Imputation of rounded data

Jan van der Laan and Léander Kuijvenhoven

Discussion paper (201108)

Statistics Netherlands

The Hague/Heerlen, 2011

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201108 X-10

Imputation of rounded data

Jan van der Laan and Léander Kuijvenhoven

Abstract: In surveys persons have a tendency to round their answers. For example, in the Labour Force Survey people are asked about the period they have been unemployed. There is clearly a tendency to give answers that are rounded to years or half years. Because of this rounding statistics based on this data tend to be biased. In this paper we introduce a method with which the rounding mechanism is modelled together with the 'true' underlying distribution. These are then used to select samples which are likely to be rounded and impute new values for these. This method is applied to the Labour Force Survey data. An investigation of robustness shows that the method is robust against misspecification of the model of the underlying distribution and to misspecification of the rounding mechanism.

Keywords: rounding, multiple imputation, Labour Force Survey, unemployment spells

1 Introduction

It is very common in surveys to be confronted by measurement errors. One type of measurement errors is caused by rounding: persons round their answers to certain 'round' values. For example, in the Dutch Labour Force Survey where persons are asked about the duration they have been unemployed, persons tend to round their answers to years or half years. Other examples are income where persons round their answers to multiples of 100 or 1,000. When one is interested in statistics such as the average income, rounding is usually not a problem. However, as rounding distorts the distribution of the variable of interest, estimates can be biased when these are dependent on the shape of the distribution. This can happen for example when performing regression (Wolff and Augustin, 2000; Augustin and Wolff, 2004), or when the shape of the distribution itself is the statistic of interest.

At Statistics Netherlands there is an interest in publishing statistics on the distribution of unemployment durations. It is for example of interest to see whether policies affect mainly persons with long unemployment spells or mainly persons with short unemployment spells. Figure 1 shows the distribution of reported unemployment spells at the moment of interviewing. Heaps are clearly visible at multiples of six months with higher heaps at multiples of twelve months. This can not be a real phenomenon as the interviewing moments are distributed (practically) randomly throughout the year. The heaps are therefore caused by memory effects: persons cannot precisely remember the exact duration and round their answer to the nearest half or full year. This effect is also

seen in other labour force surveys such as the German Socio-Economic Panel (Kraus and Steiner, 1995) and the Italian Labour Force Survey (Torelli and Trivellato, 1993). As these heaps cause a bias in the estimates based on this data, an estimation method that is able to deal with this type of data is needed.

Rounding, which in literature is often called heaping, is a special form of data coarsening, and while the amount of literature on the subject is limited, the problem has received some attention in the last twenty years. Especially Wang and Heitjan (2008) list a large amount of literature on this subject. However, most methods in the literature either assume that it is known which persons have rounded and which persons have not (Torelli and Trivellato, 1993), or focus on the effect of rounding on the estimated effect of covariates either by adapting the estimation method to obtain unbiased estimates or by estimating the bias introduced by rounding (Augustin and Wolff, 2004; Wolff and Augustin, 2000). In case of continuous data the assumption that it is known when a person has rounded is not unrealistic, as observing a specific value such as 1,000 is highly improbable unless this value is observed because a person has rounded. As the interest lies not in estimating the effect of covariates on the distribution but more in giving a description of the distribution itself, most methods discussed in literature are not directly applicable.

As we are mainly interested in the distribution of the durations, we use multiple imputation (Little and Rubin, 2002) to correct the data for the rounding. We first create a model that describes the data. This model consists of two parts. First, a model for the rounding process, and second, a model for the true underlying distribution. In order to have the choice of the model for the true underlying distribution to have as little effect as possible on the results, we choose a semi-parametric model for this. After the model has been fitted to the data, it is then used to randomly select persons that are likely to have rounded and impute new values for these persons. Multiple imputation is used for two reasons. First, this enables us to obtain variance estimates for estimates based on the imputed data. Second, as we are using a stochastic imputation method, more precise estimates can be obtained by averaging the estimates obtained from the multiple imputed datasets. An advantage of using imputation to correct the data, is that this is less sensitive for model misspecification. Only for those persons which are likely to have rounded new durations are imputed. The remainder of the data is left as observed. Any model misspecification will, therefore, only affect part of the data.

The model and imputation method are described in the next section. In section 3 the method is applied to the unemployment spell lengths obtained from the Labour Force Survey. In section 4 we investigate the robustness of the method for any misspecification in the model.

2 Model and imputation method

2.1 General heaping model

It is assumed that every person i has a true value of the variable of interest y_i that is not directly observed. Instead we observe the reported value z_i, which depends on the true value. Suppose there are k values h_j $(j = 1, \ldots, k)$ to which persons round

their reported values. We will call these values the heaps and we will assume that the number and locations of these heaps are known. Let r_{ij} indicate rounding by person i to heap j. Note that r_{ij} is not observed. The probability of rounding to heap j depends on the true value y_i: $\Pr(r_{ij}|y_i)$. For example, unemployment durations are more frequently rounded as the durations get longer.

