JIRSS (2016)
Vol. 15, No. 1, pp 1-28
DOI: 10.7508/jirss.2016.01.001

Estimation of the Conditional Survival Function of a Failure Time Given a Time-Varying Covariate with Interval-Censored Observations

M. Hossein Dehghan [1], **Thierry Duchesne** [2]

[1]Statistics Department, University of Sistan and Baluchestan, Iran
[2]Mathematics and Statistics Department, Laval University, Canada

Abstract. In this paper, we propose an approach for the nonparametric estimation of the conditional survival function of a time to failure given a time-varying covariate under interval-censoring for the failure time. Our strategy consists in modeling the covariate path with a random effects model, as is done in the degradation and joint longitudinal and survival data modeling literature, then in using a nonparametric estimator of the conditional survival function for time-fixed covariate. We derive the large sample bias and variance of the estimator under simplifying assumptions and we investigate its finite sample efficiency and robustness by simulation. We show how the proposed method can be useful in the early stages of data exploration and model specification by applying it to two real datasets, one on the time to infestation of trees by pine weevil and one on the reliability of a piece of electrical equipment. We conclude by suggesting avenues to make this data exploration method more suitable for formal inferences.

Keywords. Degradation; Generalized Kaplan-Meier estimator; Generalized Turnbull estimator; Joint modeling; Nadaraya-Watson estimator; Reliability.

MSC: Primary xx; Secondary xx.

M. Hossein Dehghan (✉)(Corresponding Author: mhdehghan@math.usb.ac.ir), Thierry Duchesne (thierry.duchesne@mat.ulaval.ca)

1 Introduction

In many fields of application of survival analysis, analysts have to build regression models to try and shed some light on the association between a time-to-event variable and some explanatory factor(s), or covariate(s), of interest. To make statistical inferences about specific features of this association, a variety of parametric and semi-parametric models are generally available. Even in the case where the time to event of interest is subject to interval censoring, a number of regression models have become available in recent years; see Sun (2006) for a review. Because different models may sometimes lead to different conclusions, one should try to choose a model that is, in some sense, compatible with the data at hand. To this end, nonparametric methods can be particularly useful, by steering the analyst in the proper direction at the early stage of data exploration. For example many survival analysis textbooks (Meeker and Escobar , 1998a; Lawless , 2003; Sun , 2006, for example) propose graphical tools based on nonparametric estimators of the conditional survival function that help the user in determining whether a proportional hazards or an accelerated failure time model should be fitted to the data. However most of these graphical tools are not available when the covariate is continuous and varies with time, or when the failure time variable is interval-censored. But this type of situation does occur in real applications. For instance in the application to electrical equipment reliability studied in Section 4.2, engineers typically want to build a parametric prediction model for the time to failure of the equipment as a function of the concentration of a certain gas in the oil of the units on the basis of a dataset where times to failure are interval censored and the gas concentration is measured on several occasions and varies with time (we also present an example in ecology in Section 4.1, and examples can easily be found in other areas such as health sciences or econometrics). Graphical tools that could help the analysts to identify an appropriate parametric form for the conditional distribution of time to failure given a gas concentration history could be helpful.

Our objective in this article is to propose a method that generalizes the different plotting procedures for interval censored data surveyed by Sun (2006, Section 10.2), so that they can handle covariates that are continuous and time-varying. In order to achieve our goal, we proceed in two steps: (i) reduce the dimensionality of the time-varying covariates and (ii) estimate nonparametrically the conditional survival function of the time to failure given the "reduced covariates". Several avenues are available to achieve step (i). Because by "high dimensionality of the covariates" here we do not mean the presence of a large number of covariates, but rather the presence of a few (often a single) covariates that are measured on many occasions, one possibility could be to use

functional principal components analysis (Ramsey and Silverman , 1997, Chap. 6) and then apply the nonparametric methods in step (ii) by using covariates as the coefficients of a basis function expansion. Unfortunately, as is often the case in survival data applications, time-varying covariates may be observed at irregular and infrequent times and this makes the application of functional principal components analysis problematic. This is particularly true in the case study presented in Section 4.2, where the times of the oil analyses tend to be extremely sparse for a long while, then frequent, then sparse again. We will instead use a strategy to handle the time-varying covariates that is more common in survival analysis contexts and that is advocated in the literature on degradation testing Lu et al. (1996), Meeker et al. (1998b) and on joint modeling of survival and longitudinal data (Tsiatis et al. , 1995; Hogan and Laird , 1997; Tseng et al. , 2005), where it is assumed that the covariates can be modeled as functions of time according to a mixed (random effects) model and the time to failure has a distribution that depends on unit-level "features" of this mixed model. As for step (ii), nonparametric estimation of a conditional distribution given continuous covariates has been studied extensively in certain contexts. When the response is observed exactly (no censoring), one may use the Nadaraya-Watson estimator (Nadaraya , 1964; Watson , 1964), while Beran (1981) introduced the generalized Kaplan-Meier (GKM) estimator when the response is subject to right-censoring. The estimator of Dehghan and Duchesne (2011a) is a generalization of the latter to the case of an interval-censored response. Hence the approach that we consider here is to apply the Dehghan and Duchesne (2011a) method using unit-level parameter estimates from the random effects as covariates.

The paper is organized as follows. In Section 2, we introduce the notation, state the required assumptions and describe the proposed estimation method. We derive some of its large sample properties under simplifying assumptions and investigate its finite sample properties in Section 3. We apply the method to the analysis of two real datasets in Section 4, one on the infestation of trees by pine weevils and one on the reliability of a piece of electrical equipment. Concluding remarks and thoughts for future research are given in Section 5.

2 Notation, assumptions and method

Suppose that for each of n independent units, we have a time to event $T_i > 0$ and a time-varying covariate $\bar{Z}_i = \{Z_i(s); s \geq 0\}$, $i = 1, \ldots, n$. We consider the case where the distribution of T_i is continuous and where we do not completely observe neither T_i nor $\bar{Z}_i$. For the latter, we assume that there is a sequence of observation times $\{t_{i1}, \ldots, t_{iN_i}\}$ that is independent of T_i and independent of $\bar{Z}_i$ where we observe $\{Z_i(t_{i1}), \ldots, Z_i(t_{iN_i})\}$.

For T_i, all we know is that it is in an interval $[L_i, R_i]$. For instance in the case study of Section 4.2, n pieces of equipment are followed. The oil of unit i will be analyzed at times $t_{i1}, \ldots, t_{iN_i}$, and the concentration of gas measured at time t_{ij} is denoted Z_{ij}. Here T_i is the time of equipment failure, and this failure time is not known exactly but rather known to have occurred somewhere in an interval $[L_i, R_i]$. Because the timing of oil analyses is mostly determined by the availability of staff and the amount of time elapsed since units in a certain geographic area have been inspected, the assumption that the sequence of inspection times is independent of $\bar{Z}_i$ and T_i appears reasonable.

To fully specify the model, we make the assumption that the censoring interval $[L_i, R_i]$ is determined according to the following mixed case censoring scheme: a censoring type variable δ_i takes on the value -1, 0 or 1 with probability π_{-1}, π_0 or $\pi_1 = 1 - \pi_{-1} - \pi_0$, respectively. If $\delta_i = -1$, then $L_i = R_i = T_i$ and the time to event is not censored. If $\delta_i = 0$, then a random right-censoring time C_i is independently of T_i generated, $L_i = \min(T_i, C_i)$, $R_i = +\infty$ if $T_i > C_i$, and $R_i = T_i$ if $T_i \leq C_i$. Finally if $\delta_i = 1$, then a (deterministic or stochastic) sequence of bracketing times $0 = v_{i0} < v_{i1} < v_{i2} < \cdots < v_{i,K_i+1} = +\infty$ independent of T_i is generated and $[L_i, R_i]$ are defined as $[v_{ij}, v_{i,j+1}]$ such that $v_{ij} \leq T_i \leq v_{i,j+1}$. Note that the sequences $\{t_{ij}; j = 1, \ldots, N_i\}$ and $\{v_{ij}; j = 1, \ldots, K_i\}$ may or may not be equal, depending on the application; they are equal in our application of Section 4.1 while they are distinct in our application of Section 4.2. Therefore, the observed data are n independent replications of the censoring intervals and observed covariate values, i.e., $\Big\{(L_i, R_i, Z_i(t_{i1}), \ldots, Z_i(t_{iN_i})), i = 1, \ldots, n\Big\}$.

