WEP issue 03-96, 26 September 1996

A Monte Carlo study into time-aggregation in continuous and discrete-time hazard models

Frenkel ter Hofstede
Universities of Groningen and Wageningen
Michel Wedel
University of Groningen

Correspondence to:
Drs. F. ter Hofstede,
Department of Marketing and Marketing Research,
University of Wageningen,
Hollandseweg 1,
6706 KN Wageningen,
the Netherlands.
Tel: +31-317484393,
fax: +31-317484361.
E-Mail:
Frenkel.terHofstede@Alg.MenM.WAU.NL

About the authors

Frenkel ter Hofstede
EC-research associate - universities of Groningen and Wageningen

Frenkel ter Hofstede is research associate for department of Marketing and Marketing Research, Wageningen University, as well as the subdepartment of Marketing and Marketing Research, department of Business Administration, Faculty of Economics, University of Groningen. He got his doctorate in Econometrics at the University of Groningen in August 1994.

Currently he is working on a doctoral thesis on international market segmentation, promotors prof.dr.ir. J.E.B.M. Steenkamp and prof.dr. M. Wedel. His main interests are in the area of international marketing and segmentation with a special focus on methodology.

Michel Wedel
University of Groningen

Michel Wedel is Professor of Methods of Marketing Research and Market Structure Analysis at the subdepartment of Marketing and Marketing Research, department of Business Administration, Faculty of Economics, University of Groningen. He got his Doctorate in Biomathematics at the University of Leiden, The Netherlands.

He finished his doctoral thesis on December 7 1990, Wageningen University, which is entitled "Clusterwise Regression and Market Segmentation - developments and applications," promotors prof.dr.ir. M.T.G. Meulenberg, prof.dr. P.S.H. Leeflang and copromotor dr.ir. J.E.B.M. Steenkamp. He is author of many articles in a broad scope of international journals. His main interests are statistical and econometric applications to marketing research.

Abstract

The purpose of this study is to investigate the effect of time-aggregation in discrete and continuous-time hazard models. A Monte Carlo study is conducted in which data are generated according to underlying continuous and discrete-time processes, which data are aggregated into daily, weekly and monthly intervals. Under each of these conditions 400 datasets are generated, that vary in the parameters of the baseline hazard. These datasets are analyzed with flexible continuous-time and discrete-time proportional hazard models. The estimates of the structural parameter and of the baseline hazard, as well as the baseline hazard predicted by those estimates, seem robust to the form of the distribution of the data generation process when the time-aggregation window is small. Both estimates of continuous-time models and of discrete-time models suffer from time-aggregation, but the estimates of the discrete-time model are more sensitive to aggregation.

Key words:
Monte Carlo simulation; Continuous-time; Discrete-time; Proportional hazard model; Time-aggregation
JEL classification:
C15, C24, C41, C43

1 Introduction

Hazard models have become widespread in their use for the analysis of duration-time data in many scientific disciplines, including biology and medicine (e.g., Cox, 1972; Kalbfleisch and Prentice, 1980), sociology (e.g., Petersen, 1995, Vermunt 1996), marketing research (e.g., Vilcassim and Jain, 1991; Wedel et al., 1995), and economics (e.g., Kiefer, 1988; Lancaster, 1990). These models overcome the problems of accounting for censored observations of duration and time-varying explanatory variables, that arise in applying standard regression type models to duration data. The basic concept in hazard models is the probability of the occurrence of an event during a certain time interval, say t to $t + \Delta t$, given that it has not occurred before t, specified as:

$$PR[t \leq T \leq t + \Delta t | T \geq t, X(t)] \quad (1)$$

with $X(t)$ the structural variables, a set of covariates at time t. (The extension of the formulation to multiple states is straightforward but will not be provided to keep notation simple.) Parametric methods for duration data typically involve assumptions about the distribution of durations, which enables estimation by maximum likelihood. Two distinct classes of hazard models arise according to whether a discrete or a continuous distribution of the durations is assumed.

The discrete-time hazard is equal to the conditional probability of an event in equation (1), specified directly for given values of Δt:

$$PR[t \leq T \leq t + \Delta t | T \geq t, X(t)] = PR[y | T \geq t, X(t)] \quad (2)$$

where y equals the number of events that has occurred during the interval $[t, t + \Delta t]$. The discrete-time formulation was initiated by Prentice and Gloecker (1978), and has been extended by Laird and Olivier (1981), Efron (1988), Meyer (1990), Kiefer (1990) and Lindsey (1995). It has the advantage that it leads to simple model formulations, which enable one to accommodate censoring, covariates that vary within spells, and multiple events in a straightforward manner. Both the logistic binomial and log-Poisson regression models have been used to model discrete durations. The latter have the advantages of allowing for more than one event occurring in the unit discrete time interval (eg., Lindsey 1995).

In the continuous-time approach, the hazard is specified by letting Δt approach zero in (1), to yield the hazard-rate:

$$\lambda[t | X(t)] = \lim_{\Delta t \downarrow 0} \frac{PR[t \leq T \leq t + \Delta t | T \geq t, X(t)]}{\Delta t} \quad (3)$$

The density function of durations is:

$$f[t|X(t)] = \lim_{\Delta t \downarrow 0} \frac{PR[t \le T \le t + \Delta t | X(t)]}{\Delta t} = \lambda[t|X(t)] \exp\left[-\int_0^t \lambda[s|X(s)]\,ds\right] \quad (4)$$

The continuous-time approach appears to be among the most commonly used approaches in the economic literature (e.g., Lancaster, 1979, 1990; Kiefer, 1988; Gritz, 1993). An important special case, is presented by the proportional hazard models (Cox, 1972), for which:

$$\lambda[t|X(t)] = \lambda_0(t) \exp[X(t)'\beta] \quad (5)$$

where $\lambda_0(t)$ denotes the baseline hazard. Because the estimates of the structural parameters are sensitive to the specification of the baseline hazard (Trussell and Richards, 1985), Flinn and Heckman (1983) proposed a flexible Box-Cox formulation of the baseline hazard. It includes many of the frequently used distribution functions as special cases, and is formulated as:

$$\lambda_0(t;\theta) = \exp\left[\alpha_0 + \sum_{i=1}^{L} \alpha_i \frac{t^{\gamma_i} - 1}{\gamma_i}\right] \quad (6)$$

Kiefer (1990) and Lindsey (1995) used similar formulations to represent the baseline hazard in discrete-time proportional hazard models, in which it can be considered a smoothing approximation to a step-functional baseline. Other flexible approaches to formulate discrete hazard models are so-called semi-parametric approaches, in which each discrete time interval is represented by a parameter in the baseline hazard (e.g., Laird and Oliver, 1981; Kiefer, 1990), and the related Cox partial likelihood approach for continuous time models (Cox, 1975). Whereas these flexible approaches are available to represent the hazard-rate in both continuous and discrete-time hazard models, the choice between these two remains an important issue.

In many cases, the choice of a discrete or a continuous distribution of durations is an empirical issue, but sometimes (economic) theory may provide guidance. Discrete time intervals occur naturally in some situations due to the discrete nature of human behavior (for example occurrence of unemployment, presidential elections and purchase behavior). If the underlying process evolves in discrete time, the discrete time hazard models seem preferable. However, even if the underlying process is truly continuous, the measurements of that process may not be, because data are collected in discrete time intervals. Often, it is only known whether an event occurred within a certain discrete time window, such as an hour, a day, a week, or a month. In estimating continuous-time models for such data, events are usually assigned to the ends of the windows. This approach may give rise to biased estimates, a problem referred to as *time-aggregation bias* (or *interval censoring*). The time aggregation problem has not yet received much attention in the economics literature, although its importance has been recognized (Kiefer, 1988; Heitjan 1989). Time-aggregation has two major effects. First, it reduces the number of observed spells and thereby reduces sample size, and second it introduces measurement errors that are negatively correlated with the durations (see Bergström and Edin, 1992).

