Statistica Sinica Preprint No: SS-2021-0118	
Title	Regression Analysis of Spatially Correlated Event Durations With Missing Origins Annotated by Longitudinal Measures
Manuscript ID	SS-2021-0118
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0118
Complete List of Authors	Yi Xiong, W. John Braun, Thierry Duchesne and X. Joan Hu
Corresponding Author	X. Joan Hu
E-mail	joanh@stat.sfu.ca

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Statistica Sinica

REGRESSION ANALYSIS OF SPATIALLY CORRELATED EVENT DURATIONS WITH MISSING ORIGINS ANNOTATED BY LONGITUDINAL MEASURES

Yi Xiong, W. John Braun, Thierry Duchesne, and X. Joan Hu

Fred Hutchinson Cancer Research Center, Simon Fraser University, University of British Columbia-Okanagan, and Université Laval

Abstract: In this study, we examine event durations when study units may be spatially correlated and the time origins of the events are missing. We develop regression models based on the partly observed durations with the aid of available longitudinal information. We use the first-hitting-time model to link the data of event durations and the associated longitudinal measures with shared random effects. We present procedures for estimating the model parameters and an induced estimator of the conditional distribution of the event duration. We apply the EM algorithm and Monte Carlo methods to compute the proposed estimators. We establish the consistency and asymptotic normality of the estimators, and present their variance estimation. We demonstrate the proposed approach using a collection of wildfire records from Alberta, Canada. We also examine its performance numerically, and compare it with that of two competitors using a simulation.

Key words and phrases: Asymptomatic event, first hitting time, EM algorithm and Monte Carlo method, joint modelling, mixed effects model

1. Introduction

Many research studies are primarily interested in relating an event duration with possible covariates for purposes of inference and/or prediction. We are particularly interested in regressions based on event-duration observations with missing time

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

origins. Such observations arise in infectious disease research, wildfire management, and other areas. For example, many studies are concerned with the incubation period of coronavirus disease 2019 (COVID-19); see, for example, Qin et al. (2020). The incubation period is defined as the duration between the infection time and the onset time of symptoms. However, the exact infection time is usually unknown, and so the time based on the individual's recollection is used as a proxy of the true infection time. As another example, the interval from the start time of a wildfire to when suppression activities (the so-called time of initial attack) begin is sometimes used as a gauge of fire management effectiveness, and provides important information for predicting the development of the fire. Once a fire is detected and reported, resources such as a fire crew or an airtanker are allocated to suppress the blaze (e.g., Martell, 2007; Morin, 2014). Because fires are not always detectable until they reach a certain size, the start time of a wildfire is always unknown; instead, the time of report is often used as a surrogate when assessing the time to the initial attack. Although various regression methods have been developed to evaluate the association of a single variable with its covariates, most approaches are not directly applicable to such observations.

Xiong, Braun, and Hu (2021) propose an approach for tackling the problem of missing time origins using longitudinal measures of an associated quantity. The well-known empirical distribution function of the event duration based on independent and identically distributed (i.i.d.) observations is adapted as an estimator of the marginal distribution, using partly observed durations caused by missing time origins. In the context of the wildfire management example, the duration of interest, which we refer to as the ISA duration, is the interval between the time when a wildfire starts and the initial attack time. Xiong, Braun, and Hu (2021) apply the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

procedure to estimate the distribution of the ISA duration in each subregion. The estimated distributions appeared to vary regionally, and the ISA durations depend on weather variables, such as relative humidity, temperature, and wind speed. We expect fires in close proximity to burn similar types of vegetation, in terms of both the understory and the canopy. Topography also plays a role, because fires tend to burn faster up a slope. Thus, the ISA durations are potentially spatially correlated. These considerations have partially motivated the research presented in this paper. We are primarily concerned with regression analyses in situations in which the response variable is an event duration and the available observations have missing time origins.

When longitudinal data are available with time-to-event data, a joint model is often considered. A typical model setting is a linear or nonlinear mixed-effects model for the longitudinal measures, and a semiparametric or parametric regression model for the time-to-event data, with the two models sharing some random effects or variables. Estimation methods for such joint models have received much attention (e.g., Tseng, Hsieh, and Wang, 2005; Wu, Liu, and Hu, 2010; Wulfsohn and Tsiatis, 1997); detailed reviews of recent works can be found in Furgal, Sen, and Taylor (2019) and Papageorgiou et al. (2019). However, no existing methods can be applied when the time origin is missing for both the event process and the longitudinal process.

We assume that the ISA duration follows a semiparametric accelerated failure time (AFT) regression model. This accounts for the potential spatial correlation of the duration times, and adjusts for the missing time origins using available longitudinal data. We specify the regression function in the AFT model as a linear combination of some or all of the predictors, together with nonparametric terms

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

for the predictors that are nonlinearly related to the response. We follow Xiong, Braun, and Hu (2021) to overcome the challenge posed by the missing time origin by employing associated longitudinal measures and specifying the longitudinal process as a Brownian motion with random drift. Longitudinal measures provide us with information about the missing time origin and assist with the inference on the unknown duration. This approach also allows us to link the longitudinal process and the regression model through shared random effects to accommodate the spatial correlations between individuals. Our approach may be viewed as an extension of the Buckley–James estimator (Buckley and James, 1979), which adapts the classical least squares estimation (LSE) to right-censored observations on the response. A similar effort is available in the interesting work of Ning, Qin, and Shen (2011).