The rounding process can be modelled as a single draw from a multinomial distribution with $k+1$ categories. Every y_i can be rounded with a probability $\Pr(r_{ij}|y_i)$ to one of the k heaps h_j and with a probability of $1-\sum_{j=1}^{k}\Pr(r_{ij}|y_i)$ not be rounded. When z_i is not on one of the heaps, it is known that z_i is not rounded and the probability of observing z_i is equal to the probability of the true value y_i being equal to z_i times the probability of y_i not being rounded. Therefore, the density function $g(.)$ of z_i is given by

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

where $f(.)$ is the density function of y_i. The goal is to retrieve the density function $f(.)$ of the unobserved variable of interest y_i from the observed z_i.

When z_i is on one of the heaps there are two possibilities. Either z_i is not rounded and therefore equal to y_i, or z_i is rounded. The probability of observing z_i in the first case is given by the previous equation. The probability of observing z_i in the second case is given by

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

where we sum over every possible h_j equal to z_i. Combining equations (1) and (2), the density function $g(.)$ of z_i is given by

$$g(z_i)=f(z_i)\left(1-\sum_{j=1}^{k}\Pr(r_{ij}|z_i)\right)+\sum_{j:z_i=h_j}\int_{-\infty}^{\infty}\Pr(r_{ij}|y)\,f(y)\mathrm{d}y \tag{3}$$

2.1.1 Estimation

In order to estimate the model, it is assumed that $f(.)$ and $\Pr(r_{ij}|z_i)$ follow known distributions parametrised by parameter vectors θ and ϕ respectively:

$$g(z_i|\theta,\phi)=f(z_i|\theta)\left(1-\sum_{j=1}^{k}\Pr(r_{ij}|z_i,\phi)\right)+\sum_{j:z_i=h_j}\int_{-\infty}^{\infty}\Pr(r_{ij}|y,\phi)\,f(y|\theta)\mathrm{d}y. \tag{4}$$

If we assume to have independent observations, the log-likelihood of observing the n values z_i ($z=(z_1,z_2,\ldots,z_n)'$) is given by

$$l(\theta,\phi|z)=\sum_{i=1}^{n}\ln g(z_i|\theta,\phi), \tag{5}$$

and the estimate $(\hat{\theta},\hat{\phi})'$ is found by maximising $l(\theta,\phi|z)$:

$$(\hat{\theta},\hat{\phi})'=\underset{(\theta,\phi)'}{\arg\max}\,l(\theta,\phi|z). \tag{6}$$

Optimization was done in the statistical program R using the Nelder-Mead (simplex) algorithm. This algorithm does not need derivatives. The integrals present in equation 4 can be evaluated after a particular choice of $\Pr(r_{ij}|y,\phi)$ and $f(y|\theta)$ (see sections 2.2 and 2.3).

For the application to the unemployment durations from the Labour Force Survey it is necessary to include weights in the model to compensate for under- or overrepresentation of certain groups. This was done using a weighted likelihood:

$$l_w(\theta,\phi|z) = \sum_{i=1}^{n} w_i \ln g(z_i|\theta,\phi), \tag{7}$$

where w_i is the weight of to the i^{th} observation.

2.2 Uniform heaping

In the previous section a general model was described with which data with rounding errors can be modelled. In order to estimate this model it is necessary to assume a model for the rounding mechanism given by $\Pr(r_{ij}|z_i,\phi)$. One possibility is to assume a constant rounding probability p_j of rounding to heap h_j inside the interval $I_j = [s_j, t_j)$. This type of rounding covers many of the commonly found types of rounding, such as true rounding (e.g. rounding to the nearest integer) and rounding down as is commonly done with age.

In this case $\Pr(r_{ij}|z_i,\phi)$ equals zero for z_i outside the interval I_j. Therefore, in equation (3) there only needs to be summed over the intervals into which z_i falls and the limits of the integral can be replaced by s_j and t_j. This leads to the following density function for the observed values

$$\begin{aligned} g(z_i|\theta,\phi) = f(z_i|\theta)\left(1 - \sum_{j:z_i\in I_j} p_j(\phi)\right) + \sum_{j:z_i=h_j} p_j(\phi)\int_{-s_j}^{t_j} f(y|\theta)\mathrm{d}y = \\ f(z_i|\theta)\left(1 - \sum_{j:z_i\in I_j} p_j(\phi)\right) + \sum_{j:z_i=h_j} p_j(\phi)\left(F(t_j|\theta) - F(s_j|\theta)\right). \end{aligned} \tag{8}$$

The heaping or rounding model is completely described by the rounding probabilities, the heaps (which we already assumed to be known) and the rounding intervals. We will also assume that the rounding intervals are given. Therefore, only the rounding probabilities need to be estimated. In principle, therefore, the parameter vector ϕ could just consist of the k rounding probabilities. However, in order to ensure that the estimated probabilities are between zero and one, we use the following transform

$$p_j = \exp\left(-\exp\left(\phi_j\right)\right), \tag{9}$$

with $\phi = (\phi_1, \phi_2, \ldots, \phi_k)'$.