A few studies have investigated the problem of time-aggregation for continuous-time models. Petersen (1991) derived the size of the bias for a constant hazard-rate model, that does not include covariates or censoring. He concluded that (1) given the discrete time-unit of measurement, the higher the hazard-rate (shorter durations) the higher the bias due to aggregation; and that (2) for a constant hazard-rate, the wider the time window (the more time units it involves) the higher the bias in the estimated hazard-rate. These results were extended by Petersen and Koput (1992) to a model with a constant hazard-rate and one covariate, and to a constant hazard model accounting for right censoring. They also showed, using synthetic data, that the bias in the estimates of a covariate in a continuous time hazard model is reduced when the events are not assigned to the ends of the window, but to the mid-point of the window. A limitation of these studies is that in order to yield analytically tractable results, the models investigated were rather simple (i.e. a constant hazard-rate). It is not known to what extent these results carry over to more complicated models.

Two studies have addressed the effects of time aggregation empirically. Narendranathan and Stewart (1990) compare the estimates from a continuous-time Weibull model with a semi-parametric grouped hazard model, and discuss the problems of daily versus weekly data of unemployment. Bergström and Edin (1992) compare semi-parametric models (Cox partial likelihood, Cox, 1975) to simple (Weibull) and more complicated (generalized gamma) parametric models, for daily, weekly, monthly and quarterly unemployment data. The conclusions are that the estimates of the structural parameters are relatively robust to the distributional assumptions. However, different distributional assumptions produced different estimates of the time dependence of the baseline hazard. Time-aggregation was found to seriously affect the estimates of the parametric models, especially at higher levels of aggregation. The study is however based on the analysis of one single (and nonrepresentative) sample, in which the true underlying process remains unknown.

Time-aggregation is not the only cause of measurement errors in duration data. Other problems, that typically are due to recall bias in retrospective surveys, are under-reporting (Mathiowetz and Duncan, 1988), and heaping (Torelli and Trivellato, 1993). These issues, as well as unobserved heterogeneity (e.g., Kiefer, 1988) and semi-parametric models formulated in continuous or discrete time (Kiefer, 1988) are not considered in this study. The purpose is to further investigate the robustness of parametric discrete and continuous-time hazard models under various assumptions of the true underlying process and under different levels of time-aggregation. This will be done in a Monte Carlo experiment in which the true parameters of the data generating process are known.

2 The design of the study

2.1 Data generation

In the Monte Carlo study, the robustness of continuous and discrete time hazard models in various conditions of time aggregation is investigated. Synthetic data are simulated for 100 subjects within a period of 420 days. For both true discrete and continuous-time processes 400 replications are generated. A proportional hazard specification is used with one (time-invariant) regressor, with the structural parameter β across all replications. The regressor X is drawn from U(ln 0.5, ln 1.5), so that it does not de- or increase the baseline hazard with more than half its size. The regressors are subject specific and are fixed for all 400 replications. The baseline hazard is formulated using the first three terms of the Box-Cox specification in equation (6) above:

$$\lambda_0(t;\gamma) = \exp\left[\gamma_0 + \gamma_1 \ln(t) + \gamma_2 t\right] \quad (7)$$

where $\nu_1 \to 0$, $\nu_2 = 1$, $\delta_0 = \gamma_0 + \gamma_2$, $\delta_1 = \gamma_1$ and $\delta_2 = \gamma_2$. The formulation (7) nests a number of distributions of duration, such as the exponential ($\gamma_1 = 0$, $\gamma_2 = 0$), the Weibull ($\gamma_2 = 0$), and the Gompertz ($\gamma_1 = 0$).

In order to investigate the robustness of the hazard models under a wide range of different forms of the hazard, the parameters of the baseline hazard are varied across replications: γ_0 is drawn from U(-12, -4), γ_1 from U(-0.5, 0.5), and γ_2 from U(0, 0.1), where γ_0 is adjusted upward or downward in steps of 0.5 to ensure a reasonable number of events. To prevent defective hazards, γ_2 is restricted to the positive domain. For the continuous-time model the durations are generated by drawing random numbers u~U(0,1) and applying the inverse cumulative distribution function $\mathbf{F}^{-1}(u)$ (since the cdf. is not a closed form expression, a binary search algorithm was applied to solve the inverse transformation). For the discrete-time model events are generated by drawing from a Poisson distribution with expectation $\lambda[t|X]$, specified according to (5) and (7), where t is set equal to the midpoint of the intervals. We assume that the true underlying hazard occurs in time periods of one day. The last spell not containing an event is treated as right-censored. In order to simulate time-aggregation, the continuous time data are aggregated into spells consisting of 420 days, 60 weeks and 14 months, where each month is considered to consist of 30 days. Similarly, the discrete time data are aggregated into weeks and months.

The datasets thus generated are each analyzed with both a continuous-time and a discrete-time proportional hazard model, with the hazard specified as in equations (5) and (7). We use the same form of the baseline hazard for continuous and discrete-time models since the purpose of the study is to compare parametric formulations of the hazard in discrete and continuous time. In both the discrete and continuous-time model the correction suggested by Petersen (1991), of assigning the events to the mid-point of the discrete time windows, is applied. The models are estimated by the method of maximum likelihood, the likelihood being maximized by the quasi-Newton method of Broyden, Fletcher, Goldfarb and Shanno, as implemented in the GAUSS (Aptech, 1995) system. Since both the discrete and continuous-time hazard models have log-linear specifications, the likelihood is a concave function in the parameters (Lancaster, 1990) and convergence to a global maximum is guaranteed.

2.2 Dependent measures

The following dependent measures are used to assess the robustness of the models (we use k = 1, ..., 400 to indicate replications):

1. The mean parameter estimates across all replications;

2. The squared error of the estimates of the baseline parameters in replication k :

$$SE(\gamma_l^{(0)}, \hat{\gamma}_l^{(k)}) = \left(\gamma_l^{(0)} - \hat{\gamma}_l^{(k)}\right)^2, l = 0, 1, 2$$

3. The squared error of the estimate of the structural parameter:

$$SE(\beta^{(s)}, \hat{\beta}^{(s)}) = \left(\beta^{(s)} - \hat{\beta}^{(s)}\right)^2$$

4. The 2.5th, 50th and 97.5th percentiles of the empirical distribution of the relative error in the baseline hazard, calculated across the replications, for a grid of points $j = 1, ..., J$ in the [0,120] duration domain:

$$RE\left(\lambda_0(t_j^{(s)}; \hat{\gamma}^{(s)}), \lambda_0(t_j^{(s)}; \gamma^{(s)})\right) = \frac{\lambda_0(t_j^{(s)}; \hat{\gamma}^{(s)}) - \lambda_0(t_j^{(s)}; \gamma^{(s)})}{\lambda_0(t_j^{(s)}; \gamma^{(s)})}$$

5. The sample log-likelihood: $l(\hat{\beta}^{(s)}, \hat{\gamma}^{(s)})$

6. A likelihood-ratio (LR) test for the difference in the actual and estimated parameters for each of the replications: $LR^{(s)} = -2\left(l(\beta^{(s)}, \gamma^{(s)}) - l(\hat{\beta}^{(s)}, \hat{\gamma}^{(s)})\right)$