The methodology of *threshold regression (TR)* provides an alternative to the joint modeling of longitudinal and time-to-event data. In a TR, the event durations are interpreted as the *first hitting times* (FHTs) of a boundary or threshold state crossed by sample paths of a longitudinal process (Lee and Whitmore, 2006). For example, in HIV studies, CD4 counts are commonly used as markers for the health status of HIV-infected individuals, and the time at which AIDS develops can be viewed as the time when the CD4 process first reaches 200 (Doksum and Normand, 1995). As another example, Xiong, Braun, and Hu (2021) estimate the distribution of the ISA duration by defining the FHT as the reporting delay, which is the time taken for the burnt area to reach the reported size.

Spatial correlation adds another layer of complexity to the analysis. Frailty models are commonly employed with spatially correlated time-to-event data (e.g., Banerjee, Wall, and Carlin, 2003; Li and Ryan, 2002; Motarjem, Mohammadzadeh, and Abyar, 2017). As in frailty models, we embed the spatial correlation in the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

random effects.

To facilitate the development of the proposed estimation procedure, we begin by assuming independence between individuals. We estimate the parameters of the longitudinal process based on the likelihood, and estimate the regression function in the AFT model using an adaptation of the LSE integrated with kernel smoothing. We also propose an estimator of the conditional distribution of the duration, given covariates, in the presence of a missing time origin. We then extend the procedure to spatially correlated study units. The approach is illustrated using a data set of wildfires in Alberta, Canada.

The rest of the paper is organized as follows. In Section 2, we describe the regression model of the duration and the stochastic model for longitudinal measures of the area burned over time. In Section 3, we present the estimation procedure for i.i.d. observations and its extension to spatial correlation, and derive their asymptotic properties. In Section 4, we apply our approach by analyzing wildfire records, and Section 5 presents simulation studies that examine the finite-sample performance of the proposed approach. Section 6 concludes the paper.

2. Notation and Modeling

2.1 Notation

We follow the notation introduced in Xiong, Braun, and Hu (2021). Consider a wildfire with its start time, report time, and initial attack time denoted by T_S, T_R, and T_F, respectively. The aforementioned ISA duration is then $L = T_F - T_S$. Let $A(a)$ be the burnt area of a fire at the elapsed time a since the start: $A(0) = 0$ and $A(S) = B$ are the respective burnt areas at the start time and report time. Here, $S = T_R - T_S$ represents the fire's elapsed time till the report time. Note that

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

$L = S + L^*$, with $L^* = T_F - T_R$. While L^* may be observed, the reporting delay S is usually unavailable, and thus so is L. Further, let $A(L) - A(S) = D$ be the increased burnt area over the interval (T_R, T_F).

We denote the location where a fire is detected by $\boldsymbol{\omega} = (\omega_1, \omega_2)'$, and a vector of environmental and spatial factors (e.g., wind speed, fuel type, and region) associated with the fire at the report time by $\mathbf{X}$. Figure 1 presents a progression description of two hypothetical wildfires using the above notation. The solid curve in each plot represents the burnt area over time of a fire that is subject to suppression after detection.

—— *Figure 1 is here.* ——

Suppose that a collection of records on $i = 1, \ldots, n$ wildfires is available, denoted by

$$\mathcal{O} = \bigcup_{i=1}^{n} \mathcal{O}_i = \big\{(T_{Ri}, T_{Fi}, B_i, D_i, \mathbf{X}_i, \boldsymbol{\omega}_i) : i = 1, 2, \cdots, n\big\}. \tag{2.1}$$

Our primary statistical interest is in estimating the conditional distribution of the ISA duration L given the covariate vector $\mathbf{X}$, $P(L \leq t|\mathbf{X} = \mathbf{x})$, using the available data. The estimated $P(L \leq t|\mathbf{X} = \mathbf{x})$ may reveal how the ISA duration $L = S + L^*$ is associated with the covariate vector $\mathbf{X}$, and can be applied for prediction.

Note that the available data $\mathcal{O} = \bigcup_{i=1}^{n} \mathcal{O}_i$ in (2.1) include observations on $L^* = T_F - T_R$, rather than on L. The burnt area records may provide information about the reporting delay S, and thus the ISA duration L. In the following subsections, we present our model of the conditional distribution $P(L \leq t|\mathbf{X} = \mathbf{x})$, along with component models that link the burnt areas to the reporting delay.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

2.2 Longitudinal Model for Repeated Measures

We first use a time-indexed stochastic process $A_i(\cdot)$ to model the longitudinal measures associated with each study unit i. In the wildfire application, these are the burnt areas of fire i over time. We assume that, conditional on the covariates $\mathbf{X}_i$ and random effects $(\delta_{1i}, \delta_{2i})$, the area burned at time u follows the mixed-effect model

$$A_i(u) = \nu_i u + \sigma W_i(u), \ \ i = 1, 2, \cdots, n, \tag{2.2}$$

where $\nu_i = \nu \exp\{\delta_{1i} + \mathbf{X}_i'\boldsymbol{\gamma} + X_i^*\delta_{2i}\}$ with constants $\nu > 0$ and $\sigma > 0$, and $W_i(\cdot)$ is the standard Wiener process. Note that we model only the longitudinal burnt areas before the initial attack in the wildfire application. As presented in Figure 1, the dashed curve in each plot shows the expected trajectory of the fire's burnt area if it had continued to burn without any suppression or intervention. Prior to the initial attack, the dashed curve coincides with the solid curve, and the growth of the fire's burnt areas can be approximated, as assumed in (2.2).

