Technometrics
2016, Vol. 58, No. 1, pp. 95-103.

Tail estimation for window-censored processes

Holger Rootzén* and Dmitrii Zholud†

Abstract

This paper develops methods to estimate the tail and full distribution of the lengths of the 0-intervals in a continuous time stationary ergodic stochastic process which takes the values 0 and 1 in alternating intervals. The setting is that each of many such 0-1 processes have been observed during a short time window. Thus the observed 0-intervals could be non-censored, right-censored, left-censored or doubly-censored, and the lengths of 0-intervals which are ongoing at the beginning of the observation window have a length-biased distribution. We exhibit parametric conditional maximum likelihood estimators for the full distribution, develop maximum likelihood tail estimation methods based on a semi-parametric generalized Pareto model, and propose goodness of fit plots. Finite sample properties are studied by simulation, and asymptotic normality is established for the most important case. The methods are applied to estimation of the length of off-road glances in the 100-car study, a big naturalistic driving experiment. Supplementary materials that include MatLab code for the estimation routines and a simulation study are available online.

Keywords: Generalized Pareto distribution; Length-biased distribution; Off-road glance; Tail estimation; Traffic safety; 100-car naturalistic driving study.

1 Introduction

Let $X(t)$ be a stationary ergodic stochastic process which takes the values 0 and 1 in alternating intervals, as illustrated in Figure 1. Here 1 could mean that a technical system is in operation and 0 that it is being repaired, or 1 could be that a person is healthy, while 0 is that he suffers from an attack of some specific recurrent disease such as the relapsing form of multiple sclerosis. In the problem of visual inattention during driving, which initiated this research, 1 means that the driver of a car looks on the road, and 0 that she looks away from the road. This paper develops methods for using window censored observations to estimate the distribution of the lengths of the 0-intervals in such processes.

Very large naturalistic driving studies, costing hundreds of millions of dollars, are used as a tool to reduce traffic risks. Visual inattention, and in particular long off road glances, i.e. with the notation above, long 0-intervals, pose the largest dangers and are at the center of interest. The statistical methods needed for efficient use of these studies are still in an early stage of development. This paper contributes one central ingredient needed for this. Analysis of data from periodic reliability inspection, and prevalence based epidemiological studies where the time of onset of disease is not recorded, are other examples of areas where our methods can be used.

*Department of Mathematical Statistics
Chalmers University of Technology and University of Göteborg, Sweden.
E-mail: hrootzen@chalmers.se
†Department of Mathematical Statistics
Chalmers University of Technology and University of Göteborg, Sweden.
E-mail: dmitrii@chalmers.se

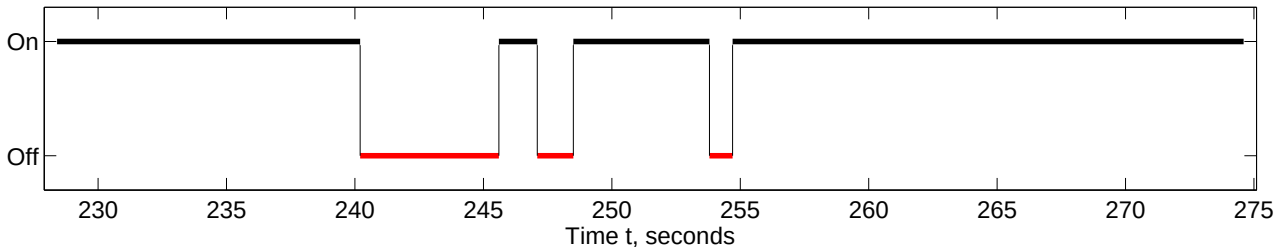

Figure 1: Visual behavior during a 45-second sideswiping near-crash (ID 8731) from the 100-Car naturalistic driving study. 'On' indicates eyes on road, 'Off'-eyes off road. The near-crash started at second 258.

Under widely applicable technical assumptions (including finiteness of the expected number of 0-intervals in finite time intervals), which we assume are satisfied, the ergodic long-run distribution function of the lengths of the 0-intervals,

$$F_0(x) = \lim_{T\to\infty} \frac{\#\ \{\text{0-intervals in } [0,T] \text{ which are shorter than x}\}}{\#\ \{\text{0-intervals in } [0,T]\}}$$

exists. In particular, if $X(t)$ is an alternating renewal process then F_0 is simply the common distribution function of the 0-intervals.

We assume that a number of independent stationary 0-1 processes of this kind have been observed during a randomly placed time window of fixed length $w > 0$ (the results also extend to windows of differing lengths). Since the process is stationary, we without loss of generality assume that this observation window is $[0, w]$. The challenge is that the 0-intervals which fall in $[0, w]$ may be left-censored, right-censored, or both left and right-censored, and that furthermore the left and doubly-censored observations are obtained by length-biased sampling. The goal is to estimate $F_0(x)$, or the tail of $F_0(x)$.

In the situation we mainly aim at here, the corresponding distribution $F_1(x)$ of lengths of 1-intervals is of less interest, or the 1-intervals are too long in relation to the window-length to make estimation reasonable. Our methods work both for long and for short 1-intervals.

We consider (i) a parametric statistical model for $F_0(x)$, and (ii) a semi-parametric generalized Pareto model, where the parametric form for $F_0(x)$ is only assumed to apply for x-values which exceed a threshold $u > 0$. For both models we develop conditional maximum likelihood estimation methods; introduce goodness of fit plots; study the methods by simulation; and apply them to data from a large naturalistic driving study, the 100-car study. For the heavytailed semi-parametric model we also show asymptotic normality.

The literature on statistical inference for window censored multistate processes is limited. Alvarez (2005) derives the maximum likelihood estimator for a stationary alternating renewal process, and, for the special case of a 0-1 continuous time Markov chain, proves asymptotic normality as the number of windows tends to infinity with window length kept constant. Karr (1994) considers maximum likelihood estimation for a 0-1 continuous time Markov chain, using observation of a single sample path in an expanding window. Non-parametric estimation for bi- and multivariate semi-Markov processes is studied in Alvarez (2006) and Ouhibi and Limnios (1999).

Starting with Cox (1969), Vardi (1982, 1985), length-bias sampling for ordinary renewal processes has received much more attention. For recent contributions see Gill and Keiding (2010), Zhao and Nagaraja (2011), Ning et al. (2013), Zhu et al. (2014), and the references in these papers. A different strand of literature considers left truncated, right-censored and size biased observations, often in connection with prevalent cohort deigns in epidemiology, see e.g. Qin et al. (2011) and the references

therein. In the literature on tail estimation, only right censoring seems to have been considered, see Einmahl et al. (2008).

