrete alternative. The discrete weibull distribution has the following probability mass function:

$$Pr(\lambda; \alpha, \beta) = \alpha^{\lambda^\beta} - \alpha^{(\lambda+1)^\beta} \qquad \lambda = 0, 1, 2, \ldots; 0 < \alpha < 1; \beta > 0 \tag{11}$$

Subsequently, the sample that is taken from this distribution is heaped. This is done by partitioning the axis in a number of intervals of length six and twelve. Thereupon, heaps are created within the distribution by increasing the probability of a certain value within each interval that should represent the heaps of the half year and year. As a result, a sample is obtained in which heaps occur in each interval. The frequency distribution of this sample is illustrated in Figure 3. The location of the heaps, the probabilities and intervals are specified in such a way that it can reflect the Dutch LFS.

Before starting the simulations the underlying true model and the imputation model

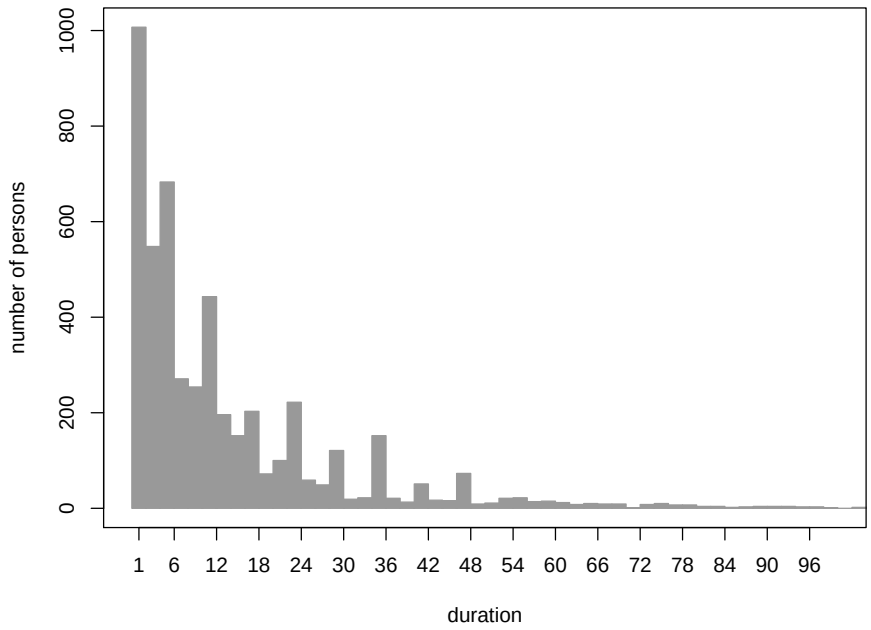

Figure 3: Frequency distribution of the simulated data.

are estimated using the likelihood function and the heaping mechanism. The shape parameter of the Weibull distribution α is first set to 0.8 and the scale parameter γ is set to 12 to obtain the parameters of the true distribution. These specific values are chosen to obtain a skewed distribution similar to the Dutch LFS.

3.1.1 Misspecification of the model parameters

Misspecification of the model parameters provides an incorrect description of the data. Hence, the effect of misspecifying the model parameters is evaluated by performing the following simulation study.

The misspecification is created by changing the shape parameter α of the Weibull distribution from 0.2 to 0.8 in steps of 0.1. The shape parameter determines the shape of the distribution. Misspecifying this parameter is logical when we take into account that the distribution of the Dutch LFS has a certain shape that should be preserved to estimate the durations using this model. The term 'misspecification' is therefore allowed in the first simulation study to cover a broad range of modelling errors for the shape parameter α.

Since the true parameters are known in the simulation, it can be examined whether

the true parameters can be retrieved after misspecifying the model parameters. The simulations will be performed by setting the sample size to 5000. For every simulation the shape parameter is changed and the parameters of the underlying true model with and without imputation are estimated. This process is repeated a number of times, so that 1000 bootstrap replicates are obtained. These replications are used to calculate the percentile confidence intervals.

3.1.2 Misspecification of the heaping mechanism

As is pointed out in Section 2.2, the heaping intervals should be specified properly by the user in order to describe the true duration distribution. The rule used for the interval of a certain duration as mentioned in the same section, is $[x-2, x+4)$. A consequence of misspecifying these interval boundaries is that the model may not identify to which interval someone has rounded his duration off.