Here, δ_{1i} with ν yields the random intercept term $\log \nu + \delta_{1i}$ of $\log \nu_i$, X_i^* is one of the components of the covariate vector $\mathbf{X}_i$, and δ_{2i} is its (random) coefficient. This model allows the drifts ν_i to accommodate individual-specific covariate effects. We assume that δ_{1i} and δ_{2i} are independent of each other and independent of $W_i(\cdot)$, for $i = 1, \cdots, n$. Furthermore, $\boldsymbol{\delta}_1 = (\delta_{11}, \cdots, \delta_{1n})'$ and $\boldsymbol{\delta}_2 = (\delta_{21}, \cdots, \delta_{2n})'$ are assumed to follow the distributions $MVN(0, \Sigma_1)$ and $MVN(0, \Sigma_2)$, respectively. We consider the following two specifications for the covariance matrices Σ_1 and Σ_2:

- *Independent Study Units.* Let $\Sigma_1 = \psi_1^2 I_{n\times n}$ and $\Sigma_2 = \psi_2^2 I_{n\times n}$, where $I_{n\times n}$ is the identity matrix of size n. Then, $\delta_{1i} \sim N(0, \psi_1^2)$ and $\delta_{2i} \sim N(0, \psi_2^2)$, for $i = 1, 2, \cdots, n$. Denote the density functions by $\phi_1(\cdot; \psi_1)$ and $\phi_2(\cdot; \psi_2)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

respectively.

- *Spatially Correlated Study Units.* Define the (i, j) element of Σ_1 and Σ_2 as

$$(\Sigma_k)_{ij} = \psi_k^2 \exp\Big(- ||\boldsymbol{\omega_i} - \boldsymbol{\omega_j}||/\rho_k \Big) \text{ for } k = 1, 2, \tag{2.3}$$

where $||\cdot||$ is the Euclidean norm. In the rest of this paper, the covariance matrices Σ_k are denoted by $\Sigma(\psi_k, \rho_k)$, for $k = 1, 2$. The spatial correlation is assumed to decay as the geographic distance increases, and $\rho_k = 0$ corresponds to the situation where the δ_{ki} are independent with variance ψ_k.

2.3 Regression Models for Event Duration

We consider the following regression model of the transformed ISA durations $Y_i = \log L_i$:

$$Y_i = h(\mathbf{X}_i, \delta_{1i}, \delta_{2i}) + \epsilon_i, \tag{2.4}$$

where the random errors ϵ_i are independent of the covariates $\mathbf{X}_i$ and the random effects $(\delta_{1i}, \delta_{2i})$, and follow an unspecified distribution function $F_\epsilon(\cdot)$, with $\mathrm{E}[\epsilon_i] = 0$. The random intercept δ_{1i} and the random slope δ_{2i} are shared terms in the longitudinal model (2.2). When study units are spatially correlated, the vector $\boldsymbol{\delta}_k = (\delta_{k1}, \cdots, \delta_{kn})'$, for $k = 1, 2$, follow a multivariate normal distribution, with the covariance matrix specified in (2.3). When study units are assumed to be independent, $\delta_{1i} \overset{i.i.d}{\sim} N(0, \psi_1^2)$ and $\delta_{2i} \overset{i.i.d}{\sim} N(0, \psi_2^2)$, for $i = 1, 2, \cdots, n$. Then, model (2.4) reduces to $Y_i = h(\mathbf{X}_i) + \epsilon_i$, for $i = 1, 2, \cdots, n$.

By specifying $h(\cdot)$ as a linear function or a partially linear function, we have the following special cases of model (2.4) with independent and correlated study units:

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

- *Special cases of model* (2.4) *for independent study units.*

$$\text{LRM-Indpt: } Y_i = \beta_0 + \mathbf{X}_i'\boldsymbol{\beta}_1 + \epsilon_i, \tag{2.4a}$$

$$\text{PLRM-Indpt: } Y_i = \mathbf{X}_i^{\dagger\prime}\boldsymbol{\beta}^{\dagger} + h_0(\mathbf{X}_i^{\ddagger}) + \epsilon_i, \tag{2.4b}$$

  where LRM and PLRM stand for linear regression model and partially linear regression model, respectively.

- *Special cases of model* (2.4) *for correlated study units.*

$$\text{LRM-Corrltd: } Y_i = \beta_0 + \mathbf{X}_i'\boldsymbol{\beta}_1 + \delta_{1i}\beta_2 + \delta_{2i}\beta_3 + \epsilon_i, \tag{2.4c}$$

$$\text{PLRM-Corrltd: } Y_i = \beta_0 + \mathbf{X}_i^{\dagger\prime}\boldsymbol{\beta}_1^{\dagger} + \delta_{1i}\beta_2^{\dagger} + \delta_{2i}\beta_3^{\dagger} + h_0(\mathbf{X}_i^{\ddagger}) + \epsilon_i, \tag{2.4d}$$

  where $\mathbf{X}_i^{\dagger}$ and $\mathbf{X}_i^{\ddagger}$ are subsets of the covariates $\mathbf{X}_i$.

2.4 Induced Distribution of First Hitting Time

Under the Wiener process model (2.2), the reporting delay S_i may be viewed as the FHT, that is, the time when the process $A_i(\cdot)$ first reaches the threshold B_i, which is the area burned at the report time: $S_i = \inf\{u : u > 0, A_i(u) > B_i\}$, which is the same as $\sup\{u : u > 0, A_i(u) < B_i\}$, almost surely. The FHT S_i follows an inverse Gaussian (IG) distribution (e.g., Chhikara and Folks, 1989) with the cumulative distribution function $G(u|B_i, \mathbf{X}_i; \mu_i, \lambda_i)$:

$$\Phi\left(\sqrt{\frac{\lambda_i}{u}}\left[\frac{u}{\mu_i} - 1\right]\right) + \exp\left(2\lambda_i/\mu_i\right)\Phi\left(-\sqrt{\frac{\lambda_i}{u}}\left[\frac{u}{\mu_i} + 1\right]\right), \tag{2.5}$$

where $\mu_i = B_i/\nu e^{\mathbf{X'}_i\boldsymbol{\gamma} + X_i^*\delta_{2i} + \delta_{1i}}$, $\lambda_i = B_i^2/\sigma^2$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. We denote the cumulative distribution of S_i by $G(u|B_i, \mathbf{X}_i, \delta_{1i}, \delta_{2i}, \nu, \sigma, \boldsymbol{\gamma})$.