7. Whether or not the null-hypothesis that the estimates equal the true values is rejected ($p < 0.05$) on the basis of the LR-test, $r^{(s)} = 0/1$

In order to investigate the performance of the hazard models under the above conditions, we analyze the squared-errors of the four parameters, the LR statistic and whether or not the null-hypothesis is rejected, with regression analyses. The mean parameter estimates and log-likelihoods are reported for descriptive purposes and the error in the baseline is depicted graphically. Each analysis is based upon 4800 observations. For the squared errors and the LR test statistic a gamma-regression with a log-link is used, for the rejection of the null-hypothesis a binomial logit regression. The explanatory variables are A. the data generating process (continuous/discrete), B. the hazard model specification (continuous/discrete), C. the level of aggregation (daily/weekly/monthly), and D. the mean length of the durations in the data, and the first and second order interactions between these variables. The mean duration is included as a factor in our analyses, because previous work by Petersen and Koput (1992) has revealed it to affect time aggregation problems. For each of the dependent variables deviance ratio-tests are reported. Deviance ratio's are the ratio's of the scaled deviance-contributions of the respective terms and the residual deviance of the full model including all terms. The deviance is minus twice the log-likelihood ratio between the fitted model and a full model that explains all the variation in the data (McCullagh and Nelder, 1989). In order to reduce capitalization on chance, effects are considered significant at $p < 0.01$.

3 Results

Table 1 provides the deviance ratio test results of the regressions of the five dependent measures. It shows, that the model specification, the level of aggregation and the interaction between these two variables significantly affect all six measures. The second order interaction of data, model and aggregation level effects all measures, except SE(β) and SE(γ_2). Note that the effect of level of aggregation is by far the largest among all measures and that the effect of the data generating process is relatively small compared to the other main effects. The remaining effects show a mixed pattern. Because of the interactive effects the means of the

dependent measures are reported for all combinations of model specification, data distribution and level of aggregation in Tables 3 and 4 below.

The mean duration appears to affect all dependent measures significantly, except for the squared error in the structural parameter. The interactive effect of duration and level of aggregation affects all six dependent measures significantly, and the second order interaction between mean duration, model and aggregation level affects all dependent measures, except for the probability of rejecting the null hypothesis. The other interaction effects of mean duration show a mixed pattern. The estimates of the regression coefficients of mean duration and aggregation level according to model type are presented in Table 5.

Table 1: Deviance ratio's of mean duration (Dur), aggregation level (Agg), data generating process (Data) and the model (mod), from the regressions of six dependent measures

Term	Df	SE(γ_0)	SE(γ_1)	SE(γ_2)	SE(β)	LR-χ	r
Data	1	69.63*	34.81*	5.50	0.01	0.54	28.81*
Mod	1	101.20*	264.57*	502.72*	490.85*	5699.35*	129.49*
Aggr	2	294.84*	747.12*	1716.62*	708.89*	11324.93*	3001.87*
Data.Mod	1	5.51	0.06	15.78*	5.35	156.37*	442.34*
Data.Agg	2	10.56*	3.11	3.39	1.35	38.87*	32.23*
Mod.Agg	2	47.15*	74.47*	132.63*	70.33*	562.22*	33.04*
Data.Mod.Agg	2	29.38*	13.93*	2.85	4.35	62.67*	9.47*
Dur	1	216.39*	56.69*	124.44*	6.08	558.08*	82.37*
Dur.Data	1	1.56	3.53	2.60	0.96	10.53*	40.41*
Dur.Mod	1	12.51*	2.00	1.14	2.82	50.33*	0.47
Dur.Agg	2	83.31*	55.71*	13.91*	23.63*	270.36*	37.01*
Dur.Agg.Dat	2	2.35	5.24*	6.11*	1.05	10.89*	0.27
Dur.Mod.Dat	1	5.42	5.43	3.20	0.65	4.55	0.34
Dur.Agg.Mod	2	32.04*	14.81*	14.42*	5.09*	29.69*	0.10
* = $p < 0.01$, residual df: 4783							

Table 2 provides the mean parameter estimates for the analyses of the synthetic data (the averages of the true parameter values across replications are also depicted), Table 3 shows the root-mean squared errors (RMSE's) of the parameter estimates, obtained by averaging the SE's across replications and taking the square-root. Table 4 shows the log-likelihoods and related statistics. Figs. 1 to 4 show the percentiles of the relative error in the baseline hazard. In the following pages (3.1 and 3.2) you'll find these as inline images.

Table 2: Average parameter estimates for discrete and continuous data, and aggregation into daily, weekly and monthly periods.

Parameter	Data: Model:	Continuous Continuous	Discrete Discrete	Continuous Discrete	Discrete Continuous
γ_0	Actual	-7.076	-6.790	-7.076	-6.790
	None	-7.236	-	-	-
	Days	-7.259	-6.705	-7.233	-6.789
	Weeks	-7.424	-7.367	-8.062	-6.921
	Months	-8.594	-8.480	-8.631	-7.329
γ_1	Actual	0.018	-0.023	0.018	-0.023
	None	0.064	-	-	-
	Days	0.075	-0.054	0.095	-0.025
	Weeks	0.151	0.300	0.529	0.087
	Months	0.731	1.093	1.183	0.610
γ_2	Actual	0.052	0.047	0.052	0.047
	None	0.052	-	-	-
	Days	0.051	0.048	0.050	0.047
	Weeks	0.049	0.035	0.034	0.044
	Months	0.028	0.000	-0.003	0.020
β	Actual	1.000	1.000	1.000	1.000
	None	1.014	-	-	-
	Days	1.013	0.992	0.998	0.991
	Weeks	1.001	0.899	0.889	0.975
	Months	0.866	0.621	0.594	0.800

Table 3: RMSE's of parameter estimates for discrete and continuous data, and aggregation into daily, weekly and monthly periods.

Measure	Data: Model:	Continuous Continuous	Discrete Discrete	Continuous Discrete	Discrete Continuous
RMSE(γ_0)	None	0.922	-	-	-
	Days	0.921	0.730	0.842	0.866
	Weeks	1.017	1.641	2.175	0.973
	Months	2.920	2.633	2.628	1.221
RMSE(γ_1)	None	0.294	-	-	-
	Days	0.294	0.246	0.287	0.280
	Weeks	0.334	0.627	0.859	0.326

	Months	1.187	1.476	1.525	0.748
RMSE(η_2)	None	0.005	-	-	-
	Days	0.005	0.005	0.006	0.005
	Weeks	0.007	0.017	0.025	0.006
	Months	0.037	0.060	0.068	0.033
RMSE(β)	None	0.129	-	-	-
	Days	0.129	0.126	0.126	0.127
	Weeks	0.129	0.169	0.180	0.132
	Months	0.212	0.441	0.470	0.269

Table 4: Average log-likelihoods and likelihood ratio-tests for discrete and continuous data, and aggregation into daily, weekly and monthly periods.