2.3 Model for the underlying distribution

For the underlying distribution we need a distribution that is capable of describing the highly skewed distribution of our data (see figure 1). From the area of survival analysis (which deals with durations such as unemployment durations) there are plenty of distributions to choose from (Marshall and Olkin, 2007; Lawless, 1982). However, in order to keep the distribution as general as possible, we choose the piecewise constant hazard model.

In our case we are dealing with discrete data as the durations are measured in months. In the discrete case the hazard is the possibility that the duration is ended given that until now the duration was not ended.

$$\lambda(t) = \Pr(y = t | y \geq t) \tag{10}$$

The distribution function follows directly from this hazard

$$f(y) = \lambda(y) \prod_{t=0}^{y-1} (1 - \lambda(t)) \tag{11}$$

In case of a piecewise constant hazard model we assume that the duration axis is divided into l intervals defined by the l times $0 = u_1 < u_2 < \ldots < u_l < \infty$. In each of these intervals we assume that the hazard is constant: $\lambda(t) = \lambda_m$ if $u_m \leq t < u_{m+1}$. The geometric distribution is a special case of this distribution for when there is only one interval.

As it is reasonable to expect that the distribution will differ for different subgroups in our population, we also want to incorporate the background properties x_i of individuals into the distribution. We have decided to use the following parametrisation

$$\lambda_{im} = \exp\left(-\exp\left(\alpha_m + \beta' x_i\right)\right), \tag{12}$$

where α_m is the base hazard that is scaled using the covariates vector x_i and parameter vector β. Therefore, the distribution is described by l plus the number of covariates parameters, and $\theta = ((\alpha_1, \ldots, \alpha_l)', \beta)'$.

2.4 Imputation

In the previous sections a model has been described that is capable of describing data that is distorted by rounding. The goal is to use this model to impute new values for some of the values that are on the location of a heap. The reason for this is that as only a part of the data exhibits rounding, only a part of the data is imputed. Therefore, a large part of the data is left untouched, causing problems with misspecification of the model to only influence part of the data. Therefore, statistics based on imputed data will be more robust against misspecification of the model than statistisc based directly on the model.

The imputation algorithm consists of a number of steps:

1. Determine for each observation and each heap the probability that the observation has been rounded to that heap. This probability is only non-zero when the observation

is on the heap. When observation i is located on heap j this probability is given by

$$\hat{f_{ij}} = \frac{p_j(\hat{\phi})\left(F(t_j|\hat{\theta},x_i) - F(s_j|\hat{\theta},x_i)\right)}{g(h_j|\hat{\theta},\hat{\phi})}. \tag{13}$$

2. For each heap the probability of belonging to that heap is known and also the probability of not belonging on a heap is known which is simply $1-\sum_{i=1}^{k}\hat{f_{ij}}$. Using a single draw from a multinomial distribution, it is randomly determined if the observation belongs to a heap and if so to which heap it belongs to.

3. If the observation does not belong to one of the heaps it is not imputed. Otherwise, a new value $\tilde{y}_i$ is drawn from

$$\tilde{y}_i \sim f(x_i|\hat{\theta}, r_{ij}). \tag{14}$$

It should be noted that this distribution is conditional on the fact that the observation is on heap j. In case of uniform rounding this means that the new value can only be between t_j and s_j as only observations from this interval can be on heap j.

The imputation algorithm discussed above is repeated multiple times generating multiple imputed datasets. Any statistics of interest can be calculated for each of the imputed datasets. Averaging these estimates creates a more accurate estimate as part of the uncertainty introduced by the stochastic imputation is averaged out. Furthermore, these multiple estimates can be used to estimate the uncertainty introduced by the imputation (Little and Rubin, 2002).

3 Application

Our method has been applied to unemployment durations obtained in the Dutch Labour Force Survey. In this survey unemployed persons are asked among other things how long they have been unemployed[1]. When answering this question they tend to round their answers to multiples of six months. Figure 1 shows the distribution of unemployment durations for 2009. Peaks are clearly visible at multiples of six months. The peaks at full years are more pronounced than those at half years.

The previously introduced model has been fitted to this data. As the peaks at full year locations are more pronounced, it looks like persons do not only round their duration to the nearest half year, but also round their duration to the nearest year, Thus, a person with for example a duration of 2.7 years, can both round his duration to 2.5 years and 3 years. We have therefore decided to introduce two types of rounding intervals: intervals of 6 months with heaps at six month intervals and interval of 12 months with heaps at 12 month intervals. It is also assumed that probability of rounding is constant after 72 months. The heap locations and intervals are shown in table 1.

For the underlying distribution the stepwise hazard model is used as was discussed in section 2.3. The duration axis was divided into nine intervals with borders u_l at 0, 2, 4, 10, 15, 20, 35, 50, 100 and 200 months. The intervals were chosen smaller for short durations as more data was available for these intervals because of the skew

[1] Actually they are not asked directly. Unemployment status and unemployment duration are derived from a set of answers to other questions.