On the other hand, the unobserved reporting delay S_i is a portion of the event duration L_i, the response variable in the desired regression analysis. By using the related longitudinal measures and the induced distribution of the first hitting time, we can overcome the inherent difficulty of the unobserved duration in the regression

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

analysis. Moreover, $(\delta_{1i}, \delta_{2i})$, the shared random effects/frailty variables, are used to connect the event duration and longitudinal measures and to capture the spatial correlations between individuals.

3. Estimation in the Presence of Missing Time Origins

In this section, we estimate the conditional distribution of the event duration $F(t|\mathbf{x}) = P(L \leq t|\mathbf{x})$ when the study units are independent or spatially correlated. We must estimate the regression function $h(\cdot)$ and the distribution of the random error. We start with the estimation procedure with independent units, and then adapt the procedure for spatially correlated units.

3.1 Estimation with Independent Study Units

The following two assumptions are made throughout this section:

ASSUMPTION (A1). *$\{(T_{Si}, T_{Ri}, T_{Fi}, B_i, D_i, \mathbf{X}_i, \boldsymbol{\omega}_i) : i = 1, 2, \cdots, n\}$ is a collection of i.i.d. realizations of $\left(T_S, T_R, T_F, B, D, \mathbf{X}, \boldsymbol{\omega}\right)$.*

ASSUMPTION (A2). *For $i = 1, 2, \cdots, n$, $L_i^* = T_{Fi} - T_{Ri} = L_i - S_i$ and $S_i = T_{Ri} - T_{Si}$ are conditionally independent given $(\delta_{1i}, \delta_{2i})$ and $\mathbf{X}_i$. In addition, B_i and $(\delta_{1i}, \delta_{2i})$ are conditionally independent given $\mathbf{X}_i$.*

In the wildfire application, the conditional independence assumption for S_i and L_i^* in (A2) is plausible, because a fire agency often assesses a reported fire in terms of its spread rate (ν_i), and then plans the initial attack accordingly. Because ν_i is specified depending on $\mathbf{X}_i$ and $(\delta_{1i}, \delta_{2i})$, L_i^* is likely associated with S_i solely through $\mathbf{X}_i$ and $(\delta_{1i}, \delta_{2i})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

3.1.1 Proposed Estimator for $F(\cdot|\mathbf{x})$

From the regression models LRM-Indpt (2.4a) and PLRM-Indpt (2.4b), $P(L \leq t|\mathbf{X}=\mathbf{x}) = F_\epsilon\big(\log(t) - h(\mathbf{x})\big)$. When all the durations L_i are observed, $F_\epsilon(\cdot)$ can be estimated using the empirical function $F_{\epsilon,n}(\zeta) = n^{-1}\sum_{i=1}^n I(\epsilon_i \leq \zeta)$, and can be written as $n^{-1}\sum_{i=1}^n I(S_i \leq e^{\zeta+h(\mathbf{X}_i)} - L_i^*)$.

Note that $\mathrm{E}\{I(\epsilon_i \leq \zeta)|\mathcal{O}_i\} = P(\epsilon_i \leq \zeta|\mathcal{O}_i) = P\Big(S_i \leq e^{\zeta+h(\mathbf{X}_i)} - L_i^*|\mathcal{O}_i\Big)$. With model (2.2) and Assumption (A2), $P\Big(S_i \leq e^{\zeta+h(\mathbf{X}_i)} - L_i^*|\mathcal{O}_i\Big)$ is

$$\iint G\Big(e^{\zeta+h(\mathbf{X}_i)} - L_i^*|B_i, \mathbf{X}_i, \delta_1, \delta_2; \nu, \sigma, \boldsymbol{\gamma}\Big)\phi_1(\delta_1|\mathcal{O}_i;\boldsymbol{\theta})\phi_2(\delta_2|\mathcal{O}_i;\boldsymbol{\theta})\mathrm{d}\delta_1\mathrm{d}\delta_2, \tag{3.1}$$

where $\boldsymbol{\theta} = (\nu, \sigma, \boldsymbol{\gamma}, \psi_1, \psi_2)'$, and $\phi_1(\cdot|\mathcal{O}_i;\boldsymbol{\theta})$ and $\phi_2(\cdot|\mathcal{O}_i;\boldsymbol{\theta})$ are the conditional distributions of δ_1 and δ_2, respectively, given the observed data $\mathcal{O}_i$. If $\boldsymbol{\theta}$ and $h(\cdot)$ are known, we can estimate $F_\epsilon(\zeta)$ by $\tilde{F}_{\epsilon,n}\big(\zeta;\boldsymbol{\theta},h(\cdot)\big) = n^{-1}\sum_{i=1}^n P\Big(S_i \leq e^{\zeta+h(\mathbf{X}_i)} - L_i^*|\mathcal{O}_i\Big)$. Thus, $P(L \leq t|\mathbf{X}=\mathbf{x}) = F_\epsilon\big(\log t - h(\mathbf{x})\big)$ can be estimated by $\tilde{F}_n\big(t|\mathbf{x};\boldsymbol{\theta},h(\cdot)\big) = \tilde{F}_{\epsilon,n}\big(\log t - h(\mathbf{x});\boldsymbol{\theta},h(\cdot)\big)$. After substituting the estimators for $\boldsymbol{\theta}$ and $h(\cdot)$, we obtain the following estimator for $F(t|\mathbf{x})$:

$$\hat{F}_n(t|\mathbf{x}) = \tilde{F}_n\big(t|\mathbf{x};\hat{\boldsymbol{\theta}},\hat{h}(\cdot)\big) = \tilde{F}_{\epsilon,n}\big(\log t - h(\mathbf{x});\hat{\boldsymbol{\theta}},\hat{h}(\cdot)\big). \tag{3.2}$$