For example, if an interval of $[x-3, x+3)$ is used instead of $[x-2, x+4)$, respondents with an observed duration 3, 4, 5, 7 and 8 are assumed to round off to duration 6. The respondents who have a duration of 10 months may now be located within the second heap. Furthermore, respondents who had reported a duration of 3 months are now located within the interval. It is now assumed that they may also round off their durations, while they didn't belong to the heaping interval before. As a consequence, the results for the table of unemployment categories (discussed in the introduction) could be biased. The totals of every category may change because some respondents may move to another category and some extra respondents are obtained from other categories.

Another simulation study is performed to evaluate whether misspecification of these heaping intervals will lead to biased estimates. For this simulation study, the specifications of the distribution used for the misspecification of the model parameters will be modified. Namely, the intervals boundaries for the heaps are shifted to the left and to the right. So, $[x-2, x+4)$ will become $[x, x+6)$ if it is shifted to the left. It is then assumed that persons tend to round their values downwards. The other misspecified interval is $[x-5, x+1)$, the right intervals are then shifted to the right. This means that it is assumed that people round their values upwards. The intervals used for the simulation study are presented in Table 1. This method will also be repeated in order to provide 1000 bootstrap replicates, which are used to calculate the percentile bootstrap confidence intervals.

Since the Weibull distribution is skewed the quantiles are appropriate to make causal inferences and therefore chosen as a measure for this analysis. The results of these simulation studies will be discussed in the following section. For both simulation studies, the quantiles of the estimated underlying model with and the imputation method are calculated in order to evaluate these two methods.

3.2 Analysis using the Dutch LFS

The purpose of this study was to find out whether it is more favourable to base the estimates on imputed data. Kuijvenhoven & Van der Laan estimated the unemployment duration using the imputation method based on true data, namely the Dutch LFS. In order to represent the difference based on true data, the unemployment durations will be predicted without making use of imputation based on the true data. This is done to examine whether there is a substantial difference between the estimates based

Table 1: Initial intervals for a certain heap and the intervals used in the second simulation study.

Heap (h_j)	interval length	[x-2, x+4)	[x, x+6)	[x-5, x+1)
6	6	4-9	1-6	6-11
12	6	10-15	7-2	12-17
12	12	7-18	4-15	9-2
18	6	16-21	13-18	18-23
24	6	22-27	19-24	24-29
24	12	19-30	16-27	21-32
30	6	28-33	25-30	30-35
36	6	34-39	31-36	36-41
36	12	31-42	28-39	33-44
42	6	40-45	37-42	42-47
48	6	46-51	43-48	48-53
48	12	43-54	40-51	45-56
60	12	55-66	52-63	57-68
72	12	67-78	34-75	69-80
⋮	⋮	⋮	⋮	⋮

directly on the estimated model and the estimates based on the imputed data. The duration based directly on the estimated model are generated by only modelling the underlying true distribution and the heaping mechanism. This process is done using a bootstrap method that takes the weights into account.

The bootstrap is a very general approach for calculating standard errors or confidence estimates of parameter estimates, or bias for sample statistic (Efron & Tibshirani, 1997). The bootstrap estimates are based on 200 resamples from the distribution as mentioned in section 3.1. The bootstrap algorithm used for estimating the standard errors is as follows:

- **step 1** The sample probability distribution $\hat{F}$ is constructed from the original dataset putting mass $1/n$ at each sampling point $x_1, x_2, \ldots, x_n$.
- **step 2** Draw a random sample with replacement of size n from

$$\hat{F} \rightarrow (x_1^*, x_2^*, \ldots, x_n^*). \tag{12}$$

- **step 3** From this bootstrap sample B independent samples $x^{*1}, x^{*2}, \ldots, x^{*B}$ each consisting of n data values are selected.
- **step 4** Approximate the statistic $\hat{\theta}(b = 1, 2, \ldots, B)$ for every bootstrap sample .
- **step 5** The standard error $se_F(\hat{\theta})$ is estimated by the sample standard deviation of the B replications

$$\hat{\sigma}_B^2 = \frac{1}{B-1} \sum_{b=1}^{B} \hat{\theta}_b^* - \hat{\theta}^*)^2. \tag{13}$$

where

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*. \tag{14}$$

Since the population $\hat{F}$ should correctly represent the Dutch population, Kuijvenhoven & Van der Laan use a specific method that enlarges the original sample into the population $\hat{F}$ using weighting. This method is described in the appendix .

The model was checked throughout for lack of convergence and no problems were detected.

4 Results

4.1 Results of the simulation study

4.1.1 Misspecification of the model parameters

In order to show that the imputation model is more robust for model misspecification compared to the model without imputation a simulation study is conducted. The results of this simulation study are summarized in Table 2. First, the true parameters are given (for α=0.8). Next, the parameter estimates based on the estimated underlying model are given whenever α is changed. Finally, the estimates of the quantiles based on imputed data whenever α is changed are given. The values within the table represent the value (duration) that is found for the quantiles. The value 3 for the true model means, that the first quantile of the distribution of x is 3.