In this paper, our aim is to propose a simple method to estimate the survival probability of T_i given one or a few key "features" of a potential covariate path $\bar{Z}_i$ that can be used in an early data exploration stage to get an idea of the form of the relationship between the distribution of T_i and the covariate. As is done in joint longitudinal and survival data analysis or in degradation modeling, we assume that as a function of time and given random coefficients $\beta_i^\top = (\beta_{i0}, \ldots, \beta_{ip})$, the covariate follows a linear model of the form

$$Z_i(t_{ij}) = \beta_i^\top \phi(t_{ij}) + \varepsilon_{ij}, \tag{2.1}$$

with

$$\phi(t_{ij})^\top = \Big(\phi_0(t_{ij}), \cdots, \phi_p(t_{ij})\Big),$$

where the $\phi_k(\cdot)$s are known functions. The survival model assumption is then that

$$P(T_i > t | \beta_i, \bar{Z}_i) = S(t|\beta_i), \tag{2.2}$$

meaning that the distribution of the time-to-event only depends on the covariate path through its "features" $\beta_{i0}, \ldots, \beta_{ip}$. We assume that the ε_{ij} are independent and identically

distributed (iid) according to a normal distribution with mean zero and variance σ_ε^2. This normality assumption is in agreement with the models presented in (Meeker and Escobar , 1998a, Chap. 13) in the case of reliability modeling and closely related to some of the models used in joint modeling of survival and longitudinal data, see for instance equations (1) and (2) in Tseng, Hsieh and Wang (2005). However the normality of the ε_{ij} can be relaxed, as what is really needed for our theoretical developments is approximate normality of estimates of the β_i. One interesting advantage of making an assumption such as (2.2) is that it allows nonparametric "prognostic" without the need to predict the future covariate path if it is not observed beyond a certain time point, in the sense that under (2.2), we get that

$$P(T_i > t | T_i > t_0, \bar{Z}_i) = \frac{S(t|\beta_i)}{S(t_0|\beta_i)}. \tag{2.3}$$

In other words, to estimate $P(T_i > t | T_i > t_0, \bar{Z}_i)$, an estimate of β_i and of $S(\cdot|\cdot)$ will suffice.

Because we want the method to be as close as possible to nonparametric, we minimize the number of model assumptions by treating the β_i as n unknown parameters in our estimation procedure. Therefore each β_i is estimated by, say, $\hat{\beta}_i^\top = (\hat{\beta}_{i0}, \ldots, \hat{\beta}_{ip})$, separately for each sample $\{t_{ij}, Z_i(t_{ij}), j = 1, \ldots, N_i\}$. One can trade off some robustness for efficiency by adding the usual assumptions of a mixed model, for example by supposing that the β_is are multivariate zero mean normal random variables and then by getting best linear predictors of the β_i; we do not pursue this here as our primary objective is to carry out completely nonparametric estimation of $S(t|\beta)$. For our purposes, we shall only require that $E(\hat{\beta}|\beta) = \beta$ and $\mathrm{Var}(\hat{\beta}|\beta) = \sigma_N^2$ such that, as $N \to \infty$, $\sigma_N \to 0$ and $N\sigma_N \to \infty$. If one were to relax the normality assumption on ε_{ij} in (2.1), then asymptotic normality of $\hat{\beta}$ would also be required.

Now let us consider estimation of $S(t|\beta)$. Dehghan and Duchesne (2011a) proposed a nonparametric estimator of this conditional survival function when the time to event is subject to the interval-censoring scheme defined above and when the covariate is continuous, which they refer to as generalized Turbull (GT) estimator. In this paper, we investigate whether the GT estimator applied to the sample $\left\{(L_i, R_i, \hat{\beta}_i), i = 1, \ldots, n\right\}$ is a good estimator of $S(t|\beta)$. The GT estimator is actually a generalization of the Nadaraya-Watson (NW) estimator (Nadaraya , 1964; Watson , 1964) of a regression function to the case where the response variable is subject to interval-censoring. Indeed, if $1_{\{A\}}$ denotes the indicator of A, then $S(t|\beta) = E\left[1_{\{T>t\}}|\beta\right]$ and thus when the T_i are observed we can use the NW estimator

$$\hat{S}_{NW}^h(t|\beta) = \sum_{i=1}^n \gamma_i w_i^h(\beta),$$

with $\gamma_i = 1_{\{T_i>t\}}$, h a smoothing parameter (bandwidth) and the weight $w_i^h(\beta)$ typically given by

$$w_i^h(\beta) = \frac{1}{h}K\left(\frac{\hat{\beta}_i - \beta}{h}\right) \Big/ \sum_{\ell=1}^{n} \frac{1}{h}K\left(\frac{\hat{\beta}_\ell - \beta}{h}\right), \tag{2.4}$$

where $K(\cdot)$ is a kernel function, usually a continuous density symmetric about zero. In theory, the NW estimator will be consistent regardless of the dimension of β. In practice, however, the choice of h rapidly becomes difficult when $\dim(\beta) > 1$. Because of this and for ease of notation, we shall hereafter consider that β is unidimensional, with the understanding that the method can readily be extended to a multidimensional β by using a multidimensional kernel function and bandwidth. Because the only covariate in the model for Z_i is time, it will usually be the case that β_i is of low dimension.

Beran (1981) extended the NW estimator to the case where the time-to-event is subject to right-censoring. He proposed the estimator, later referred to as generalized Kaplan-Meier (GKM) estimator, and defined as

$$\hat{S}^h_{GKM}(t|\beta_0) = \prod_{\{i:L_i\le t\}} \left(1 - \frac{1_{\{R_i<+\infty\}} w_i^h(\beta_0)}{\sum_{k:L_k\ge L_i} w_k^h(\beta_0)}\right).$$

The generalization of the NW and GKM estimators to the case of interval-censored data is studied in detail by Dehghan and Duchesne (2011a,b), who also provide the R package "gte" to implement their method (Dehghan et al. , 2015). Here we simply outline the algorithm that can be used to obtain this estimator. Let $0 = \tau_0 < \tau_1 < \tau_2 < ... < \tau_g$, $g \le 2n$, be the ordered distinct values of $\{L_i, R_i, i = 1, \ldots, n\}$. Set $B_j = (\tau_{j-1}, \tau_j)$, $j = 1, \ldots g$ and let $p_j(\beta)$ be the conditional probability of T_i being in B_j given that $\beta_i = \beta$.

Algorithm 1 (Generalized Turnbull estimator)**.**

Step 1 : *Set* $r = 0$ *and* $p_j^{(0)}(\beta) = 1/g$, $j = 1, \ldots, g$.

Step 2 : *Set* $r = r + 1$. *For* $j = 1, \ldots, g$, *put*

$$p_j^{(r)}(\beta_0) = \sum_{i=1}^{n} w_i^h(\beta) \frac{1\{B_j \subseteq (L_i, R_i)\} p_j^{(r-1)}(\beta)}{\sum_{k=1}^{g} 1\{B_k \subseteq (L_i, R_i)\} p_k^{(r-1)}(\beta)}, \quad j = 1, \ldots, g,$$

with $w_i^h(\beta)$ *given by (2.4).*

Step 3 : *Return to Step 2 until* $\max_j |p_j^{(r)}(\beta) - p_j^{(r-1)}(\beta)| \le \varepsilon$ *for some small number* $\varepsilon > 0$.

The algorithm above does not specify how to distribute the probability masses $p_j(\beta)$ within the intervals B_j as the data contain no information about this aspect of the distribution (Turnbull , 1976). We can thus define the following estimator by assigning all the probability masses to the left hand points of the intervals:

$$\hat{S}^h_{GTL}(t|\beta) = \begin{cases} 1, & t < \tau_1 \\ \sum\limits_{j:\tau_j > t} \hat{p}^h_j(\beta), & t \geq \tau_1. \end{cases}$$

We could also assign all the probability masses to the right hand points of the intervals and obtain the following estimator:

$$\hat{S}^h_{GTR}(t|\beta) = 1 - \sum_{j:\tau_j < t} \hat{p}^h_j(\beta).$$

Yet another way of distributing the probability masses could be to use the algorithm proposed by Li et al. (1997) whereby this distribution of the masses depends on the choice of the initial survival function used at the initial step of the algorithm. Due to its simplicity and good behavior in simulation studies, we shall only consider $\hat{S}^h_{GTL}$ in the sequel. As for the value of h, Dehghan and Duchesne (2011a) proposed a simple formula that works well when $K(\cdot)$ is the Gaussian kernel and the data are purely interval-censored ($P(\delta_i = 1) = 1$), and they propose a cross-validation method to set the value of h in more general setups.

To sum up, we suppose that $Z_i(t_{ij}) = \beta_i^\top \phi(t_{ij}) + \varepsilon_{ij}$ and $P(T_i > t|\{Z_i(s), s \geq 0\}) = S(t|\beta_i)$, and our goal is to get a nonparametric estimate of $S(\cdot|\beta)$ for a fixed value of β. To do so, we estimate β_i by an unbiased estimator $\hat{\beta}_i$ computed from the sample $\left\{(t_{ij}, Z_i(t_{ij})), j = 1, \ldots, N_i\right\}$. We then apply the GT estimator of $S(t|\beta)$ to the sample $\left\{(L_i, R_i, \hat{\beta}_i), i = 1, \ldots, n\right\}$.

3 Asymptotic properties of the estimator

3.1 Large sample properties

We first consider the large sample bias and variance of the proposed estimator. For a time-fixed covariate, the asymptotic bias and variance of the estimator are well known when there is no censoring (because it then reduces to the NW estimator) and have been obtained by Dabrowska (1987) when the response is subject to right-censoring (in which the GT estimator reduces to the GKM estimator). Under interval-censoring, the asymptotic behavior of the GT estimator is still an open problem. For time-varying

covariates, no such properties have been obtained. There are, however, results on NW-type estimators when the covariate is measured with error (Fan and Truong , 1993; Carroll et al. , 1999; Schennach , 2004). To obtain asymptotic results in this case, assumptions have to be made on the distribution of the observed and true values of the covariates (respectively the $\hat{\beta}_i$ and β_i in our case). When the variance of the error in the covariate is fixed, a convolution kernel must be used in order to obtain consistent estimates of the survival function. In our case, however, the variance of the error in the covariate, $\sigma^2_{N_i}$, diminishes as N_i increases, which allows for consistent estimation with the ordinary NW estimator. Theorem 3.1 below, whose proof is outlined in the Appendix, states the precise result.