We present the procedure for estimating $\boldsymbol{\theta}$ and $h(\cdot)$ in the next subsection.

3.1.2 Estimation Procedure for $\boldsymbol{\theta}$ and $h(\cdot)$

We now present a two-stage procedure for estimating $\boldsymbol{\theta}$ and $h(\cdot)$ under the independence assumption.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Stage I. Estimation of $\boldsymbol{\theta}$: Let $\boldsymbol{\theta} = (\nu, \sigma, \boldsymbol{\gamma}, \psi_1, \psi_2)$. Conditional on the covariates $\mathbf{X}$, the log-likelihood function with the observed data (2.1) is

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \iiint \left\{ L_{obs,i|S,\delta_1,\delta_2,\mathbf{X}_i} \right\} \mathrm{d}[S, \delta_1, \delta_2 | \mathbf{X}_i], \tag{3.3}$$

where $L_{obs,i|S,\delta_1,\delta_2,\mathbf{X}_i} = [D_i|B_i, L_i^*, \delta_1, \delta_2, \mathbf{X}_i][B_i|S, \delta_1, \delta_2, \mathbf{X}_i]$.

Let the "full data set" be the observed data (2.1), augmented by $\underline{S} = \{S_i, i = 1, 2, \cdots, n\}$, $\underline{\delta_1} = \{\delta_{1i}, i = 1, 2, \cdots, n\}$, and $\underline{\delta_2} = \{\delta_{2i}, i = 1, 2, \cdots, n\}$. The log-likelihood function conditional on the covariates $\mathbf{X}$ with the full data is given by $l_F(\boldsymbol{\theta}|\text{Observed-data}, \underline{S}, \underline{\delta_1}, \underline{\delta_2}) = l_{F_1}(\nu, \sigma, \boldsymbol{\gamma}|\underline{S}, \underline{\delta_1}, \underline{\delta_2}) + l_{F_2}(\boldsymbol{\theta}; \underline{S}, \underline{\delta_1}, \underline{\delta_2})$, where $l_{F_1}(\nu, \sigma, \boldsymbol{\gamma}|\underline{S}, \underline{\delta_1}, \underline{\delta_2})$ is

$$-n \log \sigma^2 - \sum_{i=1}^{n} \frac{\left(D_i - \nu e^{\mathbf{X}_i'\boldsymbol{\gamma} + X_i^* \delta_{2i} + \delta_{1i}} L_i^*\right)^2}{2\sigma^2 L_i^*} - \sum_{i=1}^{n} \frac{\left(B_i - \nu e^{\mathbf{X}_i'\boldsymbol{\gamma} + X_i^* \delta_{2i} + \delta_{1i}} S_i\right)^2}{2\sigma^2 S_i},$$

and

$$l_{F_2}(\boldsymbol{\theta}; \underline{S}, \underline{\delta_1}, \underline{\delta_2}) = \sum_{i=1}^{n} \log[S_i|\delta_{1i}, \delta_{2i}, \mathbf{X}_i] + \sum_{i=1}^{n} \log \phi_1(\delta_{1i}; \psi_1) + \sum_{i=1}^{n} \log \phi_2(\delta_{2i}; \psi_2).$$

The following Monte Carlo EM algorithm is used to compute the MLE $\hat{\boldsymbol{\theta}}_n$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Algorithm A

1: Assume that we have the estimate $\boldsymbol{\theta}^{(m)}$ at the mth iteration ($m \geq 0$), with $\boldsymbol{\theta}^{(0)}$ for the initial value.

2: **repeat**

2.1: *E-step.* Approximate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathrm{E}\{l_F(\boldsymbol{\theta}|\text{Observed-data}, \underline{S}, \underline{\delta_1}, \underline{\delta_2})|\mathcal{O}; \boldsymbol{\theta}^{(m)}\}$ using the sample mean $\frac{1}{J}\sum_{j=1}^{J} l_F(\boldsymbol{\theta}|\text{Observed-data}, \underline{S}^{(j)}, \underline{\delta_1}^{(j)}, \underline{\delta_2}^{(j)})$, which is given by

$$\frac{1}{J}\sum_{j=1}^{J} l_{F_1}(\nu, \sigma, \boldsymbol{\gamma}|\underline{S}^{(j)}, \underline{\delta_1}^{(j)}, \underline{\delta_2}^{(j)}) + \frac{1}{J}\sum_{j=1}^{J} l_{F_2}(\boldsymbol{\theta}; \underline{S}^{(j)}, \underline{\delta_1}^{(j)}, \underline{\delta_2}^{(j)}), \tag{3.4}$$

for $j = 1, \cdots, J$, and $(\underline{S}^{(j)}, \underline{\delta_1}^{(j)}, \underline{\delta_2}^{(j)})$ is generated from the conditional distribution given the observed data $[\underline{S}, \underline{\delta_1}, \underline{\delta_2}|\mathcal{O}; \boldsymbol{\theta}^{(m)}]$. This distribution can also be written as $[\underline{S}|\underline{\delta_1}, \underline{\delta_2}, \mathcal{O}; \boldsymbol{\theta}^{(m)}][\underline{\delta_1}, \underline{\delta_2}|\mathcal{O}; \boldsymbol{\theta}^{(m)}]$.