Statistic:	**Data: Model:**	**Continuous Continuous**	**Discrete Discrete**	**Continuous Discrete**	**Discrete Continuous**
$l(\hat{\alpha}, \hat{\gamma})$	None	-2605	-	-	-
	Days	-2607	-2597	-2620	-2550
	Weeks	-2615	-1549	-1563	-2512
	Months	-2678	-888	-909	-2431
LR (df=4)	None	3.86	-	-	-
	Days	3.94	3.50	4.25	3.62
	Weeks	6.91	15.91	28.05	15.12
	Months	111.54	922.25	1004.77	179.84
r : % reject	None	5	-	-	-
	Days	5	2	9	4
	Weeks	12	54	72	59
	Months	85	100	100	100

Table 5: Coefficients of the regression of six dependent measures on mean duration and its interaction with the level of aggregation, for discrete and continuous time models

Term	**SE(η_0)**	**SE(η_1)**	**SE(η_2)**	**SE(β)**	**LR-χ**	**r**
Continuous-time model						
Days	3.645	4.808*	10.094*	5.424*	1.521*	2.085*
Weeks	3.536	4.942	10.363	5.212	4.555*	2.806*

Months	0.181*	0.046*	3.794*	2.462*	8.183*	9.619*
Dur.Days	0.046*	0.032	0.006	0.019*	0.004	0.019
Dur.Weeks	0.329*	0.256*	0.019	0.117*	0.194*	0.250*
Dur.Months	0.009*	0.010*	0.053*	-0.003*	0.053*	0.057*
Discrete-time model						
Days	2.721*	3.839*	9.162*	5.376*	1.689*	2.044*
Weeks	4.593	4.146	7.318*	3.459*	4.515*	2.531*
Months	2.609	0.770*	4.272*	0.936*	8.202*	10.200*
Dur.Days	0.029*	0.015*	0.021*	0.018*	0.007*	0.027*
Dur.Weeks	0.520*	0.306*	0.086	0.010	0.182*	0.242
Dur.Months	0.072*	0.026	0.024	0.012*	0.024*	0.000
* = p < 0.01, relative to the estimate for days, the estimates for daily data are tested for differences from zero						

3.1 Results when the underlying process is correctly specified

Table 2 and 3 show that the bias and RMSE in the parameter estimates of the baseline hazard of both the continuous-time models and the discrete-time models are low when the models are correctly specified. In the continuous-time models the estimate of g1 is upward biased on average, but its bias in the discrete-time models appears to be (slightly) negative. As expected, the regression parameter is very close to the true value on average for both model specifications. The LR test (Table 4) supports these findings. Across all 400 replications the null-hypothesis that the estimates are equal to the true values are close to the nominal percentage of 5%.

Table 3 also shows that the RMSE of the parameter estimates of the continuous-time models hardly increases when the continuous data are aggregated into days. Figure 1 shows the relative bias in the baseline hazard: the median is close to zero across the range of durations depicted. For durations up to, say, ten days, the interval which contains 95% of the estimated hazards is relatively wide. This is caused by the growing importance of the term $\gamma \ln(t)$ and of its estimation uncertainty when t approaches zero. Moreover, due to the parametric form of the baseline, there may be relatively more hazards near zero in this range, which explodes the relative errors.

The LR statistic and the rejection percentage also show that the estimates are not much further from their true values than those obtained from the ungrouped data. The log-likelihood is quite close to that of the ungrouped data.

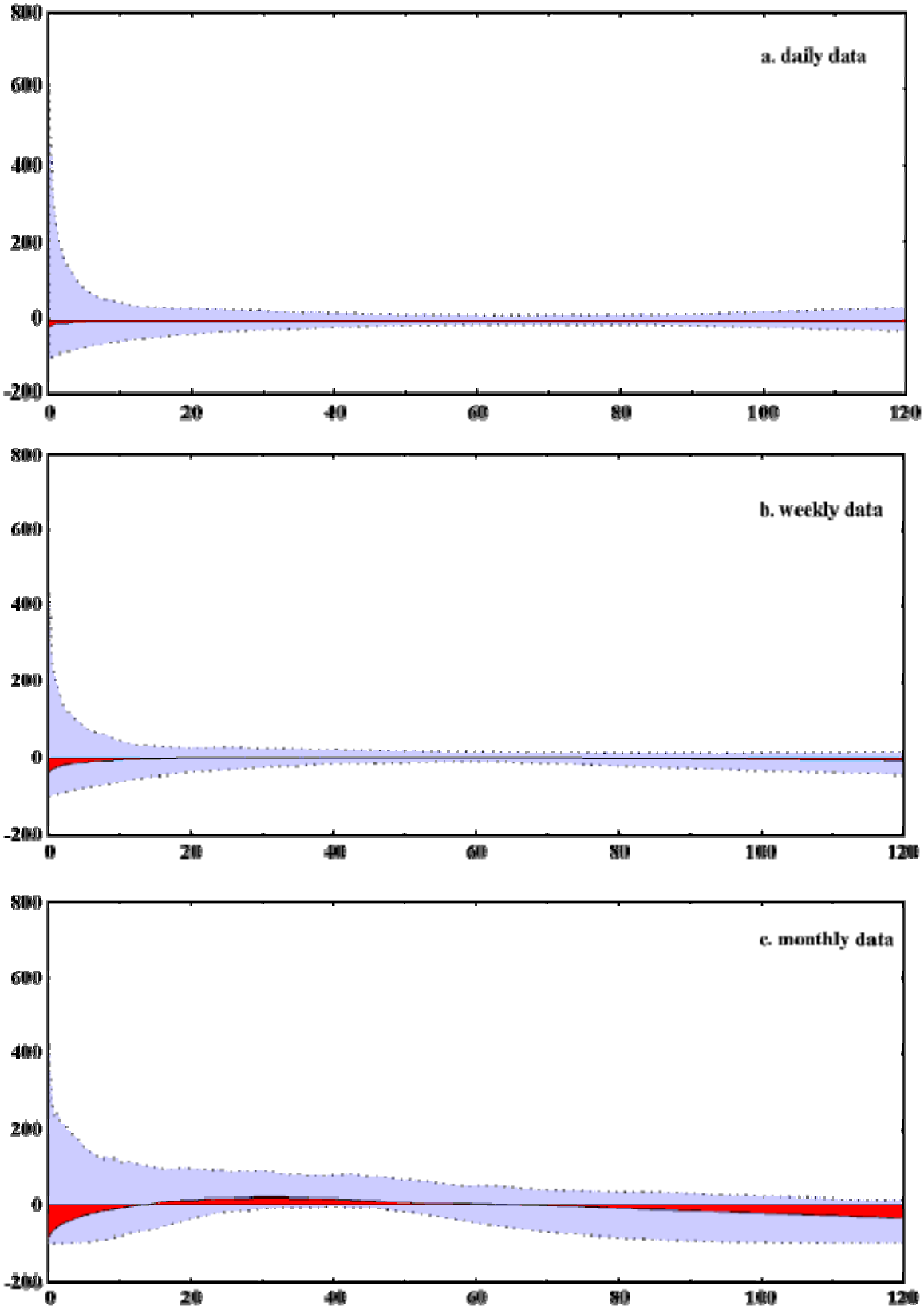


Fig 1. Relative bias in the hazard continuous data, continuous-time hazard model.

For weekly data, the bias in the estimates of the baseline parameters of the continuous-time models increases, the RMSE's increase by 10, 14 and 40%, respectively. Fig. 1 however shows that the relative bias in the baseline hazard itself hardly increases: the median of the

estimation error in the hazard is close to zero across the entire duration range, with exception of the very short durations for which it is somewhat smaller than zero. This effect is caused by the measurement errors, introduced by the aggregation, being negatively correlated with duration (very short durations being excluded from the data), which leads to an underestimation of the hazard for short durations. The RMSE of the regression parameter for weekly data (Table 3) does not increase relative to the original data. However, the average value of the LR test-statistic increases and the null-hypothesis that the estimates are equal to the true values is rejected rises to 12% of the cases. Note that due to the aggregation the log-likelihood ratio of the model with the "true" parameters is no longer distributed as Chi-square, so that the assumptions for the LR test to be valid no longer hold. Table 4 shows that the log-likelihood is lower than that for daily data, which we attribute to the measurement error induced by time aggregation.