Section 2 below introduces the models and estimators. In Section 3 small sample behavior of the methods are studied by simulation, and they are applied to the 100-car naturalistic driving study in Section 4. Results on asymptotic normality and MatLab code for the estimation routines are available in Supplementary Materials.

2 Models and likelihoods

Estimation under a full parametric specification of F_0 is considered in Subsection 2.1, and Subsection 2.2 introduces the tail estimation methods. Subsection 2.3 briefly discusses confidence intervals; regression modeling; goodness of fit plots; joint estimation of F_0 and F_1; long observation windows; and asymptotic normality.

For the estimation we only use those of the observation windows which intersect at least one 0-interval. Hence the likelihoods introduced below are conditional on the observation window intersecting at least one 0-interval. Further, in Subsections 2.1 and 2.2 we only use the first 0-interval in each window since this makes the independence assumption we need much less restrictive. This assumption is that conditionally on $X(0) = 1$, the lengths of the starting 1-interval and of the following 0-interval are mutually independent. We assume that F_0 and F_1 have continuous densities f_0 and f_1 and finite means. The finite mean condition is required for ergodicity of the process.

2.1 Conditional maximum likelihood estimation

Let S be the starting point of the first 0-interval in the observation window $[0, w]$, let L be the length of the observed part of this 0-interval, and let $\bar{F} = 1 - F$ denote the tail (or "survival") function corresponding to a distribution F. The lengths of the observed 0-intervals (see Figure 2) are classified as

nc) non-censored, i.e. with $S \in (0, w), S + L < w$: denoted $\ell_{nc,1}, \ldots \ell_{nc,n_{nc}}$,

rc) right-censored, i.e. with $S \in (0, w), S + L = w$: denoted $\ell_{rc,1}, \ldots \ell_{rc,n_{rc}}$,

lc) left-censored, i.e with $S = 0, L < w$: denoted $\ell_{lc,1}, \ldots \ell_{lc,n_{lc}}$, and

dc) doubly-censored, i.e. with $S = 0, L = w$: denoted $\ell_{dc,1} = \ldots \ell_{dc,n_{dc}} = w$,

where n_{nc} is the number of non-censored observations, n_{rc} is the number of right-censored observations, and so on. Further, if the observation window starts with an 1-interval we let $s_{nc,1}, \ldots s_{nc,n_{nc}}$ denote the lengths of those (left-censored) starting 1-intervals which are followed by a non-censored 1-interval, and we let $s_{rc,1}, \ldots s_{rc,n_{rc}}$ be the lengths of the starting (left-censored) 1-intervals which are followed by a right-censored 0-interval, see Figure 2; note that here the subscripts indicate the form of the censoring of the corresponding 0-intervals, and not of the 1-intervals themselves. It is well known that if $X(t)$ is an alternating renewal process, then the distribution of L conditional on $X(0) = 0$ (i.e. on $S = 0$) has the length-weighted "residual life" density and distribution functions

$$f_0^r(x) = \bar{F}_0(x)/\mu_0 \text{ and } F_0^r(x) = \int_0^x f_0^r(y)dy, \tag{1}$$

where the superscript "r" indicates "residual", and where $\mu_0 = \int x f_0(x)dx$ is the mean of F_0. It can be seen that (1) in fact holds not just for alternating renewal processes, but also for general stationary ergodic 0-1 processes, under conditions as in the introduction. Similarly, let $\mu_1 = \int x f_1(x)dx$ and let f_1^r and F_1^r be the residual life density and distribution functions obtained from F_1. Further (by ergodicity) $p_0 = Pr(X(0) = 0) = \mu_0/(\mu_0 + \mu_1)$ and $p_1 = Pr(X(0) = 1) = \mu_1/(\mu_0 + \mu_1)$.

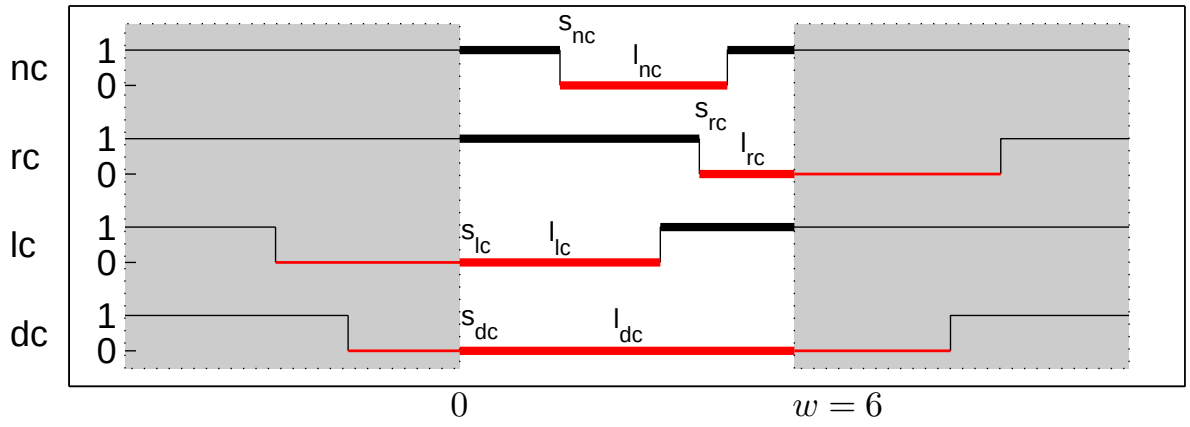

Figure 2: Censoring of the first 0-interval in the observation window: nc = no censoring, rc = right censoring, lc = left censoring, dc = double censoring. The s and l-values are the starting position and the length of the observed 0-interval, respectively. Grey indicates un-observed parts of the process.

Now, suppose additionally that there is a fully parametric specification, $F_0(x) = F_0(x;\theta)$ of the cumulative distribution function of the length of a 0-interval, with θ a finite dimensional parameter. Instead no additional conditions on F_1 are assumed. Then, using the independence of S and L for the "nc" and "rc" cases, observations ℓ of L and s of S contribute to the likelihood function with the following factors,

nc) $f_1^r(s)f_0(\ell;\theta)$ if the observation is uncensored,

rc) $f_1^r(s)\bar{F}_0(\ell;\theta)$ if the observation is right-censored,

lc) $f_0^r(\ell;\theta)$ if the observation is left-censored, and

dc) $\bar{F}_0^r(w;\theta)$ if the observation is doubly-censored.