- *E-step (i).* Generate $(\delta_{1i}^{(j)}, \delta_{2i}^{(j)})'$ from $\phi_i^{(m+1)}(\delta_1, \delta_2|\mathcal{O}_i; \boldsymbol{\theta}^{(m)})$, which is the conditional distribution of (δ_1, δ_2), given the observed data with the current parameter estimate $\boldsymbol{\theta}^{(m)}$,

$$\frac{L_{obs,i|\delta_1,\delta_2,\mathbf{X}_i}(\nu^{(m)}, \sigma^{(m)}, \boldsymbol{\gamma}^{(m)}; \delta_1, \delta_2)\phi_1(\delta_1; \psi_1^{(m)})\phi_2(\delta_2; \psi_2^{(m)})}{\iint L_{obs,i|\delta_1,\delta_2,\mathbf{X}_i}(\nu^{(m)}, \sigma^{(m)}, \boldsymbol{\gamma}^{(m)}; \delta)\phi_1(\delta_1; \psi_1^{(m)})\phi_2(\delta_2; \psi_2^{(m)})\mathrm{d}\delta_1\mathrm{d}\delta_2},$$

  with $L_{obs,i|\delta_1,\delta_2,\mathbf{X}_i} = \int\{L_{obs,i|S,\delta_1,\delta_2,\mathbf{X}_i}\}d[S|\delta_1, \delta_2, \mathbf{X}_i]$. The Metropolis–Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) can be applied to generate $(\delta_{1i}^{(j)}, \delta_{2i}^{(j)})$ with the proposed distribution being $(\delta_{1i}^{(new)}, \delta_{2i}^{(new)})' \sim MVN\Big((\delta_{1i}^{(old)}, \delta_{2i}^{(old)})', 0.5^2 I_{2\times 2}\Big)$.
- *E-step (ii).* Under Assumption (A1), we can generate $S_i^{(j)}$, the ith component of $\underline{S}^{(j)}$, independently from the conditional distribution of S_i, given the observed data with $\boldsymbol{\theta}^{(m)}$ and $\delta_{1i}^{(j)}, \delta_{2i}^{(j)}$. This distribution, $[S_i|\delta_{1i}, \delta_{2i}, \mathcal{O}_i; \boldsymbol{\theta}^{(m)}]$, is in fact the IG distribution given in (2.5), with $\nu = \nu^{(m)}, \sigma = \sigma^{(m)}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(m)}$.

2.2: *M-step.* Obtain the updated $\boldsymbol{\theta}^{(m+1)}$ by maximizing (3.4).

3: **until** $\{\boldsymbol{\theta}^{(m)} : m = 1, 2, \cdots\}$ **converges.**

Because $[S_i|\delta_{1i}, \delta_{2i}, \mathbf{X}_i]$ in $l_{F_2}(\boldsymbol{\theta}; \underline{S}, \underline{\delta_1}, \underline{\delta_2})$ does not have much additional information on the parameters ν, σ and $\boldsymbol{\gamma}$, the objective function $l_F(\boldsymbol{\theta}|\text{Observed-data}, \underline{S}, \underline{\delta_1}, \underline{\delta_2})$ can be replaced by $l_{F_1}(\nu, \sigma, \boldsymbol{\gamma}|\underline{S}, \underline{\delta_1}, \underline{\delta_2}) + \log\phi_1(\psi_1; \underline{\delta_1}) + \log\phi_2(\psi_2; \underline{\delta_2})$. This maximization procedure leads to $\tilde{\boldsymbol{\theta}}_n$, a close approximation to the MLE $\hat{\boldsymbol{\theta}}_n$. Therefore, the *M-step* can be implemented by solving $\sum_{j=1}^{J} \frac{\partial l_{F_1}}{\partial(\nu,\sigma,\boldsymbol{\gamma})}(\nu, \sigma, \boldsymbol{\gamma}|\underline{S}^{(j)}, \underline{\delta_1}^{(j)}, \underline{\delta_2}^{(j)})\Big/J = 0$, $\sum_{j=1}^{J}\sum_{i=1}^{n} \frac{\partial \log\phi_1}{\partial\psi_1}(\delta_{1i}^{(j)}; \psi_1)\Big/J = 0$ and $\sum_{j=1}^{J}\sum_{i=1}^{n} \frac{\partial \log\phi_2}{\partial\psi_2}(\delta_{2i}^{(j)}; \psi_2)\Big/J = 0$. This algorithm results in $\tilde{\boldsymbol{\theta}}_n$, a close approximation to the MLE $\hat{\boldsymbol{\theta}}_n$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Stage II. Estimation of $h(\cdot)$: Given that Model LRM-Indpt (2.4a) uses a linear function of the covariates to approximate $h(\cdot)$, the estimator of $h(\cdot)$ only requires an estimation of the parameter $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)$. If all the durations L_i were observed, a consistent estimator for $\boldsymbol{\beta}$ would be the least squares estimator $\hat{\boldsymbol{\beta}}_{LSE} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y}$, where $\mathbf{Y} = (\log L_1, \cdots, \log L_n)'$ and $\mathbb{X}$ is a matrix with the ith row being $\mathtt{X}_i = (1, \mathbf{X}_i')'$. Although L_i is not available in the data, we have $L_i = S_i + L_i^*$, with L_i^* observed, and the distribution of S_i conditional on the observed data being the IG distribution. We then consider the estimator $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = E[\hat{\boldsymbol{\beta}}_{LSE}|\mathbf{O}; \boldsymbol{\theta}]$. We replace $\boldsymbol{\theta}$ with the estimator $\hat{\boldsymbol{\theta}}_n$ from ALGORITHM A to obtain the estimator $\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_n) = \mathrm{E}[\hat{\boldsymbol{\beta}}_{LSE}|\mathcal{O}; \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}]$, denoted as $\hat{\boldsymbol{\beta}}_n$. The following algorithm, based on the Monte Carlo method, is used to compute $\hat{\boldsymbol{\beta}}_n$.