When the method with and without imputation is correctly specified, the estimates of the quantiles for both methods are more or less equal to the true parameters. The true parameters fall within the 95% confidence intervals. So far, no substantial differences are found in the estimates of the two methods. Subsequently, the effect of misspecifying the models is investigated in the simulation study.

The results of the simulation study for the method with imputation are represented in Table 2 as well. These results show that after increasing the shape parameter α the values do not change significantly. The parameter estimates of the quantiles are very close to the true parameters. Furthermore, for every simulation, the true parameters of the quantiles fall within the corresponding interval of the estimated quantiles using imputed data. Although the data is discrete, the parameters based on imputed data are given in two decimals. This a result of multiple imputation, the estimates are namely based 25 imputation. Because the average of these imputations is calculated it is allowed that the values are continuous.

The results of the simulation study of the model based on the underlying true distribution show that after changing α a significantly increase for the first quantile, a significantly increase for the median and a significantly decrease for the third quantile is observed when α is set to 0.2, 0.3, 0.4, 0.5. For these estimates the true parameter does not lie within the confidence intervals. However when α is set to 0.7 or 0.8 no significant difference is found. Then, the true parameters fall within the confidence interval for all quantiles. In order to illustrate the effect of misspecification on the models with and without imputation, figures are shown for the first, second and third quantile in Figure 4.

The bold line represents the parameters of the true model and is used as a reference for the estimates of the two methods. The first plot represents the first quantile. The full line, which represents the estimates of the imputation models, cannot be detected

Table 2: Effect of misspecifying the shape α of the distribution: Estimated quantiles with confidence intervals of the estimated model and the imputed data

		Quantile		Quantile		Quantile	
	α	0.25	95% CI	0.50	95% CI	0.75	95% CI
True model	0.8	3	-	8	-	19	-
Estimated model	0.2	1	[a]	2	1-2	35	32-33
	0.3	1	[a]	3	[a]	24	21-22
	0.4	1	[a]	4	[a]	21	21-21
	0.5	1	[a]	5	5-5	20	20-20
	0.6	2	[a]	6	[a]	19	18-19
	0.7	2	2-3	7	7-8	19	18-19
	0.8	3	[a]	8	8-9	19	19-19
True model	0.8	3	-	8	-	19	-
Imputed data	0.2	3.00	[a]	8.12	8.00-8.92	19.00	18.00-20.00
	0.3	3.00	[a]	8.10	8.00-8.88	19.00	18.00-20.00
	0.4	3.00	[a]	8.12	8.00-9.00	19.00	18.00-20.00
	0.5	3.00	[a]	8.28	8.00-8.96	19.00	18.00-20.00
	0.6	3.00	[a]	8.14	8.00-8.96	19.04	18.00-20.00
	0.7	3.00	[a]	8.22	8.00-9.00	19.00	18.00-20.00
	0.8	3.00	[a]	8.34	8.00-9.00	19.00	18.00-20.00

[a] The bootstrap replications showed no variation.

because it lies on the bold line. This means that the estimates did not increase or decrease after changing the shape parameter of the Weibull distribution. This full line seemed to follow the bold line.

Conversely, the dotted line, which represents the estimates of the models without imputation, made substantial alterations. The line increased and decreased for every time the shape parameter is changed and seemed to follow the misspecified model.

Thus, the full line showed that the estimates of the quantile based on imputed data are not affected by the misspecification of the model parameters. The dashed line, showed a lot of alterations and means that the estimates based on the estimated model are highly affected by the misspecification of the model parameters. The same results were found for the other two plots, which represented the median and the third quantile. Therefore, from these results it can be concluded that the model with imputation is more robust for model misspecification while the model based on the estimated distribution was not robust for model misspecification.