For monthly data the error in the parameters of the baseline hazard of the continuous-time model is large: the RMSE's are 3.2, 4.04 and 7.4 times those of the ungrouped data. Corresponding to the results for daily and weekly data, γ_0 and γ_1 are negatively biased on average, and γ_2 positively. Due to the underrepresentation of short spells in the data, the median error in the baseline hazard is negative at durations in the lower range (Fig. 1), and more so than for weekly data. For higher durations, the 95% coverage interval increases substantially relative to weekly data. This is caused by the fact that the number of observations for the monthly data decreases, resulting in larger estimation errors. The effect is enlarged since the different functional forms that were simulated for the hazard may invoke different directions of the bias. This effect becomes stronger as the aggregation level increases. The estimate of the structural parameter is severely biased. Table 2 shows that this parameter is underestimated by about 13% on average, and that its RMSE increases by 64% (Table 3). The likelihood-ratio test rejects the null-hypothesis that the estimates are equal to the true values in the majority (85% of the cases). This is due to the measurement errors introduced by the time-aggregation and the likelihood test statistic at the "true" parameter values not being distributed as Chi-square. The log-likelihood (Table 4) has decreased further relative to the weekly data.

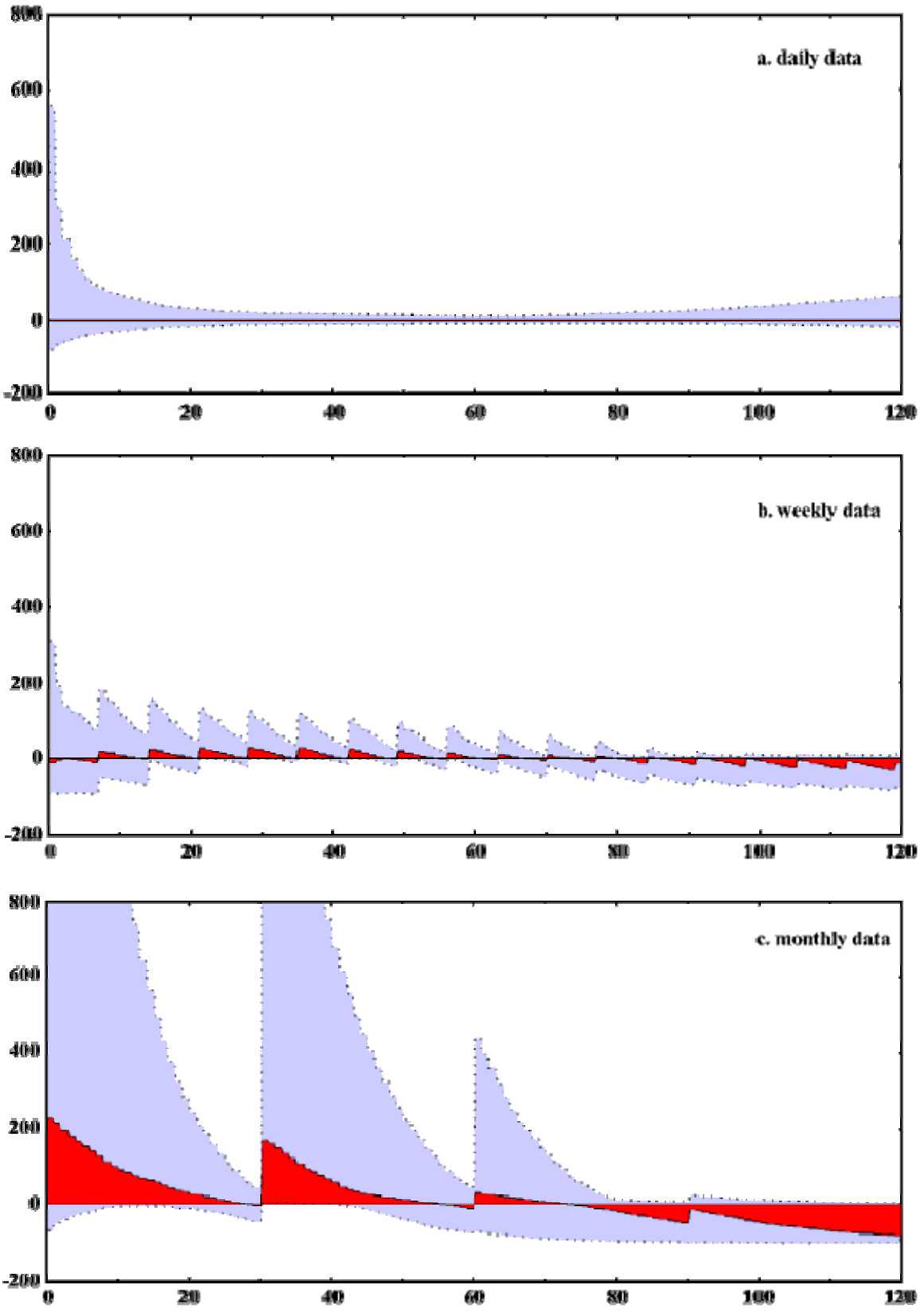


Fig 2. Relative bias in the hazard discrete data, discrete-time hazard model.

The estimates of the discrete-time model seem to be affected much more by the aggregation of durations into weeks, as compared to those of the continuous-time model. The RMSE's of the estimates of the baseline hazard increase by 125%, 155% and 240% respectively.

However, Fig. 2 shows that the effect on the error in the baseline hazard itself is modest. Due to the discrete nature of the estimated hazard, the curves of the percentiles show a spiked pattern. For shorter durations the bias within the weekly intervals decreases, whereas for longer durations it increases. In contrast to the continuous data, here one observes the tendency to overestimate the true hazard at the lower end of the duration range. Since the discrete model accounts for multiple events within one time-period, shorter spells are not discarded, and the negative bias does not occur for the shorter durations. The estimate of the structural parameter is affected by grouping the observations into weeks: Table 2 shows that on average the true value is underestimated by about 10%, Table 3 shows that the RMSE increases by 34% relative to the ungrouped data. The average value of the LR test increases substantially relative to that for the daily data, and the null-hypothesis that the estimated parameters are equal to the true parameters is rejected in more than half of the runs. As compared to the continuous-time models, the log-likelihood has increased relative to the daily data (Table 4), which is attributable to a large decrease in the number of spells due to aggregation which in this case apparently dominates te effect of measurement error induced by time aggregation (in the discrete-time models data records do not pertain to spells, but one data record is created per subject for each observation period).

For the discrete data grouped into monthly periods, the situation becomes progressively worse. Tables 2 and 3 show that the baseline hazard parameters are quite far off, and the RMSE's increase by 3.6, 6.0 and 12.0 times relative to the ungrouped situation. Similar to the continuous-time model on average  and  are negatively biased, and  positively. Fig. 2 shows that the bias in the baseline hazard is also large. As with the weekly data, the median error is positive at low and negative at high durations, but the effect is much stronger. Underestimation of the hazard of long durations is caused by the effects of censoring in combination with the rounding errors of time-aggregation, which are compounded in such situations. Additionally, the 95% coverage interval is quite wide, which is caused by a substantial reduction in the number of observations due to aggregation into months. Table 2 shows that the regression parameter is underestimated by 38% on average, the RMSE of this parameter increases 3.5 times relative to the ungrouped data. The LR statistic is very large indeed, and in all of the cases the null hypothesis is rejected. Due to a substantial reduction in the number of spells, the log-likelihood has increased dramatically relative to the weekly data (Table 4).