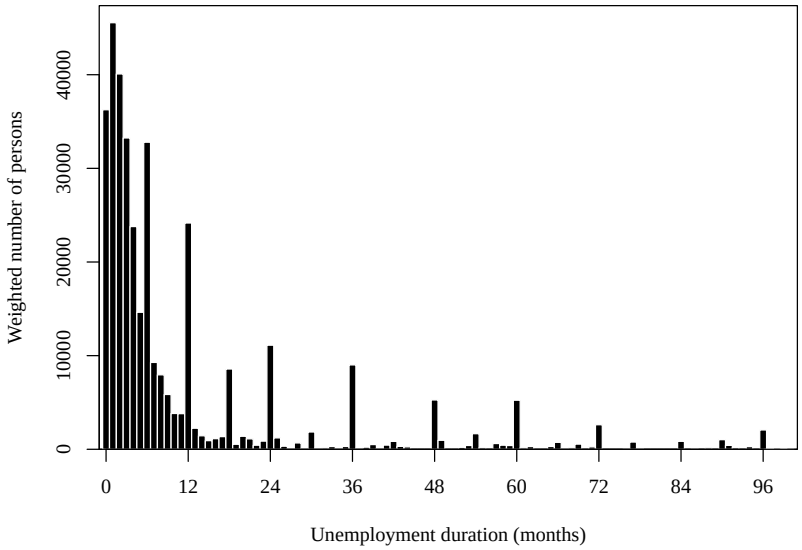

FIGURE 1 *Weighted histogram of the observed unemployment durations. Original data before imputation.*

TABLE 1 *Heap locations and intervals used for the LFS-data. After 72 months the rounding probabilities are assumed to be constant.*

Heap location	Heaping interval	$\hat{p}_j$	Confidence Interval
6 month rounding			
6	4–9	0.18	(0.13–0.23)
12	10–15	0.29	(0.00–0.52)
18	16–21	0.38	(0.24–0.52)
24	22–27	0.37	(0.10–0.65)
30	28–33	0.18	(0.00–0.35)
36	34–39	0.00	(0.00–0.32)
42	40–45	0.14	(0.00–0.30)
48	46–49	0.41	(0.00–0.85)
12 month rounding			
12	7–18	0.18	(0.05–0.31)
24	19–30	0.29	(0.09–0.45)
36	31–42	0.67	(0.47–0.79)
48	43–54	0.34	(0.00–0.67)
60	55–66	0.59	(0.37–0.79)
72	67–78	0.54	(0.38–0.69)
⋮	⋮	⋮	⋮

TABLE 2 *The estimated parameters for the underlying distribution. The parameters for the baseline hazard have been transformed back to probabilities for easier interpretation. Covariates marked with '*' are significant at the 95% level.*

Parameter	Estimate	Confidence Interval
Baseline hazard $\exp(-\exp(\hat{\alpha}_m))$		
0–2	0.16	(0.14– 0.18)*
2–4	0.16	(0.14– 0.18)*
4–10	0.13	(0.11– 0.15)*
10–15	0.08	(0.06– 0.10)*
15–20	0.08	(0.05– 0.09)*
20–35	0.06	(0.05– 0.07)*
35–50	0.05	(0.04– 0.07)*
50–100	0.06	(0.04– 0.07)*
100–200	0.04	(0.02– 0.05)*
200–∞	0.04	(0.03– 0.06)*
Covariates $\hat{\beta}$		
Ethnicity = Dutch		–
Ethnicity = Western	0.01	(-0.05– 0.05)
Ethnicity = Non-western	0.03	(-0.03– 0.07)
Gender = Male		–
Gender = Female	0.02	(-0.02– 0.06)
Age = 14–24		–
Age = 25–34	0.12	(0.05– 0.17)*
Age = 35–44	0.19	(0.12– 0.25)*
Age = 45–54	0.30	(0.23– 0.35)*
Age = 55–64	0.36	(0.29– 0.42)*
Education level = low		–
Education level = middle	−0.06	(-0.10–-0.01)*
Education level = higher	−0.04	(-0.09– 0.01)

distribution and also because the hazard shows more variation in these intervals. For the covariates we choose the following background properties: ethnicity, gender, age and education level. These were chosen because these are also used in the published tables. All variables are categorical variables. The categories are shown in table 2.

Tables 1 and 2 show the estimated parameters for the rounding process and the underlying distribution respectively. The confidence intervals of the parameters were estimated using the percentile method (Efron and Tibshirani, 1998) from a weighted bootstrap (Särndal et al., 1992) using 1000 replicates. The parameters for the rounding probabilities $\hat{\phi}_j$ and the parameters for the baseline hazard α_m were transformed back to probabilities to ease understanding. As for the parameters for the covariates β, a positive value leads to a lower hazard and therefore longer durations and a negative value leads to a higher value of the hazard and therefore to shorter durations.

For most durations persons can round both to the nearest full year heap or to the nearest half year heap. Figure 2 shows the total rounding probability: for every duration all rounding probabilities have been added. From the figure can bee seen that the probability of rounding increases the first two years until it is about constant at 60%.