Theorem 3.1. *Let h_n be a deterministic bandwidth value such that as $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$. Let $\beta_1, \beta_2, \ldots$ be an iid sample from a continuous distribution with density f_β. Given β, suppose that $\hat{\beta} \sim N(\beta, \sigma^2_N)$, where $\sigma^2_N \to 0$ and $N\sigma_N \to \infty$ as $N \to \infty$. Let $f_{\hat{\beta}}$ denotes the marginal (unconditional) density of $\hat{\beta}$. Assume that $S(t|\beta)$ is three times continuously differentiable with respect to β, uniformly in t, and put $\hat{S}(t|\beta_0) = \sum_{i=1}^n w_i^h(\beta_0)\gamma_i$, with w_i^h given by (2.4) and $K(u) = (2\pi)^{-1/2}\exp(-u^2/2)$, $u \in \mathbb{R}$, the Gaussian kernel. Define $\int_{-\infty}^{\infty} u^2K(u)du = \mu_2$ and $\int_{-\infty}^{\infty} u^jK^\ell(u)du = \kappa_j^\ell$. Then the asymptotic bias of $\hat{S}(t|\beta_0)$ is given by*

$$\frac{(h_n^2 + \sigma_N^2)\mu_2}{2f_{\hat{\beta}}(\beta_0)}\left\{2S^{01}(t|\beta_0)f_\beta^{(1)}(\beta_0) + S^{02}(t|\beta_0)f_\beta(\beta_0)\right\} + O(h_n^\ell \sigma_N^{\ell'} \big| \ell + \ell' \geq 4), \tag{3.1}$$

where $S^{jk}(x|y) = \partial^{j+k}S(x|y)/\partial x^j y^k$, $f_\beta^{(j)}(\beta) = \partial^j f_\beta(\beta)/\partial\beta^j$ and $O(x^a y^b | a + b \geq m)$ denotes a polynomial expansion in x and y where the terms $x^a y^b$ all have $a + b \geq m$.

The asymptotic variance is obtained as

$$\frac{\sigma^2_{S(t|\beta_0)}}{nh_n f^2_{\hat{\beta}}(\beta_0)}\kappa_0^2 f_\beta(\beta_0) + n^{-1}O(\sigma_N^\ell h_n^{\ell'} \big| \ell \geq 2; \ell' \geq -1), \tag{3.2}$$

with $\sigma^2_{S(t|\beta_0)} = S(t|\beta_0)\{1 - S(t|\beta_0)\}$.

The bias and variance given by (3.1) and (3.2) are quite similar to the asymptotic bias and variance of the NW estimator without error in covariates and a random sampling design for the covariate. The only difference here is the presence of the variance σ^2_N and of $f_{\hat{\beta}}$ in expressions (3.1) and (3.2). As a matter of fact, if the β are observed instead of estimated, then $\sigma^2_N = 0$, $f_{\hat{\beta}} = f_\beta$ and the expressions (3.1) and (3.2) reduce to the usual expressions for the NW estimator.

Note that the conditions imposed on σ_N^2 are not unreasonable in practice. For instance, if one has $Z_i(t_{ij}) = \beta_i t_{ij} + \varepsilon_{ij}$ and the observations are taken at times t_{ij} evenly spaced in a finite interval $[0, \tau]$, then it is easy to check that, as $N \to \infty$, the variance of the least squares estimator of β will obey the conditions set on σ_N^2 in Theorem 3.1; these calculations are also done in Appendix A.

3.2 Finite sample properties

We now investigate the integrated mean squared error (IMSE) and robustness of the proposed estimator under various finite sample scenarios. In particular, we investigate how the IMSE of the estimator varies as a function of the variance of the error terms ε_{ij} in the linear model for the covariate, of the point of estimation β, of the width of the censoring intervals and of the sample size. We also study the robustness of the estimator to misspecification of the linear model for the covariate and we look at whether the bootstrap algorithm can give reasonable estimates of the standard error of the estimator. To put the results obtained into perspective, we do as Dehghan and Duchesne (2011b) did and compare the performance of the estimator $\hat{S}^h_{GTL}$ to that of the NW (pure interval-censoring) or GKM (mixed case censoring) estimators used with midpoint imputation of the interval-censored times. To ensure valid comparisons between these estimators, for each of them we use the optimal bandwidth, as determined by pilot simulations.

As hinted above, we consider two censoring schemes in the simulations. What we shall refer to as "pure interval-censoring" corresponds to the mixed case censoring scheme described at the beginning of Section 2 with $\pi_{-1} = \pi_0 = 0$ and $\pi_1 = 1$. The other scheme that we consider is called "hybrid censoring", for which we set $\pi_1 = 0.60$, $\pi_0 = 0.40$ and $\pi_{-1} = 0.00$. For interval-censored observations the bracketing time sequences $v_{i1}, v_{i2}, \ldots$ are generated according to a homogeneous Poisson process with rate ρ. The latter parameter controls the width of the censoring intervals, as the lower the value of ρ, the more sparse the bracketing times and thus the wider the censoring intervals. Plots of typical samples obtained with different values of ρ are given in the Supplementary Material. Except when investigating the robustness of the method, we generate the covariates from a linear model without intercept, i.e., $Z_{ij} = \beta_i t_{ij} + \varepsilon_{ij}$, with the ε_{ij} iid $N(0, \sigma_\varepsilon^2)$. We generate the β_i so that the covariate paths are randomly distributed over the positive quadrant. More precisely, we simulate α_i from a uniform distribution on $(0.05, \pi/2 - 0.05)$ and then set $\beta_i = \tan \alpha_i$. As for the failure times T_i, we suppose that, given β_i, they follow a Weibull distribution with shape 3 and scale $1 + 10/\beta_i$. This specification of the scale parameter has been determined so as to ensure a non-negative scale parameter and a survival that tends to diminish when Z_{ij}

increases more steeply, which emulates what we seem to be observing in our real data applications.

In the sequel, we use the method described in Section 2 with the least squares estimator $\hat{\beta}_i = \sum_j Z_i(t_{ij})t_{ij} / \sum_j t_{ij}^2$ and the Gaussian kernel to estimate $S(t|\beta_0)$ for various values of β_0. We apply the method to samples of the form $\left\{\left(L_i, R_i, Z_i(t_{i1}), \ldots, Z_i(t_{iN_i})\right), i = 1, \ldots, n\right\}$, with n alternating between 50 and 100. Each simulation scenario is replicated 5,000 times.

As can be seen from Tables 1 and 2, the effect of the parameters is in line with what one would expect. The IMSE of the estimators increases with the variability in the error terms of the linear model and decreases when the size of the sample increases. In the case of purely interval-censored data, the IMSE also increases with the width of the censoring intervals. We also have that the estimators perform better when estimating $S(t|\beta_0)$ at a value of β_0 close to the center of the distribution of the β_i. As for the comparison between methods, as Dehghan and Duchesne (2011b) found, the GT estimator is much, much better than the NW estimator when the censoring intervals are large ($\rho = 0.01$) due to serious bias problems of the latter in this case. However the NW estimator seems to have a slight edge over the GT estimator under pure interval-censoring with short intervals. The picture is completely different under hybrid censoring, however, where the GT performs well under all conditions but not the GKM, which is badly biased all around.

To investigate the robustness of the method, we repeated the previous study, but we simulated the $Z_i(t_{ij})$ from the model $Z_i(t_{ij}) = \beta_i t_{ij} + \lambda\phi(t_{ij}) + \varepsilon_{ij}$, for various values of λ and specifications of $\phi(\cdot)$. However, when estimating $S(t|\beta_0)$, we wrongly assumed the model $Z_i(t_{ij}) = \beta_i t_{ij} + \varepsilon_{ij}$ and thus kept on using $\hat{\beta}_i = \sum_j Z_i(t_{ij})t_{ij} / \sum_j t_{ij}^2$.

Tables 3 and 4 below report the results for $\phi(x) = \log x$, $\sigma_\varepsilon^2 = 1$ and various values of λ. If we compare with the rows corresponding to $\sigma_\varepsilon^2 = 1$ in Tables 1 and 2, we can see that, when λ is small or moderate, the price to pay in terms of IMSE is small, which suggests an interesting robustness property of the estimator. Of course when the value of λ is more extreme, then the IMSE can get multiplied by four or more. However, in these cases, it should be fairly easy to see from the data that a linear model for $Z_i(t_{ij})$ as a function of t_{ij} is inappropriate. Results with another specification of $\phi(\cdot)$ lead to similar conclusions and are presented in the Supplementary Material.