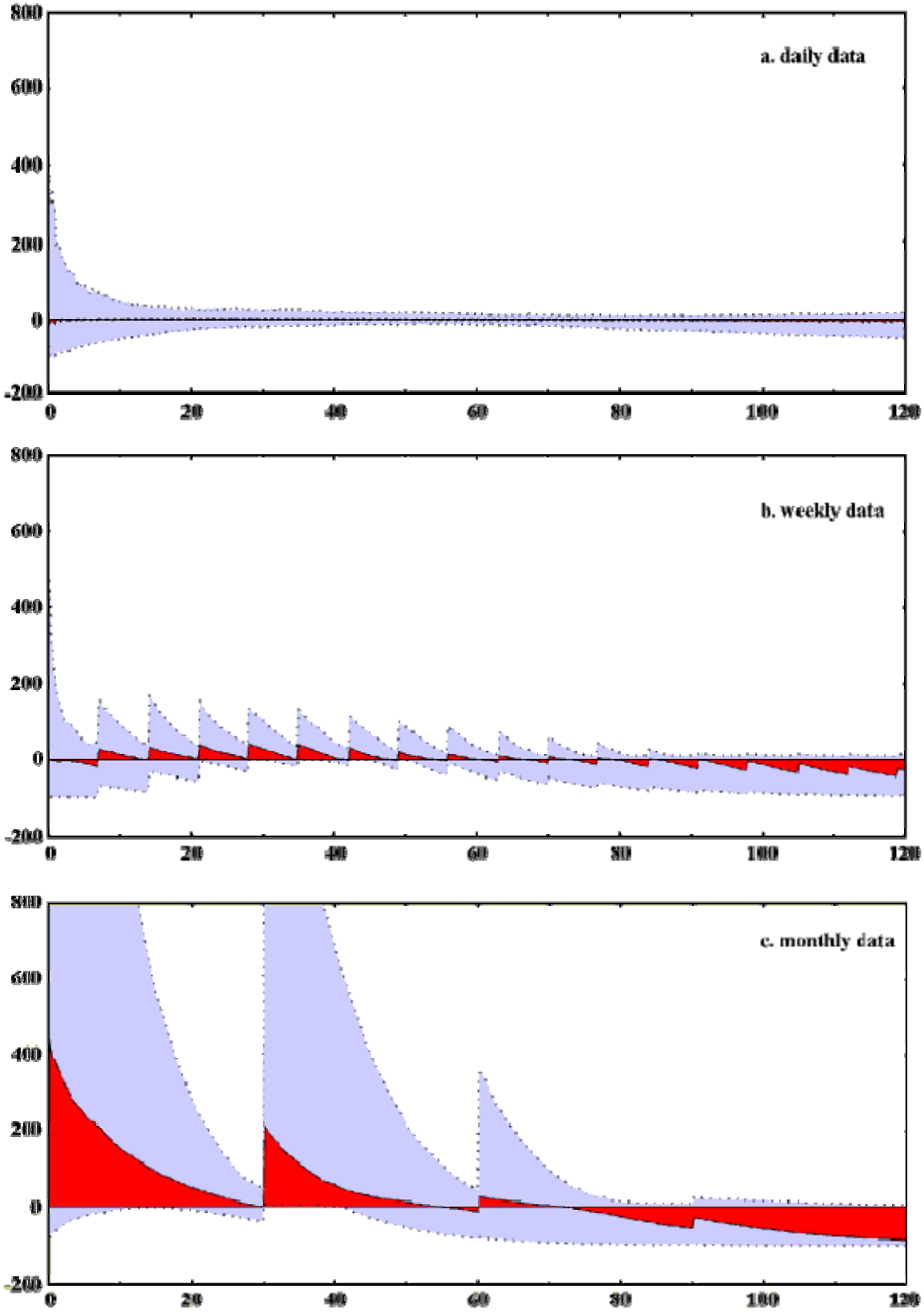
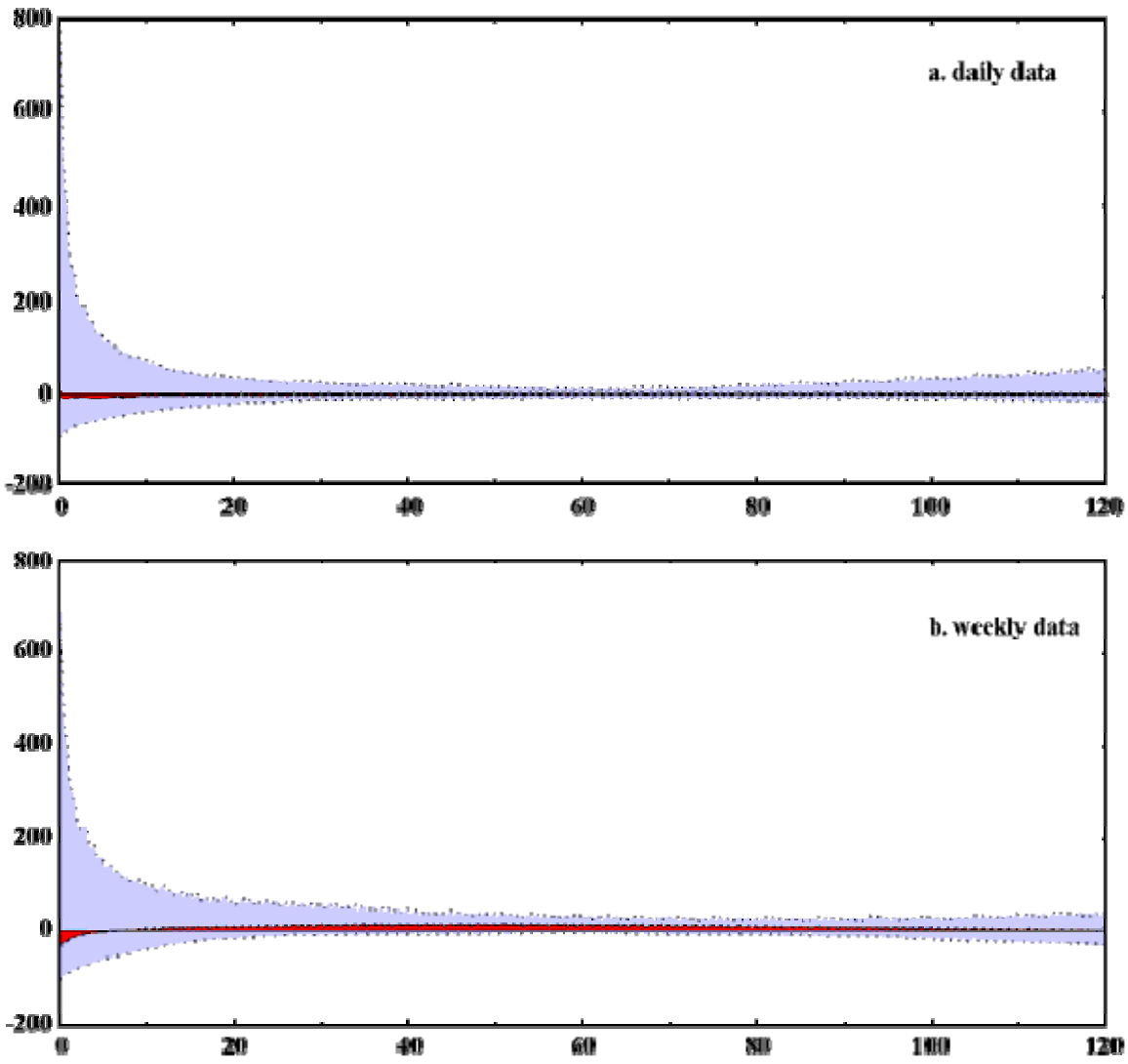


Fig 3. Relative bias in the hazard continuous data, discrete-time hazard model.