From the figure and also from table 1 it can be seen that the uncertainty in the rounding probabilities is quite large. In fact many of the confidence intervals include zero. However, from this fact we can not conclude that there is no evidence for rounding,

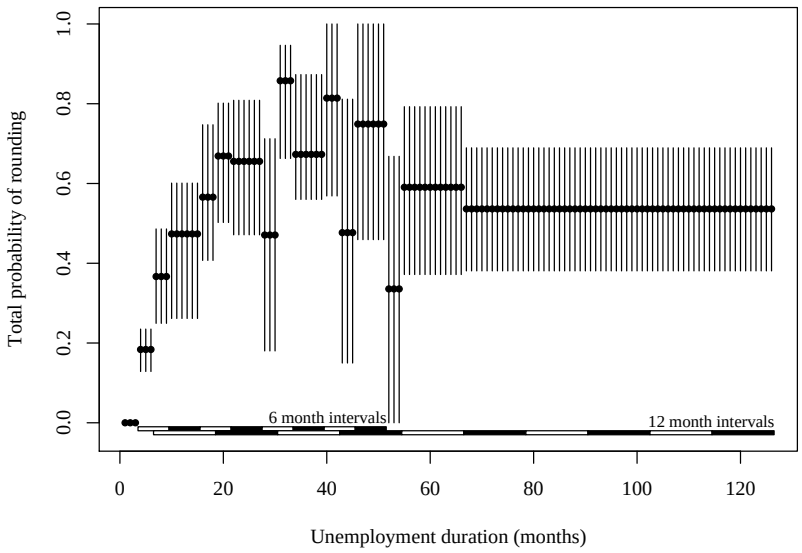

FIGURE 2 *Total probability of rounding as a function of duration. The 95% confidence intervals are indicated by vertical lines. The rounding intervals are shown in the bottom of the figure.*

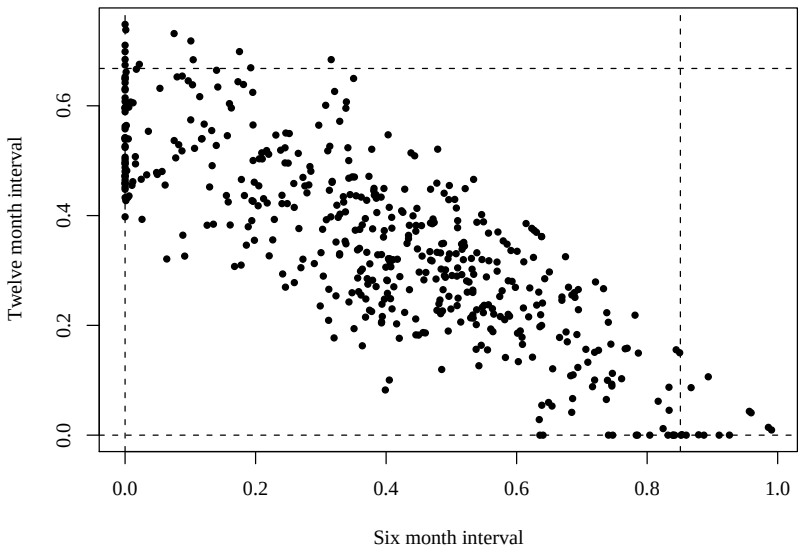

FIGURE 3 *Bootstrap estimates of the heaping probability for the two heaps at 48 months. The heaping probabilities of the heap with the twelve month interval are plotted against those of the heap with the six month interval. The dotted lines show the confidence intervals for the two heaping probabilities.*

as the rounding probabilities for rounding to full year heaps and the probabilities for rounding to half year heaps are strongly correlated. This can be seen in figure 3 where the bootstrap estimates of rounding to the full year peak at 48 months are plotted against the bootstrap estimates for rounding to the half year heap at 48 months. It is clearly visible that these estimates are strongly correlated. The fact that the rounding is significant could also be seen in figure 2, where in general zero is not included in the confidence interval for the total rounding probability. The fact that many of the estimates in table 1 include zero, indicates that the algorithm has difficulty separating full and half year rounding.

From the parameter estimates for the underlying distribution for the unemployment durations it can bee seen that the hazard decreases as the duration increases. This is to be expected as leaving an unemployment situation becomes more difficult as the unemployment duration increases. As for the covariates in the model, only age and education level are significant. Especially age has a strong effect: the durations increase as age increases. For older persons it is more difficult to leave an unemployment situation. The unemployment durations are shorter for persons with middle education level.

Although some of the parameters are not significant, we kept all of the parameters in the model. There are a couple of reasons for this. First, we want to use the same imputation model and method for all years and do not know if there will be changes in the effect of covariates in the future. Second, we do not use this model for inference but for imputation. Non-significant parameters in the model therefore only add to the imputation variance, which is not a problem as long as the imputation variance is small compared to the overall variance.

The method described in section 2.4 was used to impute the data. The resulting distribution of durations is shown in figure 4. This distribution was obtained by averaging the 25 distributions obtained from the multiple imputation. The heaps present in figure 1 have disappeared. The only heap that can be discerned namely at 54 months was not present in our model.

4 Investigation of robustness

In order to investigate the robustness of our method to misspecification of the underlying duration model or to misspecification of the heaping mechanism, simulations have been performed. In these simulations data is generated from a known distribution to which rounding is then applied. Our method is then applied to this heaped data. This is repeated to obtain confidence intervals for the estimates. This work is a continuation of the work done by El Messlaki (2010).