We have also assessed whether we could get an idea of the variance of $\hat{S}(t|\beta_0)$ by applying the bootstrap algorithm proposed by Dehghan and Duchesne (2011a). Tables 5 and 6 compare the sample variance to the average bootstrap variance estimate of $\hat{S}(t|\beta_0 = 1.5)$ for various values of t. The results are outlined in Table 5 for simulations

Table 1: IMSE of the nonparametric estimators of $S\left(t|\beta_0\right)$ for various values of the point estimation β_0, of the variance σ_ε^2 of the error terms in the covariate model, of the rate ρ of the homogeneous Poisson process that generates the interval-censoring times, and of the sample size n under the pure interval-censoring scheme. Each reported IMSE value is calculated from 5,000 replications. GT refers to estimation with $\hat{S}^h_{GTL}$ while NW refers to estimation with $\hat{S}^h_{NW}$ and midpoint imputation of the interval-censored times.

			$\beta_0 = 0.6$		$\beta_0 = 1.5$		$\beta_0 = 2.5$	
n	ρ	σ_ε	GT	NW	GT	NW	GT	NW
50	0.01	0.2	0.0192	0.0549	0.0139	0.0973	0.0163	0.1144
		1	0.0194	0.0548	0.0142	0.0972	0.0169	0.1186
		3	0.0196	0.0555	0.0158	0.1018	0.0205	0.1280
		5	0.0197	0.0561	0.0186	0.1115	0.0247	0.1373
		7	0.0198	0.0568	0.0216	0.1214	0.0282	0.1451
	0.1	0.2	0.0082	0.0058	0.0032	0.0025	0.0038	0.0026
		1	0.0083	0.0060	0.0033	0.0026	0.0043	0.0030
		3	0.0094	0.0072	0.0038	0.0029	0.0064	0.0056
		5	0.0105	0.0082	0.0051	0.0043	0.0090	0.0083
		7	0.0111	0.0090	0.0069	0.0063	0.0118	0.0111
	0.2	0.2	0.0067	0.0052	0.0025	0.0018	0.0027	0.0016
		1	0.0069	0.0054	0.0025	0.0019	0.0031	0.0019
		3	0.0078	0.0066	0.0030	0.0022	0.0053	0.0038
		5	0.0088	0.0076	0.0040	0.0031	0.0076	0.0063
		7	0.0093	0.0080	0.0055	0.0045	0.0101	0.0086
100	0.01	0.2	0.0131	0.0666	0.0077	0.1095	0.0096	0.1265
		1	0.0134	0.0667	0.0079	0.1099	0.0102	0.1316
		3	0.0142	0.0670	0.0089	0.1130	0.0128	0.1421
		5	0.0145	0.0680	0.0109	0.1220	0.0160	0.1519
		7	0.0145	0.0684	0.0131	0.1323	0.0194	0.1592
	0.1	0.2	0.0048	0.0035	0.0021	0.0016	0.0022	0.0018
		1	0.0049	0.0036	0.0021	0.0016	0.0026	0.0023
		3	0.0061	0.0048	0.0025	0.0019	0.0047	0.0046
		5	0.0072	0.0057	0.0035	0.0031	0.0072	0.0074
		7	0.0077	0.0062	0.0050	0.0047	0.0098	0.0102
	0.2	0.2	0.0038	0.0029	0.0016	0.0011	0.0015	0.0009
		1	0.0039	0.0031	0.0016	0.0011	0.0019	0.0012
		3	0.0050	0.0043	0.0019	0.0012	0.0039	0.0030
		5	0.0060	0.0052	0.0027	0.0019	0.0061	0.0050
		7	0.0064	0.0056	0.0039	0.0031	0.0085	0.0073

Table 2: IMSE of the nonparametric estimators of $S\left(t|\beta_0\right)$ for various values of the point of estimation β_0, of the variance σ_ε^2 of the error terms in the covariate model, of the rate ρ of the homogeneous Poisson process that generates the interval-censoring times, and of the sample size n under the hybrid censoring scheme. Each IMSE value reported is calculated from 5,000 replications. GT refers to estimation with $\hat{S}^h_{GTL}$ while GKM refers to estimation with $\hat{S}^h_{GKM}$ and midpoint imputation of the interval-censored times.

			$\beta_0 = 0.6$		$\beta_0 = 1.5$		$\beta_0 = 2.5$	
n	ρ	σ_ε	GT	GKM	GT	GKM	GT	GKM
50	0.01	0.2	0.0246	0.1647	0.0249	0.2300	0.0306	0.2562
		1	0.0246	0.1652	0.0252	0.2302	0.0312	0.2575
		3	0.0246	0.1668	0.0286	0.2363	0.0353	0.2651
		5	0.0246	0.1662	0.0340	0.2425	0.0400	0.2735
		7	0.0247	0.1663	0.0384	0.2508	0.0454	0.2758
	0.1	0.2	0.0207	0.1112	0.0087	0.1297	0.0100	0.1398
		1	0.0208	0.1114	0.0091	0.1303	0.0108	0.1401
		3	0.0208	0.1116	0.0112	0.1330	0.0145	0.1445
		5	0.0210	0.1118	0.0150	0.1370	0.0190	0.1501
		7	0.0210	0.1122	0.0190	0.1410	0.0238	0.1532
	0.2	0.2	0.0222	0.1122	0.0090	0.1283	0.0094	0.1378
		1	0.0225	0.1122	0.0093	0.1285	0.0103	0.1384
		3	0.0225	0.1122	0.0116	0.1319	0.0139	0.1429
		5	0.0225	0.1124	0.0153	0.1350	0.0182	0.1468
		7	0.0228	0.1126	0.0192	0.1396	0.0223	0.1505
100	0.01	0.2	0.0185	0.1752	0.0378	0.2626	0.0546	0.2910
		1	0.0187	0.1758	0.0383	0.2626	0.0590	0.2947
		3	0.0188	0.1767	0.0398	0.2628	0.0598	0.2986
		5	0.0197	0.1772	0.0430	0.2666	0.0640	0.3016
		7	0.0206	0.1796	0.0469	0.2695	0.0684	0.3026
	0.1	0.2	0.0183	0.1103	0.0204	0.1456	0.0227	0.1544
		1	0.0186	0.1108	0.0360	0.1566	0.0313	0.1638
		3	0.0190	0.1114	0.0378	0.1577	0.0491	0.1773
		5	0.0204	0.1133	0.0419	0.1611	0.0606	0.1853
		7	0.0210	0.1141	0.0408	0.1625	0.0615	0.1874
	0.2	0.2	0.0199	0.1109	0.0148	0.1386	0.0165	0.1468
		1	0.0201	0.1113	0.0310	0.1505	0.0244	0.1530
		3	0.0206	0.1117	0.0418	0.1559	0.0436	0.1684
		5	0.0229	0.1140	0.0464	0.1597	0.0654	0.1829
		7	0.0237	0.1150	0.0445	0.1617	0.0558	0.1857

Table 3: IMSE of the nonparametric estimators of $S\left(t|\beta_0\right)$ for various values of λ in the $Z(t_{ij}) = \beta_i t_{ij} + \lambda \log(t_{ij}) + \varepsilon_{ij}$, three values of ρ, two values of n, $\sigma_\varepsilon^2 = 1$ under pure interval-censoring. Each IMSE value is calculated from 5,000 replications. GT refers to estimation with $\hat{S}^h_{GTL}$ while NW refers to estimation with $\hat{S}^h_{NW}$ and midpoint imputation of the interval-censored times.

			$\beta_0 = 0.6$		$\beta_0 = 1.5$		$\beta_0 = 2.5$	
n	ρ	λ	GT	NW	GT	NW	GT	NW
50	0.01	0.1	0.0218	0.0557	0.0250	0.1202	0.0383	0.1433
		0.5	0.0219	0.0565	0.0254	0.1195	0.0406	0.1466
		1	0.0222	0.0561	0.0267	0.1230	0.0418	0.1503
		3	0.0222	0.0595	0.0299	0.1262	0.0455	0.1544
		5	0.0222	0.0613	0.0315	0.1282	0.0487	0.1580
	0.1	0.1	0.0140	0.0119	0.0088	0.0086	0.0120	0.0122
		0.5	0.0147	0.0125	0.0128	0.0132	0.0166	0.0167
		1	0.0153	0.0130	0.0136	0.0140	0.0250	0.0258
		3	0.0162	0.0136	0.0164	0.0170	0.0286	0.0298
		5	0.0170	0.0141	0.0173	0.0181	0.0301	0.0315
	0.2	0.1	0.0120	0.0109	0.0055	0.0047	0.0078	0.0067
		0.5	0.0138	0.0127	0.0119	0.0107	0.0116	0.0104
		1	0.0144	0.0134	0.0127	0.0121	0.0215	0.0200
		3	0.0154	0.0145	0.0157	0.0150	0.0276	0.0268
		5	0.0163	0.0153	0.0175	0.0169	0.0304	0.0296
100	0.01	0.1	0.0176	0.0674	0.0175	0.1317	0.0290	0.1561
		0.5	0.0180	0.0676	0.0181	0.1346	0.0311	0.1597
		1	0.0180	0.0690	0.0188	0.1358	0.0324	0.1650
		3	0.0181	0.0706	0.0214	0.1389	0.0358	0.1704
		5	0.0182	0.0735	0.0239	0.1428	0.0392	0.1734
	0.1	0.1	0.0114	0.0101	0.0069	0.0072	0.0102	0.0111
		0.5	0.0126	0.0111	0.0113	0.0126	0.0143	0.0151
		1	0.0132	0.0115	0.0120	0.0134	0.0232	0.0241
		3	0.0142	0.0121	0.0144	0.0160	0.0266	0.0291
		5	0.0151	0.0126	0.0153	0.0171	0.0281	0.0309
	0.2	0.1	0.0093	0.0084	0.0039	0.0034	0.0063	0.0058
		0.5	0.0118	0.0111	0.0096	0.0092	0.0099	0.0094
		1	0.0126	0.0119	0.0115	0.0114	0.0190	0.0184
		3	0.0138	0.0130	0.0141	0.0141	0.0262	0.0263
		5	0.0146	0.0138	0.0157	0.0157	0.0269	0.0282

Table 4: IMSE of the nonparametric estimators of $S\left(t|\beta_0\right)$ for various values of λ in the $Z(t_{ij}) = \beta_i t_{ij} + \lambda \log(t_{ij}) + \varepsilon_{ij}$, three values of ρ, two values of n, $\sigma^2_\varepsilon = 1$ under hybrid censoring. Each IMSE value is calculated from 5,000 replications. GT refers to estimation with $\hat{S}^h_{GTL}$ while GKM refers to estimation with $\hat{S}^h_{GKM}$ and midpoint imputation of the interval-censored times.