Algorithm B1

1: Assume that we have $\hat{\boldsymbol{\theta}}_n = (\hat{\nu}_n, \hat{\sigma}_n, \hat{\boldsymbol{\gamma}}_n, \hat{\psi}_{1n}, \hat{\psi}_{2n})$ obtained from ALGORITHM A.
2: *Step 1.* Generate $(\delta_{1i}^{(j)}, \delta_{2i}^{(j)})$ from $\phi_i(\delta_1, \delta_2|\mathcal{O}_i; \hat{\boldsymbol{\theta}}_n)$, for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, J^*$.
3: *Step 2.* Generate $\tilde{S}_i^{(j)}$ from the IG distribution with $\mu_i = B_i/(\hat{\nu}_n e^{\mathbf{X}_i'\hat{\boldsymbol{\gamma}}_n + X_i^*\delta_{2i}^{(j)} + \delta_{1i}^{(j)}}), \lambda_i = B_i^2/\hat{\sigma}_n^2$.
4: *Step 3.* Obtain $\tilde{\mathbf{Y}}^{(j)} = \big(\log(L_1^* + \tilde{S}_1^{(j)}), \cdots, \log(L_n^* + \tilde{S}_n^{(j)})\big)'$, for $j = 1, \cdots, J^*$, and compute the estimator $\mathrm{E}[\hat{\boldsymbol{\beta}}_{LSE}|\mathcal{O}; \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}]$ using $\hat{\boldsymbol{\beta}}_{n,J^*} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\Big(\sum_{j=1}^{J^*} \tilde{\mathbf{Y}}^{(j)}/J^*\Big)$.

The output of this algorithm, $\hat{\boldsymbol{\beta}}_{n,J^*}$ is a close approximation to the estimator $\hat{\boldsymbol{\beta}}_n$ when J^* is sufficiently large.

Under model PLRM-Indpt (2.4b), the estimation of $h(\cdot)$ requires estimating both the parametric component, that is, the parameter $\boldsymbol{\beta}^\dagger$, and the nonparametric function $h_0(\cdot)$. When all the durations are available, we can follow Speckman (1988) to derive estimators of the parametric and nonparametric components using an LSE with kernel smoothing: $\hat{\boldsymbol{\beta}}_{speck}^\dagger = (\breve{\mathbb{X}}_{\mathbf{d}}^{\dagger\prime}\breve{\mathbb{X}}_{\mathbf{d}}^\dagger)^{-1}\breve{\mathbb{X}}_{\mathbf{d}}^\dagger\breve{\mathbf{Y}}_{\mathbf{d}}$, and $\hat{h}_{0,speck}(\mathbf{x}^\ddagger) = \frac{\sum_{i=1}^n K_{\mathbf{d}}(\mathbf{x}^\ddagger - \mathbf{X}_i^\ddagger)(Y_i - \mathbf{X}_i^{\dagger\prime}\hat{\boldsymbol{\beta}}_{speck}^\dagger)}{\sum_{i=1}^n K_{\mathbf{d}}(\mathbf{x}^\ddagger - \mathbf{X}_i^\ddagger)}$, where $\mathbb{X}^\dagger$ is a matrix with the ith row being $\mathbf{X}_i^{\dagger\prime}$, $\breve{\mathbb{X}}_{\mathbf{d}}^\dagger =$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

$(\mathbf{I}-\mathbf{H_d})\mathbb{X}^\dagger$, and $\breve{\mathbf{Y}}_\mathbf{d} = (\mathbf{I}-\mathbf{H_d})\mathbf{Y}$. Here, $K_\mathbf{d}(\cdot)$ is a kernel function with bandwidth $\mathbf{d}$, and $\mathbf{H_d}$ is a smoothing matrix with the (i,j)th element being $K_\mathbf{d}(\mathbf{X}_i^\ddagger - \mathbf{X}_j^\ddagger)\Big/\sum_{j=1}^n K_\mathbf{d}(\mathbf{X}_i^\ddagger - \mathbf{X}_j^\ddagger)$. The fitted values $\hat{\mathbf{Y}} = (\hat{Y}_1, \cdots, \hat{Y}_n)'$ can be obtained by $\hat{\mathbf{Y}} = \Big(\mathbf{H_d} + \breve{\mathbb{X}}_\mathbf{d}^\dagger(\breve{\mathbb{X}}_\mathbf{d}^{\dagger\prime}\breve{\mathbb{X}}_\mathbf{d}^\dagger)^{-1}\breve{\mathbb{X}}_\mathbf{d}^{\dagger\prime}(\mathbf{I}-\mathbf{H_d})\Big)\mathbf{Y}$. Let $\mathbf{A}_d$ be the smoothing matrix calculated with the bandwidth $\mathbf{d}$ and $\mathbf{A_d} = \big(\mathbf{H_d} + \breve{\mathbb{X}}_\mathbf{d}^\dagger(\breve{\mathbb{X}}_\mathbf{d}^{\dagger\prime}\breve{\mathbb{X}}_\mathbf{d}^\dagger)^{-1}\breve{\mathbb{X}}_\mathbf{d}^{\dagger\prime}(\mathbf{I}-\mathbf{H_d})\big)$. We use the generalized cross-validation (GCV) criterion (e.g., Loader, 1999) to select the bandwidth. The GCV function is

$$GCV(\mathbf{d}) = \frac{||(\mathbf{I}-\mathbf{A_d})\mathbf{Y}||^2}{[1-n^{-1}trace(\mathbf{A_d})]^2}. \tag{3.5}$$