3.2 Results when the underlying process is misspecified

Tables 2, 3 and 4, and Figures 3 and 4 show the results when the model is incorrectly specified, i.e. when a discrete-time model is used, but the underlying process is continuous, and visa versa. When discrete-time models are used, but the underlying process is continuous, the average estimates in Table 2, the error in the hazard (Fig 3) and the RMSE's in Table 3, indicate that for daily data the discrete-time models perform comparable to the continuous-time models. The LR tests in Table 4 show that the hypothesis that the estimated parameters equal their true values is rejected in 9% of the replications. Note again that the LR test statistic is not distributed as chi-square at the "true" parameter values. The Tables and Figure show, that when the continuous data are aggregated into weeks or months, the performance of discrete-time models deteriorates rapidly. The 95% coverage intervals of the error in the baseline hazard are comparable to those when the underlying process is discrete, but the median bias seems to increase slightly due to the assumption of continuous data. The RMSE's for the parameters increase strongly, and the hypothesis that the parameters equal their true values is rejected in 72% and 100% of the cases, respectively for weekly and monthly data. The log-likelihood increases substantially due to the reduction in the number of spells.



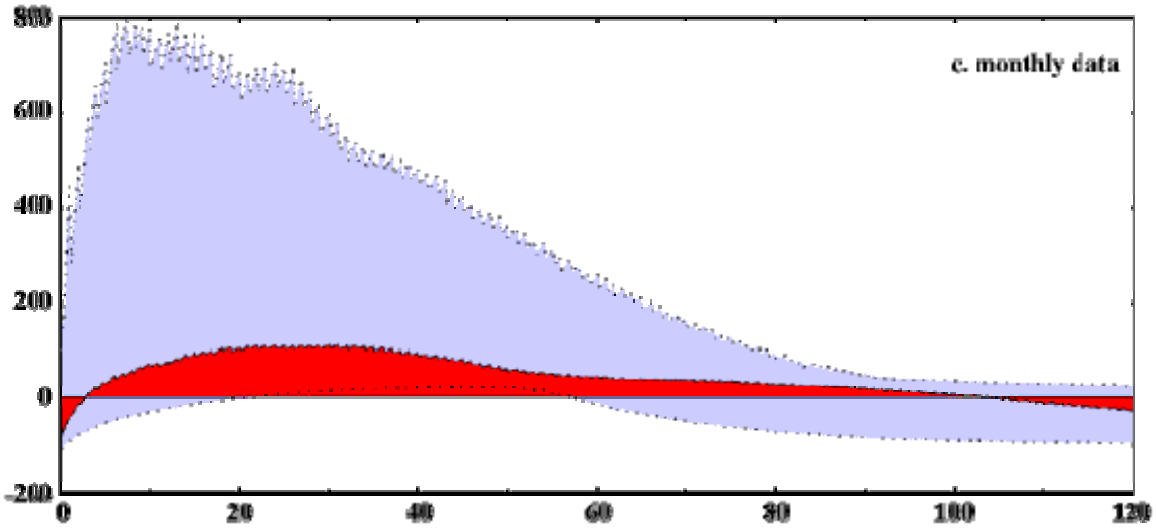


Fig 4. Relative bias in the hazard discrete data, continuous-time hazard model.

When the true underlying process is discrete and the data are not aggregated (i.e. daily data), the performance of continuous-time models is comparable to that of the discrete-time models (see Tables 2 and 3, and Fig. 4). The hypothesis that the estimates equal their true values is rejected in 4% of the replications. For weekly or monthly data, the performance of the continuous-time model deteriorates, as indicated by the RMSE's, and the LR-test statistics. Note that in these situations the RMSE's of the estimates of the continuous-time models are lower than those of the corresponding discrete-time models (Table 3), while Table 2 shows that the estimates of the parameters are on average closer to their true values. Compared to the continuous data, the median error in the baseline as well as the 95% coverage interval increases dramatically when data are aggregated to months. Note that the log-likelihood for the continuous-time models does not decrease as strongly as that for the discrete-time models if the data are aggregated. As outlined before, this is the result of the dramatic decrease in observations when applying the discrete model.

3.3 Mean duration effects

In order to investigate the influence of the mean duration on the performance of the hazard models, Table 5 presents the coefficients of the regression of the dependent measures on mean duration for continuous-time and discrete-time models. The coefficients shown are adjusted for the effects of the other factors and their interactions. First, the results show - what was already apparent from Table 3 - that the squared errors of the parameter estimates increase when the level of aggregation increases from days to months, for both discrete and continuous-time models. Except for the squared error in β and η for discrete-time models, the difference between weekly and daily data is not significant. Correspondingly, the LR-chi square value as well as the probability of rejection increase for higher levels of aggregation.

For daily data and for both model specifications, longer average durations tend to lead to larger squared errors of the parameter estimates (however, longer durations tend to improve the estimates of η, which is significant in discrete-time models. This effect may be caused by increases in the range of durations, which improves the conditions for identification of η). For discrete-time models longer durations tend to be associated with lower LR statistics and lower probabilities of rejecting the true model. Longer durations lead to less spells being

observed, with negative consequences for the precision in the estimates and the power of the LR-tests. For weekly data, the same pattern is observed, be it that the relations of mean duration and the dependent measures are stronger than for daily data. Again, this may be caused by the reduction in the number of observed spells. For monthly data, the relationship of mean duration and the squared error of the parameter estimates is mixed, but it is negative for the regression parameter in both model specifications. The results imply that longer durations in the data positively influence the estimates of the structural coefficients Apparently the effects of measurement error here dominate the sample-size effect: for data with longer durations on average, the measurement errors due to rounding are less important as compared to short durations. This is consistent with the findings of Petersen (1992), who demonstrated that for simple models this situation should occur. Correspondingly, the LR statistics and probability of rejection of the null-hypothesis are lower for longer durations.

4 Conclusions

The purpose of this paper was to investigate empirically the effects of the data generating process time-aggregation on hazard model estimates. A flexible specification of the baseline hazard, based on a Box-Cox formulation, nesting a variety of commonly used duration distributions was employed. The same flexible specification of the baseline hazard was used in the continuous and discrete-time models. By doing this the effects of the distributional assumptions per se could be investigated.

A first conclusion of the study is that both the estimates of the regression parameter and baseline hazard, as well as the baseline hazard predicted by those estimates, seem robust to the actual data generation process when the time-aggregation window is small. When continuous or discrete daily data are used, both continuous-time and discrete-time models yield estimates that are close to the true parameter values. (The error in the estimates of the parameters of the baseline hazard is partly caused by the collinearity of terms in t and $\ln(t)$ in the baseline.) Our findings support Kiefer (1990) and Lindsey (1995): the estimates of a continuous-time hazard can be approximated accurately by a discrete-time model, by taking the size of the time window to be small. This enables the estimation of hazard models using Poisson regression models, with the advantages of simple formulations in which time-varying covariates, multiple events within spells and censoring are easy to deal with.