			$\beta_0 = 0.6$		$\beta_0 = 1.5$		$\beta_0 = 2.5$	
n	ρ	λ	GT	GKM	GT	GKM	GT	GKM
50	0.01	0.1	0.0256	0.1663	0.0284	0.1709	0.0556	0.2555
		0.5	0.0257	0.1671	0.0512	0.2529	0.0743	0.2864
		1	0.0258	0.1675	0.0525	0.2539	0.0745	0.2865
		3	0.0279	0.1709	0.0556	0.2555	0.0792	0.2908
		5	0.0284	0.1709	0.0586	0.2572	0.0831	0.2932
	0.1	0.1	0.0213	0.1122	0.0258	0.1485	0.0276	0.1578
		0.5	0.0217	0.1126	0.0401	0.1581	0.0353	0.1648
		1	0.0222	0.1132	0.0414	0.1592	0.0546	0.1796
		3	0.0240	0.1155	0.0449	0.1624	0.0648	0.1867
		5	0.0242	0.1155	0.0467	0.1642	0.0672	0.1889
	0.2	0.1	0.0225	0.1123	0.0185	0.1410	0.0197	0.1492
		0.5	0.0233	0.1131	0.0362	0.1556	0.0275	0.1568
		1	0.0237	0.1136	0.0452	0.1581	0.0480	0.1718
		3	0.0259	0.1160	0.0503	0.1617	0.0692	0.1849
		5	0.0271	0.1166	0.0523	0.1634	0.0727	0.1877
100	0.01	0.1	0.0185	0.1752	0.0378	0.2626	0.0546	0.2910
		0.5	0.0187	0.1758	0.0383	0.2626	0.0590	0.2947
		1	0.0188	0.1759	0.0398	0.2628	0.0598	0.2986
		3	0.0192	0.1773	0.0427	0.2665	0.0635	0.2980
		5	0.0206	0.1796	0.0469	0.2695	0.0684	0.3026
	0.1	0.1	0.0183	0.1103	0.0204	0.1456	0.0227	0.1544
		0.5	0.0186	0.1108	0.0360	0.1566	0.0313	0.1638
		1	0.0190	0.1114	0.0378	0.1577	0.0491	0.1773
		3	0.0204	0.1133	0.0419	0.1611	0.0606	0.1853
		5	0.0210	0.1141	0.0421	0.1625	0.0615	0.1874
	0.2	0.1	0.0199	0.1109	0.0148	0.1386	0.0165	0.1468
		0.5	0.0201	0.1113	0.0310	0.1505	0.0244	0.1530
		1	0.0206	0.1117	0.0418	0.1559	0.0436	0.1684
		3	0.0229	0.1140	0.0464	0.1597	0.0654	0.1829
		5	0.0237	0.1150	0.0465	0.1617	0.0658	0.1857

with sample size $n = 50$ and in Table 6 for simulations with $n = 100$. As we can see, the bootstrap variance estimates tend to underestimate the true variance when the censoring intervals are extremely large ($\rho = 0.01$). For moderate interval lengths ($\rho = 0.1$), the bootstrap performs well when $n = 100$, but not so well for $n = 50$. For narrower intervals, the bootstrap gave satisfying results even with $n = 50$. These two tables also illustrate why we prefer $\hat{S}^h_{GTL}$ to $\hat{S}^h_{GTR}$, the bias of the latter being obvious.

4 Real data applications

4.1 Pine weevil data

The white pine weevil is a type of beetle that is regarded as one of the most important pests to conifers. Nathoo (2009) studied its impact on the growth of 4330 spruce trees using a joint spatial model for the recurrence of pine weevil infestation and tree growth. The height of the trees was measured yearly during 11 years. For each tree and each year we know whether pine weevil infestation occurred during the year, but not the exact time at which it started. If we let T_i denote the time at which tree i is first infested by pine weevils and by $Z_i(t_{ij})$ the height of tree i at time t_{ij}, then we are in the presence of a dataset of the form $\Big\{(L_i, R_i, Z_i(t_{i1}), \ldots, Z_i(t_{i,11})),\ i = 1, \ldots, 4330\Big\}$. A first order approximation to a non-spatial version of the growth model considered by Nathoo (2009) yields $Z_i(t_{ij}) = \beta_i t_{ij} + \varepsilon_{ij}$, where β_i can be interpreted as the growth rate of tree i.

The method proposed in this paper with $\hat{\beta}_i$ chosen as the least squares estimate of β_i can give an initial picture of the relationship between the age at first pine weevil infestation and the growth rate of the tree. A plot of $\hat{S}^h_{GTL}(t|\beta)$ as of function of t for various values of β and $h = 1.2$ (equation (7) of Dehghan and Duchesne (2011a)) is shown in Figure 1. As was observed by Nathoo (2009), the probability of infestation before age 5 is virtually null. Then we observe that trees with a higher growth rate are expected to get infested later. This result cannot be directly compared with Nathoo's approach that modeled the intensity of the number of infestations as a function of tree height, and not the time of first infection as a function of growth rate.

Note that, to make such inferences formal, one should fit a parametric or semi-parametric model to the data. The plot of $\log[-\log\{\hat{S}^h_{GTL}(t|\beta)\}]$ vs $\log t$ given in Figure 2 are close to parallel straight lines, suggesting that a proportional hazards or accelerated failure time model with a Weibull baseline distribution could be an appropriate model for the time until first infestation as a function of the growth rate.

Finally if one were interested to estimate the probability that a tree that has yet to

Table 5: Monte Carlo variances of $\hat{S}(t|\beta_0 = 1.5)$ at five values of t corresponding to the 10th, 25th, 50th, 75th and 90th percentiles of the Weibull distribution together with bootstrap estimates of this variance. Results are given for $\sigma_\varepsilon = 1$, $Z(t_{ij}) = \beta_i t_{ij} + \varepsilon_{ij}$, four values of ρ, $n = 50$ and under pure interval-censoring. All values are based on 5,000 replications. S_{True} denotes the true value of $S(t|\beta_0 = 1.5)$, $\bar{S}_{GTL}$ and $\bar{S}_{GTR}$ respectively denote the average values of the $\hat{S}^h_{GTL}$ and $\hat{S}^h_{GTR}$ estimators, and V_{EGT} and V_{BGT} respectively denote the empirical (Monte Carlo) variance and average bootstrap variance of the $\hat{S}^h_{GTL}$ estimator.

ρ	S_{True}	$\bar{S}_{GTL}$	$\bar{S}_{GTR}$	V_{EGT}	V_{BGT}
	0.8927	0.9059	0.9325	0.0342	0.0126
	0.7744	0.7737	0.8751	0.0691	0.0264
0.01	0.5115	0.5097	0.6493	0.0916	0.0638
	0.2334	0.2772	0.3683	0.0611	0.06070
	0.1033	0.1670	0.2134	0.0494	0.0502
	0.8927	0.9175	0.9279	0.0080	0.0045
	0.7744	0.7594	0.7800	0.0158	0.0125
0.1	0.5115	0.4977	0.5298	0.0206	0.0190
	0.2334	0.2903	0.2910	0.0162	0.0151
	0.1033	0.1590	0.1599	0.0095	0.0091
	0.8927	0.9087	0.9218	0.0063	0.0054
	0.7744	0.7729	0.7861	0.0124	0.0109
0.2	0.5115	0.5249	0.5329	0.0171	0.0156
	0.2334	0.2786	0.2882	0.0124	0.0120
	0.1033	0.1484	0.1573	0.0071	0.0070
	0.8927	0.9025	0.9073	0.0032	0.0031
	0.7744	0.7703	0.7751	0.0065	0.0063
4	0.5115	0.5250	0.5268	0.0090	0.0088
	0.2334	0.2826	0.2828	0.0071	0.0070
	0.1033	0.1515	0.1516	0.0041	0.0041

Table 6: Monte Carlo variances of $\hat{S}(t|\beta_0 = 1.5)$ at five values of t corresponding to the 10th, 25th, 50th, 75th and 90th percentiles of the Weibull distribution together with bootstrap estimates of this variance. Results are given for $\sigma_\varepsilon = 1$, $Z(t_{ij}) = \beta_i t_{ij} + \varepsilon_{ij}$, four values of ρ, $n = 100$ and under pure interval-censoring. All values are based on 5,000 replications. S_{True} denotes the true value of $S(t|\beta_0 = 1.5)$, $\bar{S}_{GTL}$ and $\bar{S}_{GTR}$ respectively denote the average values of the $\hat{S}^h_{GTL}$ and $\hat{S}^h_{GTR}$ estimators, and V_{EGT} and V_{BGT} respectively denote the empirical (Monte Carlo) variance and average bootstrap variance of the $\hat{S}^h_{GTL}$ estimator.