One of the main reasons imputation was used to obtain the estimates, is that we suspect that imputation is less sensitive to misspecification. Only a fraction of the data is rounded and only for this fraction of the data new values are imputed. Therefore, depending on the fraction of rounding any misspecification of the model has only a limited influence on the estimate, while estimates based solely on the estimated model will be strongly affected by the misspecification. In order to investigate if this is in-

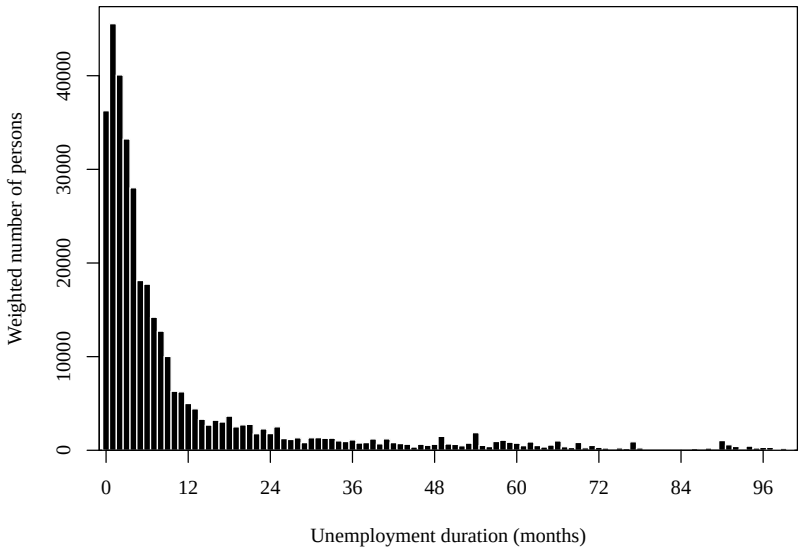

FIGURE 4 *Weighted histogram of the observed unemployment durations after imputation*

deed the case, estimates based on the model will be compared to those of the imputed dataset.

The statistics that were used to compare the estimated and imputed distributions to the true distribution, were the fractions of persons with durations between 0 and 6, 6 and 12, 12 and 24, and longer than 24.

For the true underlying distribution we used a discrete Weibull distribution with a shape parameter of 0.8 and a scale parameter of 12. This distribution was chosen because it has a similar shape as that of the unemployment durations. From this distribution 1,000 samples were drawn. For the half year rounding a rounding probability of 0.2 was used, for the full year rounding a rounding probability of 0.3 was used except for the first year where a probability of 0.1 was used. The simulation was repeated 1,000 times. The lower and upper bounds of the confidence intervals were determined by taking the 2.5 and 97.5 percentiles respectively.

4.1 Misspecification of underlying model

In order to introduce misspecification in the underlying model, the shape parameter of the estimated model is fixed to different values. Therefore, only the scale parameter was estimated. The shape parameter is changed in steps of 0.1 from 0.1 to 1.7.

Figure 5 shows each of the statistics as a function of the fixed shape parameter for the estimated model (dashed line with solid points) and for the imputed dataset (dotted line with open points). The confidence intervals are indicated using the grey areas. The statistics based on the imputed dataset are practically constant and almost equal to the true value, while the estimates based directly on the estimated model vary strongly as a function of shape and are also in most cases significantly different from the true