Thus the full likelihood function based on the observed lengths of the zero-intervals and on the information if $X(0)$ is 0 or 1 is

$$\begin{aligned} L(\theta) &= p_1^{n_{nc}+n_{rc}} \prod_{k=1}^{n_{nc}} f_1^r(s_{nc,k}) f_0(\ell_{nc,k};\theta) \times \prod_{k=1}^{n_{rc}} f_1^r(s_{rc,k})\bar{F}_0(\ell_{rc,k};\theta) \\ &\quad \times p_0^{n_{lc}+n_{dc}} \prod_{k=1}^{n_{lc}} f_0^r(\ell_{lc,k};\theta) \times \bar{F}_0^r(w)^{n_{dc}}. \end{aligned} \tag{2}$$

Now, recall that $p_1 = \mu_1/(\mu_0+\mu_1)$ and $p_0 = \mu_0/(\mu_0+\mu_1)$ where $\mu_0 = \mu_0(\theta)$ and μ_1 is determined by F_1. Thus the factors p_1 and p_0 in the likelihood function couple information about F_1 with information about θ.

However, μ_1 is determined by the tail behavior of F_1, and in the situations we aim at the 1-intervals are much longer than the 0-intervals, and hence data collected in the short observation window $[0, w]$ contains little information about the tail of F_1. For a special case, a continuous time 0-1 Markov chain, Alvarez (2005) indicates that the loss of information from using the conditional likelihood instead of the full likelihood can be sizeable if the 0- and 1-intervals are of comparable lengths, but that the loss is small if the lengths are substantially different. The loss should be even smaller in the present situation. Thus p_1 and p_0 contain little useful information about θ.

We hence use a conditional log likelihood function, which in addition to being conditional on the observation window intersecting at least one 0-interval, also is conditional on the observed values

$n_1 = n_{nc} + n_{rc}$ and $n_0 = n_{lc} + n_{dc}$. Omitting factors which do not depend on θ and hence do not play a role in ML-estimation, the conditional log likelihood function then is

$$\begin{aligned} \ell(\theta) = \ell(\theta|\, n_0, n_1) &= \sum_{k=1}^{n_{nc}} \log f_0(\ell_{nc,k};\theta) + \sum_{k=1}^{n_{rc}} \log \bar{F}_0(\ell_{rc,k};\theta) \\ &+ \sum_{k=1}^{n_{lc}} \log f_0^r(\ell_{lc,k};\theta) + n_{dc} \log \bar{F}_0^r(w). \end{aligned} \tag{3}$$

Estimates of θ are obtained as $\hat{\theta} = \arg\max_\theta \ell(\theta)$.

Example. A simple and important special case is when F_0 is an exponential distribution, $F_0(x) = 1 - e^{-x/\sigma}$, $f_0(x) = \sigma^{-1}e^{-x/\sigma}$, where we have written $\theta = \sigma$. Then the residual life distribution is the same as the original distribution, and hence

$$\ell(\sigma) = -\sigma^{-1}\left(\sum_{k=1}^{n_{nc}} \ell_{nc,k} + \sum_{k=1}^{n_{rc}} \ell_{rc,k} + \sum_{k=1}^{n_{lc}} \ell_{lc,k} + n_{dc}w\right) - (n_{lc} + n_{nc}) \log \sigma.$$

Hence the conditional maximum likelihood estimate of σ_0 is the standard estimate of the scale parameter for a right-censored exponential distribution,

$$\hat{\sigma} = \frac{\sum_{k=1}^{n_{nc}} \ell_{nc,k} + \sum_{k=1}^{n_{rc}} \ell_{rc,k} + \sum_{k=1}^{n_{lc}} \ell_{lc,k} + n_{dc}w}{n_{lc} + n_{nc}}.$$

The standard error may be estimated by the inverse $\hat{\sigma}/\sqrt{n_{lc} + n_{nc}}$ of the observed information. □

If the exponential distribution does not fit, a possibility could be to use a gamma distribution: the exponential distribution can be thought of as the time it takes for performing one task, and the gamma distribution is then the time it takes to perform a number of identical tasks. The next, more general and, perhaps, most natural model would be a phase type distribution, i.e. the distribution of the time to absorption in a finite state Markov chain. In reliability applications, the Weibull distribution may instead be more useful. For the convenience of the reader, formulas for the gamma and Weibull distributions are given in Supplementary Materials.

2.2 Semi-parametric tail estimation

In this section we consider the situation where the tail of the distribution of the length of a 0-interval is at the center of interest.

Using all of the data and a full parametric model produces an estimated distribution which is mainly determined by the shape of the center of the distribution. This often leads to bad fit in the tail of the distribution. We hence here instead only assume a parametric model for the tails of F_0, i.e. for $F_0(x)$ for values of x greater than some threshold u, and use non-parametric Kaplan-Meier estimates of $F(x)$ for $x < u$. Specifically, we use the Peaks over Thresholds model with a generalized Pareto Distribution (GPD) for the excesses of u, see Coles (2001). We thus assume that

$$\bar{F}_0(x) = \bar{F}_0(u)\bar{G}_0(x - u), \quad \text{for} \quad x > u, \tag{4}$$

where the threshold $u < w$ is chosen large enough to make model fit acceptable. Here G_0 is a generalized Pareto cumulative distribution function,

$$G_0(y) = G_0(y;\sigma,\gamma) = 1 - \left(1 + \frac{\gamma}{\sigma}y\right)_+^{-1/\gamma}, \quad y > 0,$$

where the + signifies that the expression in parentheses should be replaced by 0 if it is negative. Thus the distribution has a finite right endpoint $\sigma/|\gamma|$ if $\gamma < 0$, and an infinite right endpoint otherwise. To ensure a finite mean, we assume that $\gamma < 1$. For $\gamma = 0$ the expression should be interpreted as its limit as $\gamma \to 0$,

$$G_0(y; \sigma, 0) = e^{-y/\sigma},$$

i.e. as an exponential distribution with scale parameter σ. Typically the choice of u is aided by diagnostic data plots. Coles (2001), Section 4.3 contains a discussion of data driven methods for this threshold choice, and of the ramifications surrounding it. Methods for checking model fit in the present situation are discussed in Section 2.3.