ρ	S_{True}	$\bar{S}_{GTL}$	$\bar{S}_{GTR}$	V_{EGT}	V_{BGT}
	0.8927	0.9059	0.9325	0.0342	0.0126
	0.7744	0.7737	0.8751	0.0691	0.0264
0.01	0.5115	0.5097	0.6493	0.0916	0.0638
	0.2334	0.2772	0.3683	0.0611	0.06070
	0.1033	0.1670	0.2134	0.0494	0.0502
	0.8927	0.9175	0.9279	0.0080	0.0045
	0.7744	0.7594	0.7800	0.0158	0.0125
0.1	0.5115	0.4977	0.5298	0.0206	0.0190
	0.2334	0.2903	0.2910	0.0162	0.0151
	0.1033	0.1590	0.1599	0.0095	0.0091
	0.8927	0.9087	0.9218	0.0063	0.0054
	0.7744	0.7729	0.7861	0.0124	0.0109
0.2	0.5115	0.5249	0.5329	0.0171	0.0156
	0.2334	0.2786	0.2882	0.0124	0.0120
	0.1033	0.1484	0.1573	0.0071	0.0070
	0.8927	0.9025	0.9073	0.0032	0.0031
	0.7744	0.7703	0.7751	0.0065	0.0063
4	0.5115	0.5250	0.5268	0.0090	0.0088
	0.2334	0.2826	0.2828	0.0071	0.0070
	0.1033	0.1515	0.1516	0.0041	0.0041

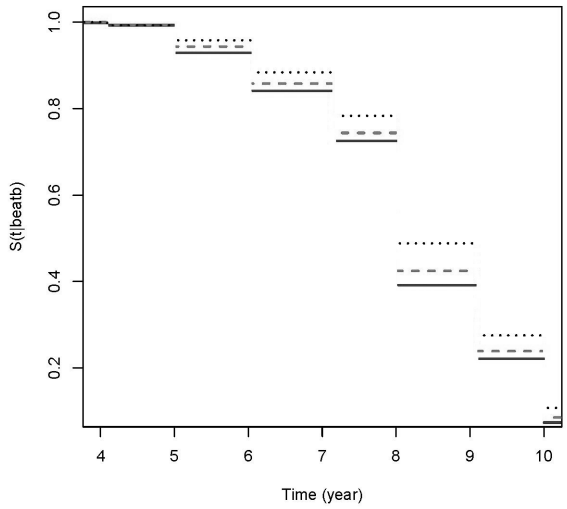

Figure 1: Plot of $\hat{S}^{h=1.2}_{GTL}(t|\beta)$ as a function of time (in years) for the pine weevil data for the quartile values of β: the solid line corresponds to $\beta = 1.56$, the dashed line to $\beta = 1.81$ and the dotted line to $\beta = 2.11$.

be infested at age 7 will remain free of pine weevil by time t as a function of its growth rate β, one only needs to plot $\hat{S}(t|\beta)/\hat{S}(7|\beta)$, as suggested by equation (2.3). Figure 3 shows such a plot for three values of β.

4.2 Reliability of electrical equipment

For confidentiality reasons, we cannot give the complete details of the dataset that we are about to discuss. But the data observed are of the form $\Big\{(L_i, R_i, Z_i(t_{i1}), \ldots, Z_i(t_{iN_i})), i = 1, \ldots, 738\Big\}$, where T_i is the time until failure of a piece of electrical equipment and $Z_i(t_{ij})$ is the concentration of a gas that is dissolved in the oil of this piece of equipment at time t_{ij}. The number N_i of oil analyses varies greatly from unit to unit (min. 20, 1st quartile 22, median 25.5, 3rd quartile 33, max. 148). The question of interest is whether a proportional hazards or an accelerated failure time model would be appropriate for the relationship between the time to failure and some features of the gas concentration history. Following (Meeker and Escobar , 1998a, Chap. 13), we will assume that the

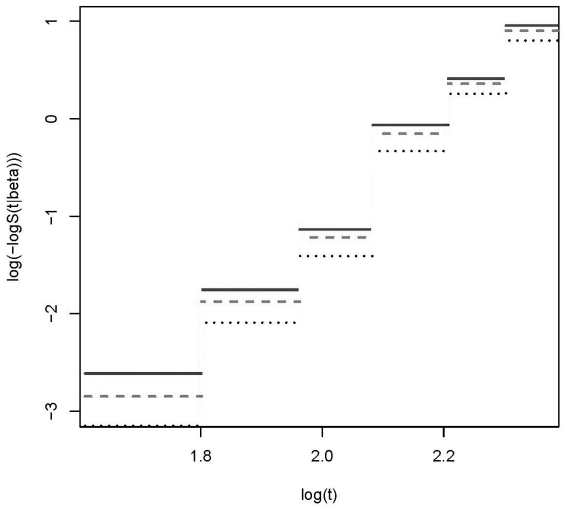

Figure 2: Plot of $\log[-\log\{1-\hat{S}^{h=1.2}_{GTL}(t|\beta)\}]$ as a function of log(time) (in years) for the pine weevil data for the quartile values of β: the solid line corresponds to $\beta = 1.56$, the dashed line to $\beta = 1.81$ and the dotted line to $\beta = 2.11$.

time to failure given the gas concentration history only depends on some features of the path. Data exploration suggests a model of the form $Z_i(t_{ij}) = \beta_{i0} + \beta_{i1}t_{ij} + \varepsilon_{ij}$. The correlation between the least squares estimates $\hat{\beta}_{i0}$ and $\hat{\beta}_{i1}$ being very strong (−93%), it is reasonable to assume that $S(t|\beta_{i0}, \beta_{i1}) = S(t|\beta_{i1})$. The value of $\hat{S}^h_{GTL}(t|\beta)$ with $h = 0.6$ (equation (7) of Dehghan and Duchesne (2011a)) is given for various values of β in Figure 4. As we can see, when the concentration of the dissolved gas increases more rapidly (or decreases more slowly), the pieces of equipment tend to fail more rapidly.

The plot of $\log[-\log\{\hat{S}^h_{GTL}(t|\beta)\}]$ vs $\log t$ in Figure 5 does not seem to show curves that are horizontal or vertical shifts of each other, suggesting that simple proportional or accelerated failure time models are unlikely to fit these data well. Perhaps a proportional hazards model with an interaction between β_i and an indicator that $\log(t) > 4.42$ could work well.

Again, if one is interested in prognostic, say on the probability of survival of a piece of equipment still working at time 76, then plots of $\hat{S}(t|\beta)/\hat{S}(76|\beta)$, as the one shown in Figure 6 for various values of β, could be interesting.

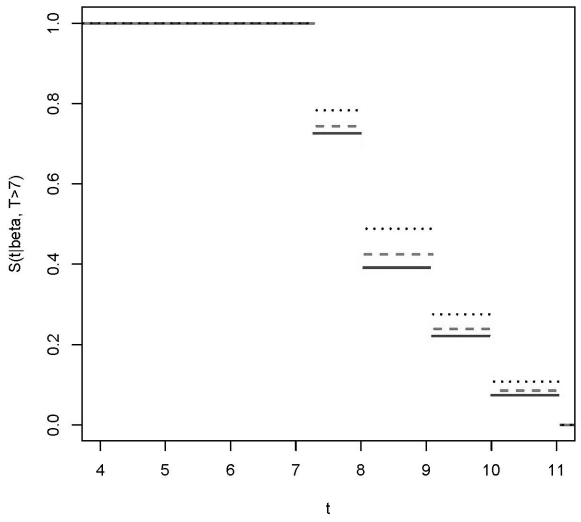

Figure 3: Plot of $\hat{S}^{h=1.2}_{GTL}(t|t > 7, \beta)$ as a function of time (in years) for the pine weevil data for the quartile values of β: the solid line corresponds to $\beta = 1.56$, the dashed line to $\beta = 1.81$ and the dotted line to $\beta = 2.11$.

5 Discustions

In this paper, we have proposed a nonparametric estimator of the conditional survival function when the time-to-event variable is subject to interval-censoring and the covariate is continuous and time-varying. Our approach is inspired from the literature on joint modeling of survival and longitudinal data and on reliability modeling from degradation data. It consists in applying the nonparametric estimator of the conditional survival function proposed by Dehghan and Duchesne (2011a) to estimates of the features of the covariate path, for example as obtained with a linear model. Our simulation study showed that the method works well in finite samples, and this even when the linear model for the covariate path is misspecified. We also obtained the asymptotic bias and variance of the estimator under the assumption that the time-to-event is uncensored. We demonstrated how the method can be used in the early stages of data exploration by applying it to two real datasets.