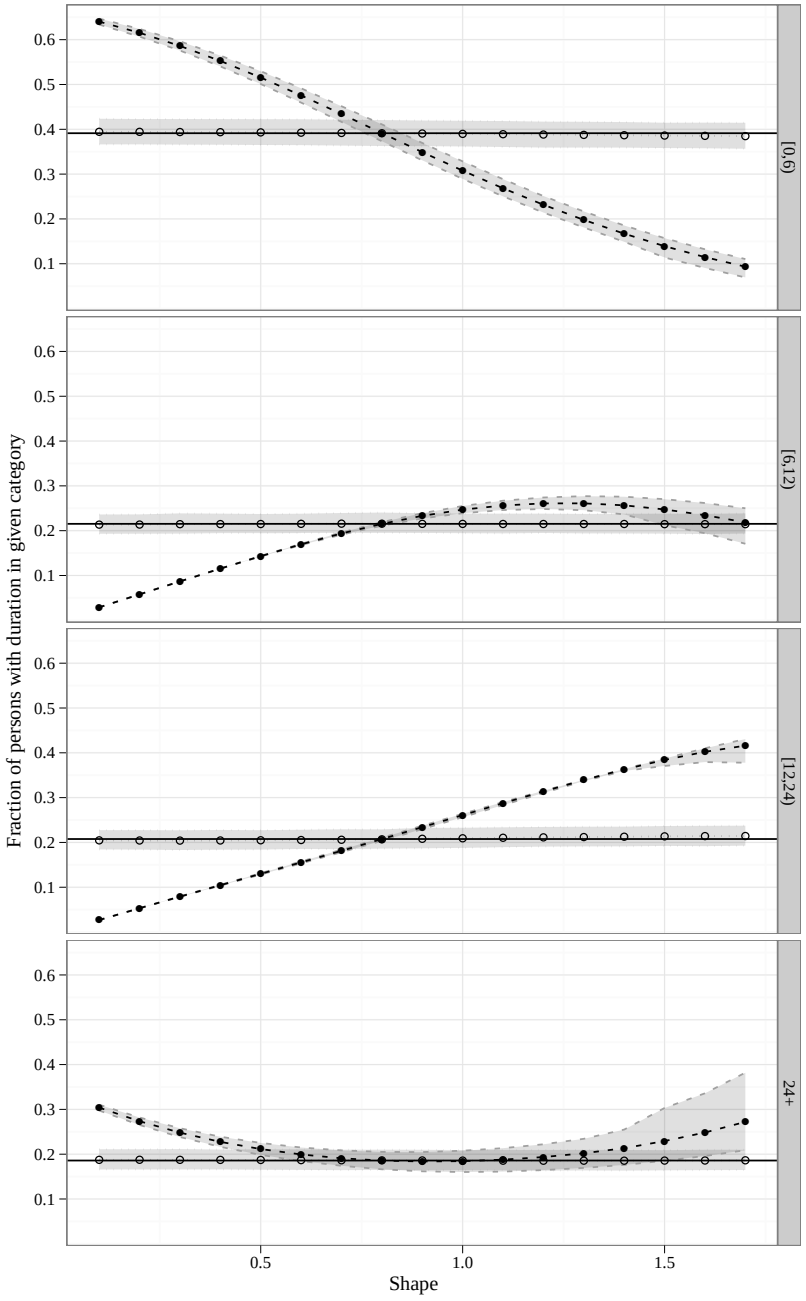

FIGURE 5 *Fraction of persons in each of the categories estimated directly from the model (dashed line with solid points) and estimated from the imputed dataset (dotted line with open points) as a function of the fixed shape. The true values are shown by the horizontal solid lines.*

value. The estimates based directly on the estimated model are therefore sensitive to misspecification while the estimates based on the imputed dataset are insensitive to misspecification of the underlying distribution.

4.2 Misspecification of rounding model

Previously it was assumed that the rounding intervals were centered around the heaps: persons round to the nearest half year or to the nearest full year. In order to investigate the sensitivity of the results to this assumption, the rounding intervals are shifted to the left and right. The intervals used when generating the data are kept centered around the heaps. When the intervals are shifted to the left, persons have a tendency to round their durations up; when the intervals are shifted to the right, persons have a tendency to round their durations down. Since we want to have the heaps inside the intervals, the maximum shift to the left is three (-3) and the maximum shift to the right is two (2).

Figure 6 shows the fraction of persons in each of the four categories as a function of the shift estimated from the model (dashed line with solid points) and estimated from the imputed dataset (dotted line with open points). The difference between the estimates based on the model and those based on the imputed dataset are negligible. Only when the rounding intervals are shifted three months to the left (persons round mainly up) is there a significant difference between the estimates and the true value.

The sign of the bias introduced by the misspecification is as would be expected. A negative shift causes the fraction of persons in the first category to increase since persons in the heap at six months will have rounded mainly up. Therefore, the imputed values will be mainly smaller than six months.

In case of misspecification in the underlying true distribution there was a clear difference between estimates based directly on the model and estimates based on the imputed dataset. In both cases the bias is small and only significant when assuming the answers are rounded mainly up (which is not a realistic assumption).

5 Conclusion

In the previous sections a method has been introduced with which rounding errors can be corrected. This method assumes that there is a true underlying distribution of the variable of interest. However, this underlying distribution is not observed. Instead a distribution distorted by a rounding mechanism is observed. By assuming a model of both the underlying distribution and the heaping mechanism, it is possible to estimate both the parameters of the underlying distribution and the parameters of the heaping mechanism. These parameters can then be used to (randomly) select values that are probably rounded and impute now values for these from the underlying distribution. The distortion introduced by the rounding is thereby removed from the dataset and estimates based on this imputed dataset are unbiased.

One of the main reasons to use multiple imputation and not base the estimates directly on the estimated model, is that it was expected that using imputation is more robust to