Secondly, both continuous-time models and discrete-time models suffer from time-aggregation. The effect of time aggregation is twofold: it reduces the number of observed spells -an effect that is much larger in discrete-time models- and it introduces measurement errors that are negatively correlated with observed durations. In both discrete and continuous-time models the estimates of the baseline hazard become progressively biased when the level of aggregation increases. On average, the bias in the intercept and in the terms involving $\ln(t)$ was negative and that in t positive. The results support the findings of Bergström and Edin (1992), that in complex hazard models time aggregation produces dramatic changes in the baseline parameters. However, the results showed that the effects on the predicted baseline hazard itself were much less severe. Apparently, the reduction in the number of spells due to aggregation induces collinearity in the estimates. When data are aggregated into weeks or months discrete time-models overestimate the baseline hazard for short durations, while the hazard is underestimated for continuous-time models. Whereas the latter may be attributed to the exclusion of short spells, the discrete models, which do not suffer from this problem, seem

to overestimate the effect of these short durations because they incorrectly assume them to occur randomly in the intervals.

Time-aggregation causes the estimate of the structural parameter to be biased towards zero. This confirms previous results of and Petersen and Koput (1992). Note that their results were derived from one synthetic dataset, while the results in this study are obtained from a large number of datasets with a large variety of different forms of the baseline hazard. For monthly data the aggregation effects are quite dramatic: the bias is around 13% for the continuous-time models, and around 38% for the discrete-time models, on average. Overall, the estimates of the discrete-time model are more sensitive to time-aggregation than those of the continuous-time model. Bergström and Edin (1992) found a discrete model to be less sensitive to aggregation than a continuous-time model. However they used parametric formulations of the baseline for the continuous-time model, and a semi-parametric formulation for the discrete-time model. Their result is potentially due to the more flexible formulation of the discrete-time vis-à-vis the continuous-time model. They also did not use an adjustment for the time-aggregation in the continuous-time model. In part, the relative robustness of the continuous-time models to time-aggregation in our study may be caused by the use of the mid-point adjustment suggested by Petersen (1992).

Our study revealed that if the mean duration of the data increases, the estimates for daily and weekly data (and both model specifications) become less precise. However, for monthly data the precision of most estimates, but especially those of the structural parameter, increases with mean duration. This may be explained from the counteracting effects of the reduction of sample size with aggregation, and the negative correlation of measurement errors and mean durations. When time aggregation windows are relatively small (days/weeks) the error in the parameter estimates increases with duration, because the effects of the reduction in the number of data points dominates the effect of the measurement error. When the time window is large (months) the effect of the measurement errors dominates, and longer average duration is associated with smaller biases in the parameters. Thus, our results supplement the analytical results from Petersen (1991) and Petersen and Koput (1992), who demonstrated analytically for simple models that the estimation bias should decrease with time-aggregation.

Finally, LR tests seem inappropriate when data are aggregated. The log-likelihood, even at the "true" parameter values, is not distributed as Chi-square, invalidating the use of LR tests to investigate nested models.

References

Aptech, 1995,
GAUSS maximum likelihood users manual, Aptech systems Inc (Maple Valley, WA).

Bergström, R. and P.A. Edin, 1992,
Time aggregation and the distributional shape of unemployment duration, Journal of Applied Econometrics 7, 5-30.

Cox, D.R. 1972,
Regression models and life tables, Journal of the Royal Statistical Society, B 34, 187-400.

Cox, D.R. 1975,
Partial likelihood. Biometrika, 62, 269-276.

Efron, B. 1988,

Logistic regression, survival analysis, and the Kaplan Meier curve, Journal of the American Statistical Association, 1988, 414-425.

Flinn, C. and J. Heckman 1983,
The likelihood function for the multi state multi-episode model, in: Advances in Econometrics, Vol. 2 (JAI-Press Inc., Greenwich, C.T) 225-231.

Gritz, R.M., 1993,
The impact of training on the frequency and duration of employment, Journal of Econometrics 57, 21-51.

Heijtjan, D.F. 1989,
Inference from grouped continuous data: a review, Statistical Science 4, 164-183.

Holford, T. L. 1980,
The analysis of rates and of survivorship using log-linear models, Biometrics 36, 299-305.

Kalbfleisch, J. D. and R.L. Prentice, 1980,
The statistical analysis of failure time data (John Wiley & Sons Inc., New York).

Kiefer, N.M., 1988,
Economic duration data and hazard functions, Journal of Economic Literature 26, 646-679.

Kiefer N.M., 1990,
Econometric methods for grouped duration data, in: J. Hartog, G.Ridder and J. Theeuwes, eds., Panel data and labor market studies, (North-Holland, Amsterdam) 97-117.

Laird, N. and D. Olivier, 1981,
Covariance analysis of censored survival data using log-linear analysis techniques, Journal of the American Statistical Association 76, 231-240.

Lancaster, T., 1979,
Econometric methods for the duration of unemployment, Econometrica 47, 939-956.

Lancaster, T., 1990,
The econometric analysis of transit data (Cambridge University Press, Cambridge, U.K).

Lindsey, J.K., 1995,
Fitting parametric counting processes using log linear models, Applied Statistics 44, 201-212.

Mathiowetz, N.A. and G.J. Duncan, 1988,
Out of work, out of mind: response errors in retrospective reports on unemployment, Journal of Business and Economic Statistics 6, 221-229

Meyer, B., 1990,
Unemployment insurance and unemployment spells, Econometrica 58, 757-782.

McCullagh, P. and J.A. Nelder, 1989,
Generalized linear models (Chapman and Hall, London).

Narendranathan, W. and M. Stewart, 1990,
An examination of the robustness of models of the probability of finding a job for the unemployed, in: J. Hartog, G.Ridder and J. Theeuwes, eds., Panel data and labor market studies (North Holland, Amsterdam) 135-156.

Petersen, T., 1991,
Time aggregation bias in continuous-time hazard-rate models, in: P.V. Marsden, ed., Sociological Methodology, Vol 21 (Basil Blackwell, Cambridge MA) 263-290.

Petersen, T., 1995,
Analysis of Event Histories, in: G. Arminger, C.C. Clogg and M.E. Sobel, eds., Handbook of statistical modeling for the social and behavioral sciences (Plenum Press, New York) 453-517.

Petersen, T. and K.W. Koput, 1992,
Time aggregation bias in hazard-rate models with covariates, Sociological Methods and Research, 21, 25-51.
Prentice, R., and L. Gloecker, 1978,
Regression analysis of grouped survival data with applications to breast cancer data, Biometrika, 67, 145-153.
Torelli, N., and U. Trivellato, 1993,
Modeling inaccuracies in job-search duration data, Journal of Econometrics, 59, 187-211.
Trussell, J. And T. Richards, 1985,
Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure, In: Sociological Methodology, N.B. Tuma (ed.), Jossey-Bass (San Fransisco), 242-276.
Vermunt, J.K., 1996,
Log-linear event history analysis (Tilburg University Press, Netherlands).
Vilcassim, N. J. and D.C. Jain, 1991,
Modeling purchase timing and brand switching behavior incorporating explanatory variables and unobserved heterogeneity, Journal of Marketing Research, 28, 29-41.
Wedel, M., W.A. Kamakura, W.S. DeSarbo and F. ter Hofstede, 1995,
Implications for asymmetry, non proportionality and heterogeneity in brand switching models from piece-wise exponential mixture hazard models, Journal of Marketing Research, 32, 457-463.