Because our intent is to propose a simple graphical tool that requires as few mod-

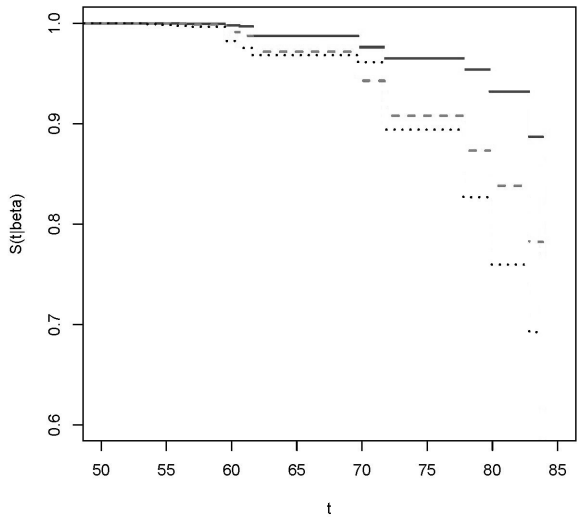

Figure 4: Plot of $\hat{S}^{h=0.6}_{GTL}(t|\beta)$ as a function of time for the electrical equipment data for the quartile values of β: the solid line corresponds to $\beta = -1.3$, the dashed line to $\beta = -0.3$ and the dotted line to $\beta = 0.05$.

eling assumptions as possible, we do not advocate its use for formal inferences, but rather as a tool to help in the model specification process. Though semi- or completely parametric models would be preferable for formal inferences, perhaps nonparametric inferences could be possible, but at the cost of greatly complexifying the method. For starters, confidence bands would have to be derived. But without a rigorous large sample theory for the GT estimator, which is still an open problem, this is not possible. Second, nonparametric regression methods with errors in covariate have been shown to be inconsistent when the variance of the error is fixed and a convolution kernel is not used. We have shown here that, when this variance does shrink to zero, then it is possible to obtain consistent estimates with the ordinary NW estimator. But perhaps gains in efficiency could be achievable with a method that does take the error in the covariate into consideration. Another possible way to gain efficiency would be to postulate a complete mixed model on the β_i and then use inference tools for mixed models to get more efficient estimators $\hat{\beta}_i$. For instance, one could assume that the β_i are iid from a zero mean multivariate normal distribution and then use the best linear unbiased

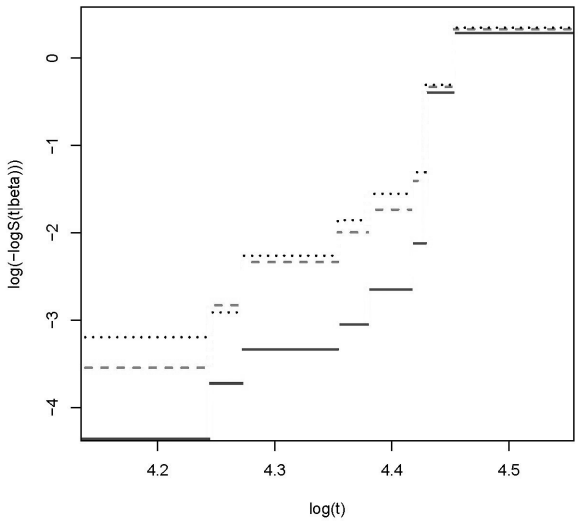

Figure 5: Plot of $\log[-\log\{1-\hat{S}^{h=0.6}_{GTL}(t|\beta)\}]$ as a function of log(time) for the electrical equipment data for the quartile values of β: the solid line corresponds to $\beta = -1.3$, the dashed line to $\beta = -0.3$ and the dotted line to $\beta = 0.05$.

predictor of β_i to compute $\hat{S}(t|\beta)$ rather than the ordinary least squares estimator of β_i as was done here.

Acknowledgements

The authors are grateful to the Natural Sciences Canada and research center of University of Sistan and Baluchestan for financial support of this work. We also wish to thank Dr Farook Nathoo for sharing the pine weevil data with us and JFB from a Canadian company for giving us access to the data on the reliability of electrical equipment.

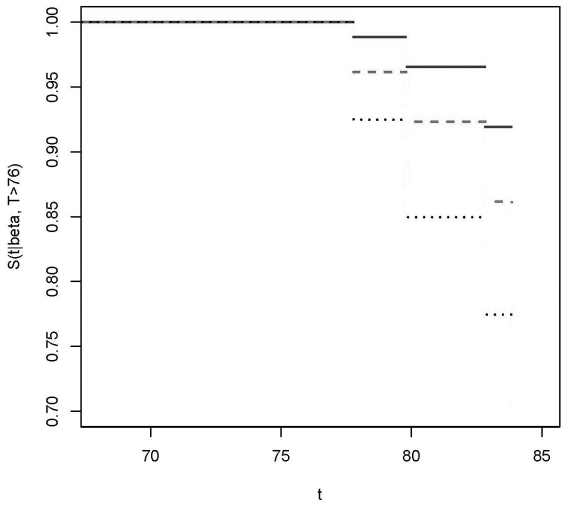

Figure 6: Plot of $\hat{S}^{h=1.2}_{GTL}(t|t > 76, \beta)$ as a function of time for the electrical equipment data for the quartile values of β: the solid line corresponds to $\beta = -1.3$, the dashed line to $\beta = -0.3$ and the dotted line to $\beta = 0.05$.

References

Beran R (1981). *Nonparametic regression with randomly censored survival data*. Technical report, Department of Statistics, U. of California, Berkeley.

Carroll RJ, Maca JD, Ruppert D (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, 541-554.

Dabrowska, D. M. (1987). Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*, 181-197.

Dehghan, M. H., and Duchesne, T. (2011a). A generalization of Turnbulls estimator for nonparametric estimation of the conditional survival function with interval-censored data. *Lifetime data analysis*, **17**(2), 234-255.

Dehghan, M. H., and Duchesne, T. (2011). On the performance of some non-parametric

estimators of the conditional survival function with interval-censored data. *Computational Statistics and Data Analysis*, **55**(12), 3355-3364.

Dehghan MH, Duchesne T, Baillargeon S (2015). gte: Generalized Turnbull's Estimator. R Package Version 1.2-2.

Fan, J., and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics*, 1900-1925.

Hansen B (2009). Nadyara-Watson Local Linnear Regression. Lecture Notes, http://wwww.ssc.wisc.edu/~bhansen/718/NonParametrics2.pdf, Accessed on May 11, 2012.

Hogan, J. W., and Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in medicine*, **16**(3), 239-257.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley and Sons.

Li, L., Watkins, T., and Yu, Q. (1997). An EM Algorithm for Smoothing the Selfconsistent Estimator of Survival Functions with Interval-censored Data. *Scandinavian Journal of Statistics*, **24**(4), 531-542.

Lu, C. J., Meeker, W. Q., and Escobar, L. A. (1996). A comparison of degradation and failure-time analysis methods for estimating a time-to-failure distribution. *Statistica Sinica*, 531-546.

Meeker WQ, Escobar LA (1998a). *Statistical Methods for Reliability Data*. Wiley, New York, NY.

Meeker, W. Q., Escobar, L. A., and Lu, C. J. (1998). Accelerated degradation tests: modeling and analysis. *Technometrics*, **40**(2), 89-99.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, **9**(1), 141-142.

Nathoo, F. S. (2010). Joint spatial modeling of recurrent infection and growth with processes under intermittent observation. *Biometrics*, **66**(2), 336-346.

Ramsey JO, Silverman BW (1997). *Functional Data Analysis*. Springer, New York, NY.

Schennach, S. M. (2004). Nonparametric regression in the presence of measurement error. *Econometric Theory*, **20**(06), 1046-1093.

Sun J (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York, NY.

Tseng, Y. K., Hsieh, F., and Wang, J. L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, **92**(3), 587-603.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, **90**(429), 27-37.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290-295.

Watson, G. S. (1964). Smooth regression analysis. *Sankhy: The Indian Journal of Statistics, Series A*, 359-372.
Watson GS (1964). Smooth regression analysis. Sankhya Ser A 26:359-372.

Appendix

Example where $\sigma_N^2 \to 0$ and $N\sigma_N \to \infty$ when $N \to \infty$

Because we consider the estimator of β_i, $\hat{\beta}_i = \frac{\sum_{j=1}^{N_i} Z_{ij}t_{ij}}{\sum_{j=1}^{N_i} t_{ij}^2}, i = 1, \ldots, n$ and $\sigma_{N_i}^2 = \frac{\sigma_\varepsilon^2}{\sum_{j=1}^{N_i} t_{ij}^2}, i = 1, \ldots, n$ where σ_ε^2 is a value. We assume that $t_{ij} \in (0, [\tau < \infty]), j = 1, \ldots, N_i, i = 1, \ldots, n$. When $N_i \to \infty$, $\frac{1}{N_i}\sum_{j=1}^{N_i} t_{ij}^2 \to \overline{t_i^2}\ \forall i$, in other word when $N \to \infty$, $\frac{1}{N}\sum_{j=1}^{N} t_j^2 \to \overline{t^2}$ (fixed value). Therefore it is easy to show ($\sigma_N^2 = \frac{\sigma_\varepsilon^2}{N\overline{t^2}}$) $\to 0$ and ($N\sigma_N = \frac{\sqrt{N}\sigma_\varepsilon}{\sqrt{\overline{t^2}}}$) $\to \infty$ when $N \to \infty$.