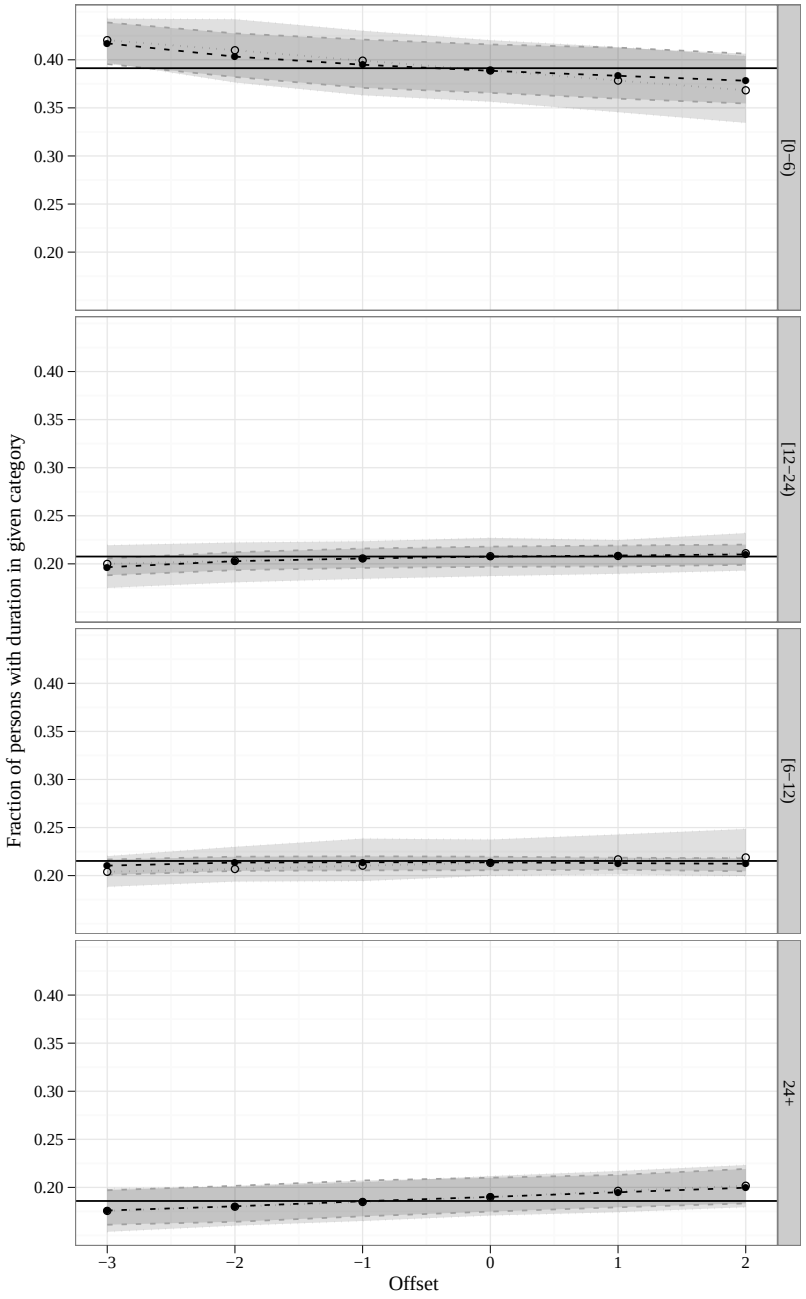

FIGURE 6 *Fraction of persons in each of the categories estimated directly from the model (dashed line with solid points) and estimated from the imputed dataset (dotted line with open points) as a function of the shift in the heaping intervals. The true values are shown by the horizontal solid lines.*

model misspecification. In section 4 the robustness of the method to model misspecification was investigated. The method is seems to be robust to both misspecification of the underlying true distribution and to misspecification of the heaping mechanism. Especially, for misspecification of the underlying distribution the differences between the estimates based on the model and those based on the imputed dataset are large. The estimates based on imputation are hardly affected by the misspecification, while those based directly on the model are strongly affected. It is therefore preferable to use multiple imputation and not base the estimates directly on the model.

The method has been applied to unemployment durations obtained from the Dutch Labour Force Survey (LFS). In the LFS persons have a tendency to round their durations to the nearest half year or full year causing a bias in estimates based on these values. After applying the method to this data the heaps were removed from the data. Although this method has been applied only to unemployment durations, there is no reason why this method would not be equally applicable to other data with rounding present. The only assumptions are that the data is discrete and that the observed distribution can be modelled as an true underlying distribution to which a heaping mechanism is applied.

References

Augustin, T. and Wolff, J. (2004), 'A bias analysis of Weibull models under heaped data', *Statistical Papers* **45**, pp. 211–229.

Efron, B. and Tibshirani, R. J. (1998), *An introduction to the bootstrap*, Chapman & Hall/CRC, Boca Raton.

El Messlaki, F. (2010), Making use of multiple imputation to analyze heaped data, Master's thesis, University of Utrecht.

Kraus, F. and Steiner, V. (1995), 'Modelling heaping effects in unemployment duration models - with an application to retrospective event data in the German Socio-Economic Panel', *Discussion Papers of the Zentrum für Europäische Wirtschaftstorschung* **09**, pp. 1–29.

Lawless, J. F. (1982), *Statistical models and methods for lifetime data*, John Wiley & Sons, New York.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, 2 edn, Wiley, New York.

Marshall, A. W. and Olkin, I. (2007), *Life distributions: structure of nonparametric, semiparametric and parametric families*, Springer, New York.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992), *Model assisted survey sampling*, Springer, New York.

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

Wang, H. and Heitjan, D. F. (2008), 'Modeling heaping in self-reported cigarette counts', *Statistics in Medicine* **27**(19), pp. 3789–3804.

Wolff, J. and Augustin, T. (2000), 'Heaping and its consequences for duration analysis', *Sonderforschungsbereich* **386**(203), pp. 1–31.