The only part of the observations to be used for estimation of σ and γ are the excesses $X = L - u$ of u. Thus we will use

nc) the $\bar{n}_{nc}$ excesses $x_{nc,1}, \dots x_{nc,\bar{n}_{nc}}$ of u by non-censored observations,

rc) the $\bar{n}_{rc}$ excesses $x_{rc,1}, \dots x_{rc,\bar{n}_{rc}}$ of u by right-censored observations,

lc) the $\bar{n}_{lc}$ excesses $x_{lc,1}, \dots x_{lc,\bar{n}_{lc}}$ of u by left-censored observations,

dc) the $\bar{n}_{dc}$ values $w - u$ which come from doubly-censored variables.

It follows from (4) that the cumulative distribution function of X, conditional on $L > u$, is G_0, and that the corresponding probability density function, for $\gamma \neq 0$, is

$$g_0(x; \sigma, \gamma) = \begin{cases} \sigma^{-1} \left(1 + \frac{\gamma}{\sigma} x\right)_+^{-1/\gamma - 1}, & \text{for } x > 0 \text{ and } \gamma \neq 0 \\ \sigma^{-1} e^{-x/\sigma}, & \text{for } x > 0 \text{ and } \gamma = 0. \end{cases}$$

By (1) and (4), the residual life density f_0^r has the form

$$f_0^r(x; \sigma, \gamma) = \bar{F}_0(u) \bar{G}(x - u; \sigma, \gamma)/\mu_0, \quad \text{for} \quad x > u,$$

and, for $\gamma \neq 0$, the probability density of the excess of a residual life time over u is

$$g_0^r(x; \sigma, \gamma) = \frac{\bar{F}_0(u) \bar{G}(x; \sigma, \gamma)/\mu_0}{\int_u^\infty \bar{F}_0(u) \bar{G}(y - u; \sigma, \gamma) dy/\mu_0} = \frac{1}{e(\sigma, \gamma)} \bar{G}(x; \sigma, \gamma), \tag{5}$$

for $x > 0$, where $e(\sigma, \gamma) = \int_0^\infty \bar{G}(y; \sigma, \gamma) dy = \sigma/(1 - \gamma)$ is the mean of the GPD. Integration then gives the residual life tail function

$$\bar{G}_0^r(x; \sigma, \gamma) = \left(1 + \frac{\gamma}{\sigma} x\right)_+^{-1/\gamma + 1}. \tag{6}$$

For $\gamma = 0$, instead $g_0^r(x) = \sigma^{-1} e^{-x/\sigma}$ and $\bar{G}_0^r(x) = e^{-x/\sigma}$.

In this model, the form of the cumulative distribution function for $x < u$ is supposed not to be connected with the parametric form assumed for $x \geq u$, and thus the numbers of excesses do not to contain any information about μ or σ. Hence we condition on $\bar{n}_{nc}, \bar{n}_{rc}, \bar{n}_{lc}, \bar{n}_{dc}$, and obtain the conditional log likelihood function

$$\begin{aligned} \ell_u(\sigma, \gamma) \quad = \quad & \sum_{k=1}^{\bar{n}_{nc}} \log g_0(x_{nc,k}; \sigma, \gamma) + \sum_{k=1}^{\bar{n}_{rc}} \log \bar{G}_0(x_{rc,k}; \sigma, \gamma) \\ & + \sum_{k=1}^{\bar{n}_{lc}} \log g_0^r(x_{lc,k}; \sigma, \gamma) + \bar{n}_{dc} \log \bar{G}_0^r(w - u; \sigma, \gamma). \end{aligned}$$

Thus, if $\gamma \neq 0$, then

$$\begin{aligned}\ell_u(\sigma,\gamma) &= -(1/\gamma+1)\sum_{k=1}^{\bar{n}_{nc}}\log\Big(1+\frac{\gamma}{\sigma}x_{nc,k}\Big) - 1/\gamma\sum_{k=1}^{\bar{n}_{rc}}\log\Big(1+\frac{\gamma}{\sigma}x_{rc,k}\Big)\\ &\quad -1/\gamma\sum_{k=1}^{\bar{n}_{lc}}\log\Big(1+\frac{\gamma}{\sigma}x_{lc,k}\Big) - \bar{n}_{dc}(1/\gamma-1)\log\Big(1+\frac{\gamma}{\sigma}(w-u)\Big)\\ &\quad -(\bar{n}_{nc}+\bar{n}_{lc})\log\sigma+\bar{n}_{lc}\log(1-\gamma). \qquad (7)\end{aligned}$$

For exponential case $\gamma = 0$ the log likelihood function, the estimate of σ and of its standard error is the same as in the example at the end of Section 2.1.

It remains to find an estimator of $\bar{F}_0(x)$ for $x \le u$, and in particular for $\bar{F}_0(u)$. For this we use the Kaplan-Meier estimator based on the non-censored and right-censored observations. (Cf. Gill and Keiding (2010), p. 576). Also left and doubly-censored observations contain information about $\bar{F}_0(u)$. However, to use them requires non-parametric estimation of a density function. This can only be done with much less precision, and using left and doubly-censored observations for non-parametric estimation is expected to add little to the precision of the estimate.

2.3 Complements

Confidence intervals: The simplest approach often is to use the inverse of the observed information matrix to estimate standard errors, and then to construct confidence intervals based on the assumption of normality. However, for some models bootstrap or parametric bootstrap methods may be simpler, and are also expected to lead to more accurate intervals if interest is centered at non-linear functions of the parameters, such as high quantiles.

Regression type modeling: Covariate dependence may be handled by making the parameters of the distribution be functions of the covariates. For non-censored observations and the Peaks over Thresholds model this is extensively discussed in, e.g., Coles (2001). In particular, models of the popular Accelerated Failure Time type may be obtained by letting the scale parameter σ in the generalized Pareto distribution depend on covariates $\mathbf{x} = (x_1, \ldots, x_d)'$ and parameters $\beta = (\beta_1, \ldots, \beta_d)$ as $\sigma = \exp(\beta\mathbf{x})$.