Proof of Theorem 1

Let $K_h(u) = K(u/h)/h$, $\gamma_i = 1\{T_i > t\}$ and $\xi_{\hat{\beta}}(\beta_0) = 1/\{n^{-1}\sum_{i=1}^{n} K_h(\hat{\beta}_i - \beta_0)\}$. Then $\xi_{\hat{\beta}}(\beta_0)$ is one over the classical kernel estimator of the marginal density $f_{\hat{\beta}}(\beta_0)$ and thus

$$\xi_{\hat{\beta}}(\beta_0) = \frac{1}{f_{\hat{\beta}}(\beta_0)} + o_p(1).$$

Let $f_{\hat{\beta}_i|\beta_i}$ denote the conditional density of $\hat{\beta}_i$ given the value of β_i and let f_{β_i} denote the marginal density of β_i, so that the joint density of $\hat{\beta}_i$ and β_i is given by $f_{\hat{\beta}_i,\beta_i} = f_{\hat{\beta}_i|\beta_i} f_{\beta_i}$. Now for the expectation of $\hat{S}$ we have

$$
\begin{aligned}
E(\hat{S}^{h_n}(t|\beta_0)) &= E_{\hat{\beta},\beta}\left(E\left[\frac{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)\gamma_i}{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)}\Big|\hat{\beta}_i,\beta_i\right]\right)\\
&= E_{\hat{\beta},\beta}\left(\frac{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)S(t|\beta_i)}{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)}\right) \qquad (5.1)\\
&= E_{\hat{\beta},\beta}\Big[S(t|\beta_0)+\frac{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)(\beta_i-\beta_0)}{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)}S^{01}(t|\beta_0)\\
&+\frac{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)(\beta_i-\beta_0)^2}{2\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)}S^{02}(t|\beta_0)\\
&+\frac{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)O([\beta_i-\beta_0]^3)}{\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)}\Big]\\
&= E_{\hat{\beta},\beta}\Big(S(t|\beta_0)+S^{01}(t|\beta_0)\xi_{\hat{\beta}}(\beta_0)n^{-1}\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)(\beta_i-\beta_0)+\\
&\frac{S^{02}(t|\beta_0)}{2}\xi_{\hat{\beta}}(\beta_0)n^{-1}\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)(\beta_i-\beta_0)^2\\
&\quad+\xi_{\hat{\beta}}(\beta_0)n^{-1}\sum_{i=1}^n K_{h_n}(\hat{\beta}_i-\beta_0)O([\beta_i-\beta_0]^3)\Big) \qquad (5.2)\\
&= S(t|\beta_0)+S^{01}(t|\beta_0)f_{\hat{\beta}}(\beta_0)^{-1}E_{\hat{\beta},\beta}\Big(K_{h_n}(\hat{\beta}-\beta_0)(\beta-\beta_0)\Big)\\
&\quad+\frac{S^{02}(t|\beta_0)f_{\hat{\beta}}(\beta_0)^{-1}}{2}E_{\hat{\beta},\beta}\Big(K_{h_n}(\hat{\beta}-\beta_0)(\beta-\beta_0)^2\Big)\\
&\quad+f_{\hat{\beta}}(\beta_0)^{-1}E_{\hat{\beta},\beta}\Big(K_{h_n}(\hat{\beta}-\beta_0)O([\beta-\beta_0]^3)\Big) \qquad (5.3)\\
&= S(t|\beta_0)+S^{01}(t|\beta_0)f_{\hat{\beta}}(\beta_0)^{-1}\int\int K_{h_n}(\hat{\beta}-\beta_0)(\beta-\beta_0)f_{\hat{\beta}|\beta}(\hat{\beta})f_\beta(\beta)d\hat{\beta}d\beta\\
&\quad+\frac{S^{02}(t|\beta_0)f_{\hat{\beta}}(\beta_0)^{-1}}{2}\int\int K_{h_n}(\hat{\beta}-\beta_0)(\beta-\beta_0)^2 f_{\hat{\beta}|\beta}(\hat{\beta})f_\beta(\beta)d\hat{\beta}d\beta\\
&\quad+f_{\hat{\beta}}(\beta_0)^{-1}\int\int K_{h_n}(\hat{\beta}-\beta_0)\{O([\beta-\beta_0]^3)\}f_{\hat{\beta}|\beta}(\hat{\beta})f_\beta(\beta),d\hat{\beta}d\beta. \qquad (5.4)
\end{aligned}
$$

where the $O(\cdot)$ term in the last integral denotes a polynomial in $(\beta-\beta_0)$ with terms of degree 3 or more.

Under the theorem assumptions and keeping in mind that $K(\cdot)$ is the standard

Gaussian density, we have that $f(\hat{\beta},\beta) = K_{\sigma_N}(\hat{\beta}-\beta)f_\beta(\beta)$. Now let $\mu_{2m}^M = \int u^{2m}K^M(u)du$, where m and M are positive integers and define

$$\Gamma^m = \int\int (x-\beta_0)^m K_{h_n}(y-\beta_0)K_{\sigma_N}(y-x)f_\beta(x)dydx.$$

From the symmetry of $K(\cdot)$ and with the change of variable $(y-\beta_0)/h_n = u, (x-y)/\sigma_N = v$, we can rewrite Γ^m as

$$\Gamma^m = \int\int (h_n u + \sigma_N v)^m K(u)K(v)f_\beta(x)dudv.$$

Taking Taylor expansion about $h_n = 0$ and $\sigma_N = 0$, we get

$$\begin{aligned}\Gamma^m = &\int\int (h_n u+\sigma_N v)^m K(u)K(v)dudv f_\beta(\beta_0)\\ &+\int\int (h_n u+\sigma_N v)^{m+1}K(u)K(v)dudv f_{\beta_i}^{(1)}(\beta_0)\\ &+\int\int \frac{(h_n u+\sigma_N v)^{m+2}}{2}K(u)K(v)dudv f_\beta^{(2)}(\beta_0) + O(h_n^l\sigma_N^{l'}\big|l+l'\geq 4).\end{aligned}$$

Substituting Γ^1 and Γ^2 into (5.4) we get

$$\begin{aligned}E[\hat{S}(t|\beta_0)] &= S(t|\beta_0) + \frac{\mu_2(h_n^2+\sigma_N^2)}{2f_{\hat{\beta}}(\beta_0)}\Big[2S^{01}(t|\beta_0)f_\beta^{(1)}(\beta_0) + S^{02}(t|\beta_0)f_\beta(\beta_0)\Big]\\ &\quad + \frac{3\mu_2^2h_n^2\sigma_N^2S^{02}(t|\beta_0)f_\beta^{(2)}(\beta_0)}{f_{\hat{\beta}}(\beta_0)} + \frac{\mu_4(h_n^4+\sigma_N^4)S^{02}(t|\beta_0)f_\beta^{(2)}(\beta_0)}{2f_{\hat{\beta}}(\beta_0)}\\ &\quad + O(h_n^\ell\sigma_N^{\ell'}\big|\ell+\ell'\geq 6)\\ &= S(t|\beta_0) + \frac{\mu_2(h_n^2+\sigma_N^2)}{2f_{\hat{\beta}}(\beta_0)}\Big[2S^{01}(t|\beta_0)f_\beta^{(1)}(\beta_0) + S^{02}(t|\beta_0)f_\beta(\beta_0)\Big]\\ &\quad + O(h_n^\ell\sigma_N^{\ell'}\big|\ell+\ell'\geq 4),\end{aligned}$$

which leads to the bias expression in the theorem. To calculate the variance of the estimator, we follow Hansen (2009) and let

$$\hat{S}(t|\beta_0) = S(t|\beta_0) + \frac{\hat{m}_1}{\hat{f}_{\hat{\beta}_i}(\beta_0)} + \frac{\hat{m}_2}{\hat{f}_{\hat{\beta}_i}(\beta_0)}, i = 1,\ldots,n,$$

where

$$\hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\hat{\beta}_i - \beta_0)\{S(t|\beta_i) - S(t|\beta_0)\} \text{ and}$$

$$\hat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\hat{\beta}_i - \beta_0)\epsilon_i \quad \text{with} \quad \epsilon_i = \gamma_i - S(t|\beta_i).$$

Under the assumptions of the Theorem 1, $E(\epsilon|\hat{\beta},\beta) = E\left(\gamma - S(t|\beta)\,\middle|\hat{\beta},\beta\right) = 0$ and thus $E(\hat{m}_2|\hat{\beta},\beta) = 0$ and $\text{Var}(\hat{m}_2|\hat{\beta},\beta) = E(\hat{m}_2^2|\hat{\beta},\beta)$.

In the supplementary material, we drive expression for $\text{Var}[\xi_{\hat{\beta}}(\beta_0)\hat{m}_2]$, $\text{Var}[\xi_{\hat{\beta}}(\beta_0)\hat{m}_1]$ and $Cov(\hat{m}_1, \hat{m}_2)$ which we combine to obtaine the desired variance formula.