4.1.2 Misspecification of the heaping mechanism

The results of the second simulation study represent the effect of misspecification of the heaping mechanism and are summarized in Table 3. The results show that there is no substantial difference between the performance of estimated underlying model and the imputation method. The estimates based on imputed data are not significantly different from the true estimates. The confidence intervals of the estimates contain the true parameter for all quantiles. The parameter estimates are somewhat overestimated when it is assumed that persons tend to round downwards (i.e. make use of backward

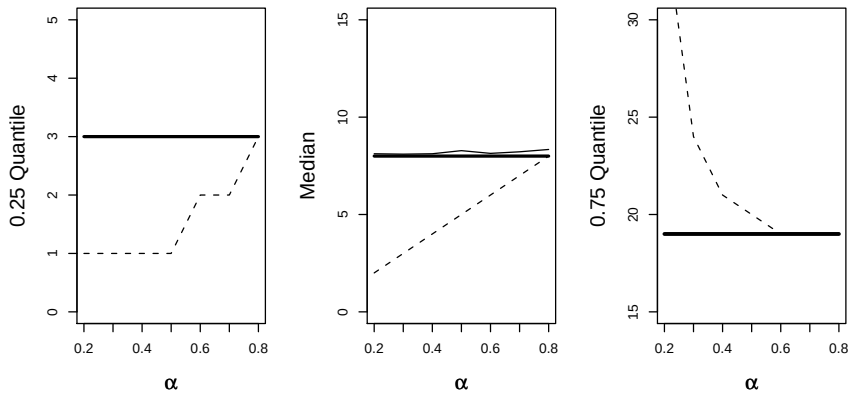

Figure 4: The effect of changing the shape parameter α from 0.2 to 0.8 for the three quantiles.

Table 3: Effect of misspecifying the heaping intervals: Estimated quantiles with confidence intervals of the estimated model and the imputed data

		Quantile		Quantile		Quantile	
	Interval	0.25	95% CI	0.50	95% CI	0.75	95% CI
True model		3	-	8	-	19.00	-
Estimated	Rounding up	3.00	2-3	8	7-8	18	17-19
model	Rounding down	3.00	[a]	9	8-9	20	19-21
Imputed	Rounding up	2.19	2.00-3.00	7.36	7.00-8.00	19.00	17.92-19.00
data	Rounding down	3.00	[a]	9.00	8.00-9.00	20.00	18.00-21.00

[a] The bootstrap replications showed no variation.

telescoping). This is caused by replacing the intervals to the right. Also somewhat underestimated results were found when it is assumed that persons tend to round upwards (i.e. make use of forward telescoping). Again, this is caused by replacing the intervals (in this case) to the left.

The estimated underlying model shows good results as well. The estimated parameters are not significantly different from the true parameters. The same results can be made here about the overestimated result of the parameters when it is assumed that persons tend to round down.

From these results it can be concluded that both methods are robust to misspecification of the heaping mechanism. It should be stated that the given confidence intervals indicate that the bootstrap did not always show variance. In a number of cases the variation was smaller than 5 % and in other cases no variation was found at all.

4.2 Analysis of true data

The durations generated from the estimated underlying model are given in Table 4. This table represents the number of persons that belong to each unemployment category according to the estimates directly on the underlying model and the estimates based on the imputed data. The estimate of the bootstrap standard errors are represented as well and are based on 200 resamples. Statistics Netherlands uses complex data processing, (i.e. registers, weights) which makes it time consuming to increase the number of resamples to 1000. No confidence intervals could therefore be calculated to examine the difference between the estimates. Nevertheless, in order to obtain a rough idea of the difference between the estimates, the standard error is used. These standard errors seem to indicate that the estimates of the unemployment duration do not differ significantly.

Table 4: Number of unemployed persons for different duration categories based on the estimated true distribution and on the imputed data. The standard errors are shown between parenthesis.

Year	Duration (months)	Estimated model $\times$ 1000	Imputation model $\times$ 1000
2002	0-5	150(3.9)	152.5(5.8)
	6-11	56(2.1)	54.6(3.1)
	12-23	32.5(2.2)	33.9(2.7)
	24+	48.5(3.1)	46.9(3.3)
2003	0-5	180.6(4.1)	184.3(4.2)
	6-11	84.5(2.3)	86.9(3.2)
	12-23	59.2(2.8)	57.8(3.0)
	24+	61.4(2.9)	58.3(3.6)
2004	0-5	188.9(3.7)	190.3(5.4)
	6-11	107.0(2.4)	103.6(3.8)
	12-23	94.4(3.1)	98.8(4.0)
	24+	74.6(3.1)	72.0(3.5)
2005	0-5	176.9(4.1)	177.3(5.0)
	6-11	101.1(2.5)	101.6(3.9)
	12-23	93.9(3.7)	94.2(4.4)
	24+	96.9(3.2)	95.8(3.8)
2006	0-5	148.7(4.0)	149.0(4.9)
	6-11	75.0(2.2)	73.0(3.2)
	12-23	70.8(3.5)	76.0(4.0)
	24+	104.3(3.4)	100.8(4.2)
2007	0-5	128.5(3.6)	129.5(4.6)
	6-11	60.8(2.2)	57.7(3.2)
	12-23	52.2(2.5)	57.4(3.3)
	24+	85.1(3.5)	82.1(4.0)

The reason for the lack of difference in the estimates of the two models is probably due to the fact that the estimated model is correctly specified.