Goodness of fit plots: These are less standard. There are two types of observations of lengths of 0-intervals, a) left-censored and doubly-censored observations and b) non-censored and right-censored ones. We propose to do model control separately in two plots, one for the type a) lengths and one for type b) lengths, as follows. Let $\hat{\theta}$ be the estimate of the parameters. The first plot displays the empirical tail function of the type a) observations and the parametric estimate $\bar{F}_0^r(\cdot;\hat{\theta})$ of the length-biased tail function. The second plot shows the non-parametric Kaplan-Meier estimate of the tail function of the interval lengths for the type b) observations together with the corresponding parametric estimate $\bar{F}_0(\cdot;\hat{\theta})$.

The goodness of fit plots can also indicate whether the assumption of independence between the length of a starting 1-interval and the following 0-interval is reasonable, or if it is violated in ways which influence estimation. This assumption may also be checked by estimating the parameters separately from the type a) and the type b) observations, and comparing the results. If these estimates are similar it is an indication that the independence assumption is acceptable. Finally, making these plots for different thresholds u can be used to help choosing an appropriate u.

Maximum likelihood using the first two interval lengths: Write n_{no} for the number of observation windows which contain no observed 0-intervals, and write s_{lc}^{nc} and s_{lc}^{rc} for the observations from windows which start with $X(0) = 0$, then have a 0-interval which is shorter than w, and then an 1-interval

which is fully observed, or right-censored, respectively. Further, write n_{lc}^{nc} for the number of s_{lc}^{nc}-observations and n_{lc}^{rc} for the number of s_{lc}^{rc}-observations. Then, assuming θ also includes a fully parametric specification of F_1, and with obvious notation, the full likelihood for the first two observed interval lengths is

$$\begin{aligned} L(\theta) &= p_1^{n_{nc}+n_{rc}+n_{no}} \prod_{k=1}^{n_{nc}} f_1^r(s_{nc,k}) f_0(\ell_{nc,k};\theta) \times \prod_{k=1}^{n_{rc}} f_1^r(s_{rc,k}) \bar{F}_0(\ell_{rc,k};\theta) \times \bar{F}_1^r(w)^{n_{no}} \\ &\quad \times p_0^{n_{lc}^{nc}+n_{lc}^{rc}+n_{dc}} \prod_{k=1}^{n_{lc}^{nc}} f_0^r(\ell_{lc,k};\theta) f_1(s_{\ell c,k}^{nc};\theta) \times \prod_{k=1}^{n_{lc}^{rc}} f_0^r(\ell_{nc,k}) \bar{F}_1(s_{\ell c,k}^{rc};\theta) \times \bar{F}_0^r(w)^{n_{dc}}. \end{aligned}$$

Maximizing this likelihood gives an estimate of the parameters of the distribution of the lengths of both 0-intervals and 1-intervals. This approach is suitable for situations where the 0-intervals and 1-intervals are of similar or shorter lengths than the observation window.

Long windows: If the observation windows are long compared with the 0- and 1-intervals, there will often be several 0-intervals in an observation window, and it is wasteful to only use the first observed 0-interval. For the fully parametric specification of F_0 and an alternating renewal process the conditional likelihood (3) still applies if one uses all 0-intervals in the window, and not just the first one, and confidence intervals can be obtained from the observed information matrix. If there instead is dependence between the intervals in an observation window (3) is not a true conditional likelihood if all 0-intervals in the windows are used. But then (3) can instead be used as an estimating equation. This will provide consistent estimators, but standard deviations typically will be larger than those given by the inverse of the observed information matrix. However, since measurements from different observation windows are assumed to be independent, standard deviations and confidence intervals may be obtained using the sandwich method, or a block bootstrap with blocks equal to observation windows.

Unless windows are very long or there is a quite high dependence between long 0-intervals, it is rare that windows contain more than one 0-interval which is longer than the threshold u. Still, there might sometimes be windows which start with one or a few short 0-intervals, and then comes a long one. If one only uses the first 0-interval, such windows will not contribute to the estimation of the parameters of the generalized Pareto distribution. However, for such data sets one can instead of the first 0-interval use the first 0-interval, which exceeds the threshold u and consider (7) as an estimating equation. Again the resulting estimators are consistent, and confidence intervals may be obtained using block bootstrap.

Asymptotic normality: Using numerical computation, see Supplementary Materials, we show asymptotic normality of the GPD parameter estimates, for the most important case, $\gamma > 0$, when the distribution is heavytailed. For the exponential sub-model, $\gamma = 0$, asymptotic normality follows from standard results about right-censored observation of an exponential distribution. The asymptotics appropriate for the present problem is $n \to \infty$ with u and w fixed, rather than the extreme value type asymptotics where also u and w would tend to infinity. It may be noted that maximum likelihood estimation of GPD parameters is non-regular for $\gamma \leq -0.5$, see Drees et al. (2004).

3 Simulation study

In this section the small sample precision and coverage probabilities for confidence intervals are studied by simulation. The simulation was set up to resemble the visual inattention data discussed in the next section. In particular, throughout the observation window had length 6 s, and the 0-1 process was an alternating renewal process, with a mean 6 s exponential distribution for the length of the 1-intervals.

Further simulations, not included in the paper, produced very similar results for short (mean 0.1 s) 1-intervals. We used the MatLab fmincon minimization algorithm to find the ML-estimators. Confidence intervals were computed from the observed information matrix using a custom made MatLab function to compute the Hessian at the ML-estimates (MatLab fmincon estimates of the Hessian led to unsatisfactory results). For each choice of parametric and non-parametric distribution, and choice of N, the number of simulated glances that intersect the observation window w, we estimated the parameter(s) of the distribution from 10,000 replicates of the experiment.

For the fully parametric method the simulations were from exponential, gamma and Weibull distributions with means of the 0-intervals $\mu = 0.2, 1$, and 5, representing mild, medium and severe censoring, respectively.

For the semi-parametric tail estimation method we simulated from a mixture of a uniform $[0, 1]$ and (1+ a GPD) distributions. The mixing probabilities were 0.5 and 0.5, respectively. For the GPD distribution we used the shape parameters $-0.25, 0, 0.25$ and scale parameters $0.75, 1, 1.25$. These choices made the density continuous and led to the means $1.15, 1.25, 1.42$ for the mixture distributions. In estimation we used the threshold $u = 2$.

Table 1 shows that for the fully parametric exponential model the root mean square error (RMSE) was less than 15% of the true parameter value, except for samples of 50 observation windows for the most heavily censored case $\mu = \sigma = 5$. The coverage probabilities of the confidence intervals were close to the nominal value 0.95.

Table 1: Bias, standard deviation (STD), and root mean square error (RMSE) of the ML-estimator of σ in the full exponential model, and coverage probability (CP) for 95% confidence intervals.

	N	Bias	STD	RMSE	CP
$\mu = 0.2$	50	0.000	0.028	0.028	0.94
	250	0.000	0.013	0.013	0.95
	1000	0.000	0.006	0.006	0.95
$\mu = 1$	50	0.003	0.149	0.149	0.94
	250	0.000	0.067	0.067	0.95
	1000	0.000	0.034	0.034	0.95
$\mu = 5$	50	0.091	0.952	0.956	0.95
	250	0.018	0.416	0.416	0.95
	1000	0.007	0.205	0.206	0.95

For the gamma distribution the scale parameter was $k = 3$, roughly resembling the value in the visual inattention data. Since (with parametrization as given in Supplementary Materials) $\mu = k\sigma$ for the gamma distribution, σ was $0.067, 0.33$, and 1.67. From Table 2 it can be seen that for sample sizes 250 and 1000 the estimates of both parameters had RMSE-s which were smaller than 15% of the true parameter value, and that the coverage probabilities of confidence intervals were close to 95%.

For the Weibull simulation the shape parameter was $k = 1/2$, which leads to a heavier than exponential tail. Since $\mu = \sigma\Gamma(1 + 1/k)$ for this distribution, the values for σ were $0.1, 0.5$, and 2.5. The estimates of both parameters had RMSE less than 19% of the estimated parameter for sample sizes 250 and 1000. The coverage probabilities of the confidence intervals were close to 95% in all cases.

It was thus possible to find reasonable estimates even for the heavily censored case, $\mu = 5$ for some sample sizes. Nevertheless the method may still often not be practical in such cases, since the fit of the tail part of the model cannot be checked.

Table 2: Bias and root mean square error (RMSE) of the ML-estimators of k and σ for gamma and Weibull distributions, and coverage probability (CP) for 95% confidence intervals.

		N	Bias		RMSE		CP	
			$\hat{k}$	$\hat{\sigma}$	$\hat{k}$	$\hat{\sigma}$	$\hat{k}$	$\hat{\sigma}$
Gamma	$\mu = 0.2$	50	0.184	−0.001	0.677	0.014	0.96	0.92
		250	0.035	−0.000	0.267	0.006	0.95	0.94
		1000	0.007	−0.000	0.131	0.003	0.95	0.95
	$\mu = 1$	50	0.233	−0.008	0.794	0.076	0.95	0.90
		250	0.040	−0.001	0.293	0.033	0.95	0.95
		1000	0.011	−0.000	0.145	0.017	0.95	0.95
	$\mu = 5$	50	0.550	−0.038	2.285	0.589	0.96	0.89
		250	0.083	−0.007	0.462	0.253	0.95	0.94
		1000	0.020	−0.002	0.217	0.125	0.95	0.95
Weibull	$\mu = 0.2$	50	0.013	0.003	0.059	0.032	0.95	0.93
		250	0.002	0.001	0.024	0.014	0.95	0.95
		1000	0.001	0.000	0.012	0.007	0.95	0.95
	$\mu = 1$	50	0.010	0.022	0.058	0.170	0.95	0.93
		250	0.002	0.005	0.025	0.074	0.95	0.95
		1000	0.000	0.001	0.012	0.037	0.95	0.94
	$\mu = 5$	50	0.011	0.212	0.077	1.184	0.95	0.93
		250	0.002	0.042	0.032	0.477	0.95	0.95
		1000	0.001	0.013	0.016	0.237	0.95	0.95

For the semi-parametric tail estimation simulations, on average only around 11-18% of the observation windows contained an observed glance off road which was longer than the threshold $u = 2$ seconds, and hence, e.g., in the simulations with N=500 observation windows the estimators were based on roughly 55-90 off-road glances in excess of 2 seconds. Together with the information loss caused by censoring, this explains the rather low precision (in particular for $\gamma = -0.25$) for $N = 500$.

For samples consisting of $N = 2,500$ and $10,000$ glances, the RMSE-s of the γ-estimators were less than 0.07 and the RMSE-s of the σ-estimators were less than 10% of the true value, and the coverage probabilities of the confidence intervals were close to the nominal value, see Table 3.

Table 3: Bias and root mean square error (RMSE) of the semi-parametric tail parameter estimators of γ and σ, and coverage probability (CP) for 95% confidence intervals.

		N	Bias		RMSE		CP	
			$\hat{\gamma}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\sigma}$
Generalized Pareto	$\gamma = -0.25$, $\sigma = 0.75$	500	−0.078	0.061	0.211	0.207	0.89	0.94
		2500	−0.015	0.011	0.067	0.069	0.93	0.94
		10000	−0.004	0.003	0.030	0.032	0.94	0.95
	$\gamma = 0$, $\sigma = 1$	500	−0.020	0.027	0.169	0.260	0.89	0.89
		2500	−0.006	0.007	0.066	0.093	0.94	0.95
		10000	−0.002	0.001	0.032	0.045	0.95	0.95
	$\gamma = 0.25$, $\sigma = 1.25$	500	−0.019	0.043	0.135	0.284	0.94	0.95
		2500	−0.003	0.007	0.056	0.119	0.95	0.95
		10000	−0.001	0.002	0.028	0.059	0.95	0.95

Finally, except for the smallest sample size, the empirical distribution of the estimators were close to a normal distribution, both for the fully parametric models and for the semi-parametric models (plots not shown here).

4 Visual inattention in driving

In a naturalistic driving study ordinary cars with ordinary drivers are equipped with cameras which film driver behavior and the surrounding traffic; radars which measure the distance to road edges and other cars; GPS instruments; and sensors which measure things like brake and gas pedal actions. The vehicles are then used in everyday driving, just as if the instrumentation was not there, and driving behavior is recorded in extensive detail, both for normal driving and when accidents occur. In this section we analyze visual behavior in the 100-car naturalistic driving study (Wu and Jovanis (2012), data may be downloaded at http://forums.vtti.vt.edu). For 4,803 randomly chosen 6 second long observation windows obtained during normal driving, human annotators have used web camera recordings of the driver's face to construct a 6 second 0-1 process, where 0 means that the driver looks away from the road, and 1 that she looks at the road. Out of these windows, 2,602 contained at least one off-road glance.