5 Concluding remarks

The goal of this paper was to evaluate the method used by Kuijvenhoven & Van der Laan to analyze heaped data and test whether it is more favorable to base the estimates of the unemployment durations on imputed data.

The first simulation study showed that when a model parameter is misspecified the imputation method is more robust compared to the estimated underlying model. Making use of imputation resulted in more precise and consistent estimates. It can therefore be concluded that the method with imputation performs better when a model parameter is misspecified compared to the estimated underlying model.

The second simulation study showed that when the heaping intervals are misspecified, both the imputation method and the underlying true model perform well. There is no clear conclusion about which method works best. Misspecifying the heaping mechanism did not have a substantial influence on the parameter estimates and therefore both methods are robust to this form of misspecification. This means that the assumptions forward and backward telescoping can both be made when the heaping mechanism is specified. Changing the direction of rounding did not influence the estimates of the durations.

Overall, the imputation method works better because it is robust to the misspecification of the model and to the misspecification of the heaping intervals as well. As such, the practical conclusion from this analysis is that multiple imputation is necessary to avoid bias caused by misspecification of the model.

Using misspecification in the model and in the heaping mechanism is a flexible way to analyze heaped unemployment data. This novel method could also be used to data that is distorted by another kind of heaping (e.g. age heaping) because the user is free to specify the heaping intervals and model parameters.

Another way to test the imputation method is to use another underlying true model for the heaped data. Using splines instead of using the stepwise geometric model. In this way, it can be tested whether the use of imputation is dependent on the method that is used for the underlying true model.

References

Chambers, R., & Skinner, C. (2003). *Analysis of survey data.* Chicester: Wiley.

Heijtjan, D., & Rubin, D. (1990). Inference from coarse data via multiple imputation with application to age heaping. *American Statistical Association*, *85 (410)*, 304-314.

Klein, J. P., & Moeschberger, M. (2003). *Survival analysis: Techniques for censored and truncated data.* Springer.

Kraus, F., & Steiner, V. (1993). Modelling heaping effects in unemployment duration models: With an application to retrospective event data in the german soci0-economic panel. *Jahrbiicher fur NationalSkonomie und Statistik*, *59 (1-2)*, 187-211.

Kuijvenhoven, L., & Laan, J. Van der. (2009). *Een methode voor het corrigeren van afgeronde werkloosheidsduren.* (Rapport technique). Centraal Bureau voor de Statistiek.

Lynn, P. (2009). *Methodology of longitudinal survey.* John Wiley & Sons.

Petoussis, K., Gill, R., & Zeelenberg, C. (s. d.). *Statistical analysis of heaped duration data.* (Retrieved September 2009, from http://sciencestage.com/d/1103321/statistical-analysis-of-heaped-duration-data-(2004).html)

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63(3)*, 581-592.

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

Wolff, J., & Augustin, T. (2000). *Heaping and it's consequences for duration analysis.* (Rapport technique). Ludwig-Maximilians Universität München.

Appendix

Construction of the artificial population for the bootstrap

In order to construct the population $\hat{F}$, the sample s should be enlarged. This done by making copies of the sample elements using the weights. For every $k \in s$, w_k copies are formed from the k^{th} sample element. The sum of all copies made for every k^{th} sample element of every sample is then equal to the artificial population $\hat{F}$.

Once the population is obtained, bootstrap samples should be taken. This is done using a sample design that should be identical to sample design of the original sample. Samples are taken without replacement using unequal selection probabilities and a sample size that is equal to s. The selection probability refers to the probability of a person being selected in a sample and is defined for every sample element as the reciprocal of its weight w_k. Subsequently, the estimate of unemployment duration described in (subsection 2.4) can be calculated.

It should be addressed that the sample s is now equal to the entire labour force of a certain sample year. However, this paper concerns the unemployed persons and restriction should be made. Since the population $\hat{F}$ is based on copies of s, the obtained bootstrap sample should reduce to the unemployed persons. The size of the bootstrap sample on which the estimates are based, is therefore smaller than the bootstrap sample obtained from the process before in this appendix. Furthermore, this bootstrap sample size is now stochastic.

A consequence might be that the bootstrap sample size is not identical to the correct size of the population. The weights should then be computed again. Unfortunately, there is no information available to do this. However, the effect of this recalculation is expected to be small.