In addition to individual off-road glance lengths, "task duration" periods which may include several off-road glances, such as the time period used to find a new program on the car radio, have important safety and economic consequences. Here we simply define a task as starting with an off-road glance and continuing until there is a on-road glance which is longer than 1 second (thus off-road glances separated by less than 1 second are joined). A more refined task duration analysis could (sometimes) be made by using further information from the annotation. However this is beyond the scope of the present analysis.

This section illustrates our methods by using them to find the (tail) distribution of off-road glances and task durations during normal driving. As a first step we used the method from Section 2.1 to fit gamma and Weibull distributions to the lengths of off-road glances and to task durations. The Weibull distribution gave visually a slightly better fit. The parameter estimates for the Weibull distribution were $\hat{k} = 1.39$ and $\hat{\sigma} = 0.99$ for the lengths of off-road glances, and $\hat{k} = 1.16$ and $\hat{\sigma} = 1.26$ for the task durations.

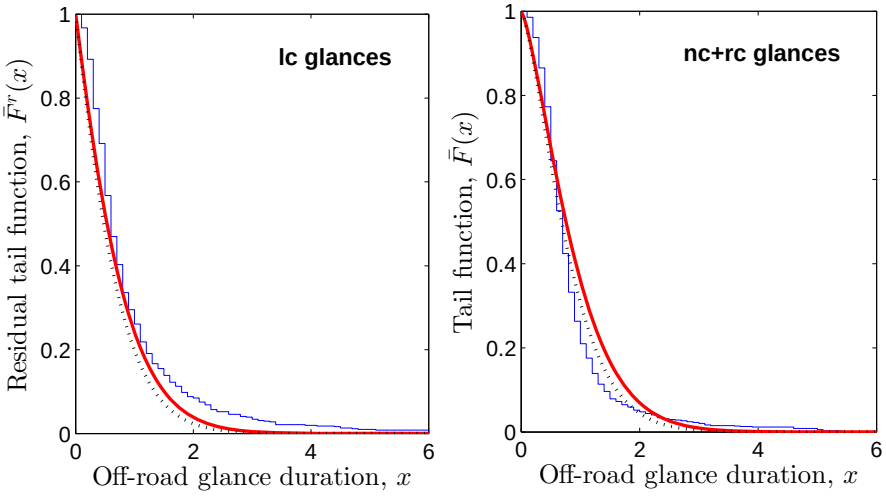

Figure 3: *Left:* Empirical tail function for left-censored off-road glances (jagged line) and fitted residual life Weibull tail function (smooth line). *Right:* Kaplan-Meier tail function estimate for non-censored plus right-censored off-road glances (jagged line) and fitted Weibull tail function (smooth line). *Left and Right:* Measurement resolution was 0.1 s. Dotted line is the fit obtained when not taking censoring and size-bias into account.

From the goodness of fit plots in Figure 3 it can be seen that the Weibull fit did not catch tail behavior well for the off-road glances. This was even more pronounced for the task durations (plots not shown here).

In the literature it is often stated that off-road glances longer than 1.8-2 seconds are dangerous, and hence we used the tail estimation method from Section 2.2 to fit a GPD distribution to the excesses of 2 seconds. There were all in all 124 off-road glances longer than 2 seconds, and the GPD parameter estimates were $\hat{\gamma} = 0.13$ and $\hat{\sigma} = 1.09$, with 95% confidence intervals $(-0.07, 0.33)$ and $(0.72, 1.46)$, respectively. Hence γ was not significantly different from 0. Fitting the model with $\gamma = 0$, i.e. assuming that the excess lengths of off-road glances longer that 2 seconds have an exponential distribution, gave an estimated value of $\sigma = 1.30$ with 95% confidence interval $(1.05, 1.57)$.

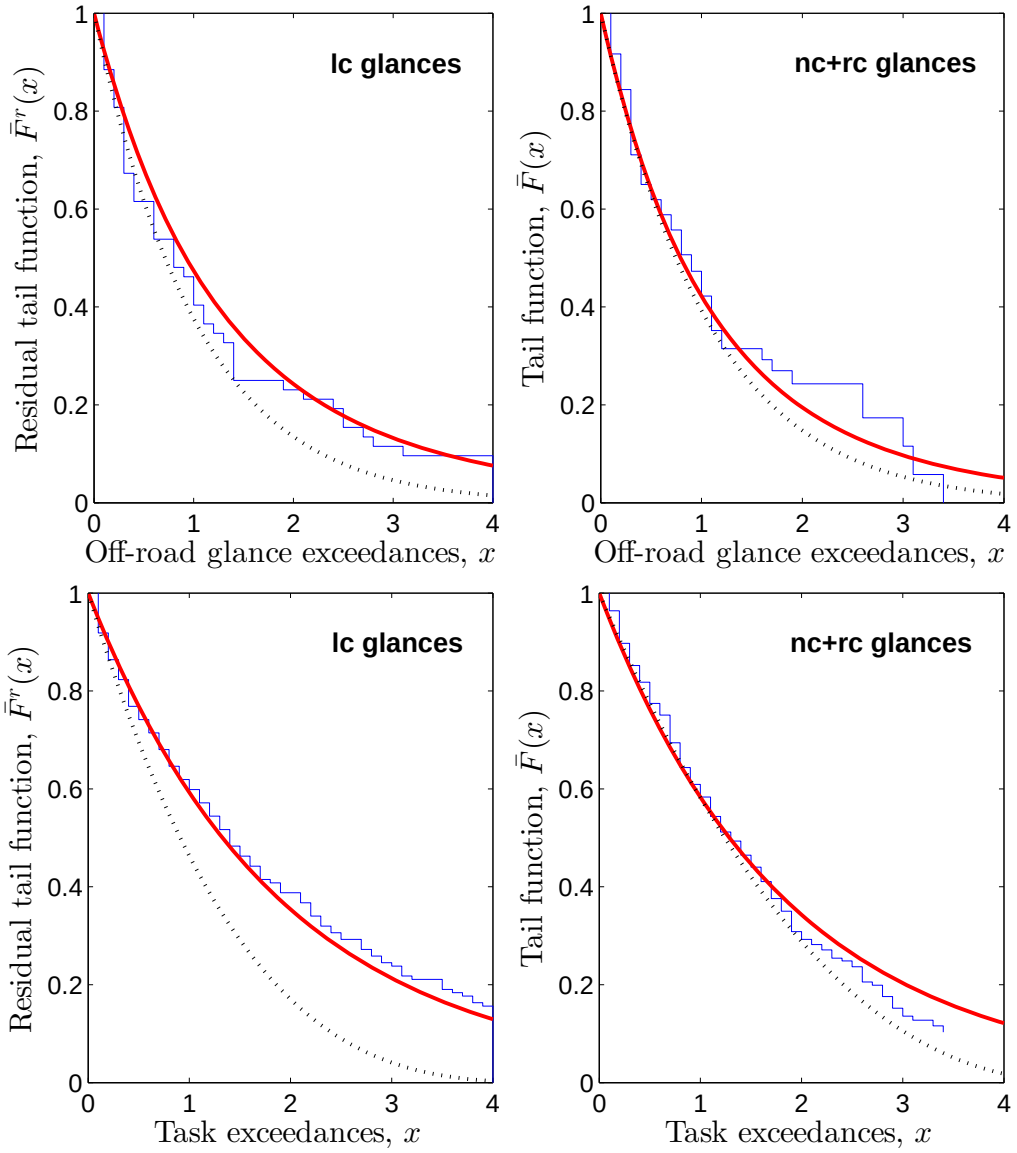

Figure 4: Excess lengths longer than 2 seconds. *Top row, left:* Empirical tail function for left-censored off-road glances (jagged line) and estimate from fitted residual life GPD distribution (smooth line). *Top row, right:* Kaplan-Meier estimate of tail function for non-censored plus right-censored off-road glances (jagged line) and estimate from fitted GPD (smooth line). *Bottom row:* the same plots for task durations. *All plots:* Measurement resolution was 0.1 s. Dotted line shows fit obtained when not taking censoring and size-bias into account.

The plots in Figure 4 show good fit of the GPD. The fit of the exponential distribution was almost identical. Further, the Kaplan-Meier estimate of the probability that an off-road glance was longer than 2 seconds, based on the non-censored and right-censored observations (Subsection 2.2), was 0.048, with a 95% confidence interval $(0.038, 0.058)$.

The same analysis for the task durations gave the GPD parameter estimates $\hat{\gamma} = 0.03$ and $\hat{\sigma} = 1.84$, with 95% confidence intervals $(-0.11, 0.17)$ and $(1.51, 2.17)$, respectively, based on 424 tasks longer than 2 seconds. This is almost exactly an exponential distribution. Figure 4 shows good fit. The Kaplan-Meier estimate of the probability that a task duration was longer than 2 seconds, based on the non-censored and right-censored observations, was 0.17, with a 95% confidence interval $(0.16, 0.19)$.

The conclusion of this analysis, that the excess length (=length - 2 seconds) of off-road glances and task durations both follow an exponential distribution seems both simple and useful to us. It is also somewhat surprising since we expect that glance behavior is different in different traffic situations and for different drivers. This presumably could lead to a relation between the length-biased distribution and the distribution itself which is different than for identically distributed observed 0-intervals. However, Figure 4 shows little indication of this (although Figure 3 perhaps points to such an effect). Still, the next step in the analysis of visual inattention will be to investigate how glance distributions depend on traffic situation or driver characteristic covariates. We will pursue this further in a paper directed at traffic safety research.

Finally, the goodness of fit plots show that the estimates which are obtained if one ignores censoring and size-bias can be quite bad.

Acknowledgements

We thank Niels Keiding, Olle Nerman, Jeff Steif, two referees and an associate editor for helpful comments. Research supported by the Knut and Alice Wallenberg foundation and by Vinnova.

Supplementary Materials

Main file: Contains exact expressions for residual life for gamma and Weibull distributions, additional results on asymptotic normality, and a brief description of numerical routines used throughout the analysis.

MatLab and Wolfram Mathematica scripts: Estimation algorithms, simulation study, analysis of 100-Car data, and numerical verification of asymptotic normality.

References

Alvarez, E. E. (2005). Smoothed nonparametric estimation in window censored semi-Markov processes. *J. Statist.Planning and Inference*, 131:209–229. 2, 4

Alvarez, E. E. (2006). Maximum likelihood estimation in alternating renewal processes under window censoring. *Stochastic Models*, 22:55–76. 2

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London. 5, 6, 7

Cox, D. (1969). On extreme values in stationary random fields. In Johnson, N. and Smith, H., editors, *New developments in survey sampling*, pages 506–527. Wiley, New York. 2

Drees, H., Ferreira, A., and de Haan, L. (2004). On maximum likelihood estimation of the extreme value index. *Ann. Appl. Probab.*, 14:1179–1201. 8

Einmahl, J., Fils-Villetard, A., and Guillou, A. (2008). Statistics of extremes under random censoring. *Bernoulli*, 14:207–227. 3

Gill, R. and Keiding, N. (2010). Product-limit estimators of the gap time distribution of a renewal process under different sampling patterns. *Lifetime Data Anal.*, 16:571–579. 2, 7

Karr, A. (1994). Estimation and reconstruction for zero-one Markov processes. *Stochastic Process. Appl.*, 16:219–255. 2

Ning, J., Qin, J., Asgharian, M., and Shen, Y. (2013). Empirical likelihood-based confidence intervals for length-biased data. *Statistics in medicine*, 32:2278–2291. 2

Ouhibi, B. and Limnios, N. (1999). Nonparametric estimation for semi-Markov processes based on its hazard rate functions. *Statist. Inf. Stoch. Proc.*, 2:151–173. 2

Qin, J., Ning, J., Liu, H., and Shen, Y. (2011). Maximum likelihood estimations and em algorithms with length-biased data. *Journal of the American Statistical Association*, 106(496):1434–1449. 2

Vardi, Y. (1982). Non-parametric estimation in the precence of length bias. *Ann. Statist.*, 10:616–620. 2

Vardi, Y. (1985). Empirical distribution in selection bias models. *Ann. Statist.*, 13:178–203. 2

Wu, K.-F. and Jovanis, P. (2012). Crashes and crash-surrogate events: Exploratory modeling with naturalistic driving data. *Accident Analysis and Prevention*, 45:507–516. 11

Zhao, Y. and Nagaraja, H. N. (2011). Fisher information in window censored renewal process data and its applications. *Ann. Inst. Stat. Math.*, 63:791–825. 2

Zhu, Y., Yashchin, E., and Hosking, J. R. M. (2014). Parametric estimation for window censored recurrence data. *Technometrics*, 56:55–66. 2