Following the idea of ALGORITHM B1, we can also consider estimators $\hat{\boldsymbol{\beta}}_n^\dagger = \mathrm{E}[\hat{\boldsymbol{\beta}}_{speck}^\dagger|\mathcal{O};\hat{\boldsymbol{\theta}}_n]$ and $\hat{h}_{0n}(\cdot) = \mathrm{E}[\hat{h}_{0,speck}(\cdot)|\mathcal{O};\hat{\boldsymbol{\theta}}_n]$ when the duration is not observed. We can then adapt *Step 3* in ALGORITHM B1 to the PLRM-Indpt (2.4b). This yields the following algorithm.

Algorithm B2

1: Assume that we have $\hat{\boldsymbol{\theta}}_n$ obtained from ALGORITHM A.
2: *Steps 1 and 2.* Follow *Steps 1 and 2* of ALGORITHM B1 to generate $(\delta_{1i}^{(j)}, \delta_{2i}^{(j)})$, and $S_i^{(j)}$, for $i = 1, \cdots, n$ and $j = 1, \cdots, J^*$.
3: *Step 3.* For $j = 1, 2, \cdots, J^*$, obtain the vector $\tilde{\mathbf{Y}}^{(j)}$ with the ith element being $\tilde{Y}_i^{(j)} = \log(L_i^* + \tilde{S}_i^{(j)})$.

- *Step 3(i)* For $j = 1, \cdots, J^*$, compute the bandwidth $\mathbf{d}^{(j)}$ with $\{\tilde{Y}_i^{(j)}, i = 1, \cdots, n\}$ using the GCV function in (3.5). Compute the vector $\breve{\tilde{\mathbf{Y}}}_{\mathbf{d}^{(j)}}^{(j)} = (\mathbf{I}-\mathbf{H}_{\mathbf{d}^{(j)}})\tilde{\mathbf{Y}}^{(j)}$, where the smoothing matrix $\mathbf{H}_{\mathbf{d}^{(j)}}$ is obtained using the selected bandwidth $\mathbf{d}^{(j)}$.
- *Step 3(ii)* The estimator $\hat{\boldsymbol{\beta}}_n^\dagger$ can be approximated by $\hat{\boldsymbol{\beta}}_{n,J^*}^\dagger = \sum_{j=1}^{J^*} \Big((\breve{\mathbb{X}}_{\mathbf{d}^{(j)}}^{\dagger\prime}\breve{\mathbb{X}}_{\mathbf{d}^{(j)}}^\dagger)^{-1}\breve{\mathbb{X}}_{\mathbf{d}^{(j)}}^{\dagger\prime}\breve{\tilde{\mathbf{Y}}}^{(j)}\Big)\Big/J^*$, and $\hat{h}_{0n}(\cdot)$ is approximated by $\hat{h}_{0n,J^*}(\mathbf{x}^\ddagger) = \frac{1}{J^*}\sum_{j=1}^{J^*}\frac{\sum_{i=1}^n K_{\mathbf{d}^{(j)}}(\mathbf{x}^\ddagger - \mathbf{X}_i^\ddagger)(\tilde{Y}_i^{(j)} - \mathbf{X}_i^{\dagger\prime}\hat{\boldsymbol{\beta}}_{n,J^*}^\dagger)}{\sum_{i=1}^n K_{\mathbf{d}^{(j)}}(\mathbf{x}^\ddagger - \mathbf{X}_i^\ddagger)}$.

When J^* is sufficiently large, the estimators $\hat{\boldsymbol{\beta}}_{n,J^*}^\dagger$ and $\hat{h}_{0n,J^*}$ are close approximations to $\hat{\boldsymbol{\beta}}_n^\dagger$ and $\hat{h}_{0n}(\cdot)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Alternatively, we can estimate the nonparametric function $h_0(\cdot)$ using splines. For example, to estimate $h_0(x^\ddagger)$ using a natural cubic spline with M knots, we may use $h_0(x^\ddagger) = \sum_{j=1}^{M} \alpha_j b_j(x^\ddagger)$. The associated spline coefficients $\alpha_1, \cdots, \alpha_M$ can then be estimated together with $\boldsymbol{\beta}^\dagger$ in the regression function by the LSE. This yields an estimator of $F(t|\mathbf{x})$ if we substitute the obtained estimators $\hat{\boldsymbol{\theta}}_n$ and $\hat{h}(\cdot)$ into $\tilde{F}_n(t|\mathbf{x}; \boldsymbol{\theta}, h(\cdot))$. The estimator involves the double integral presented in (3.1). We can compute this numerically as $\sum_{j=1}^{J^*} G\Big(e^{\zeta+h(\mathbf{X}_i)} - L_i^*|B_i, \mathbf{X}_i, \delta_{1i}^{(j)}, \delta_{2i}^{(j)}; \nu, \sigma, \boldsymbol{\gamma}\Big) \Big/ J^*$. Here, $(\delta_{1i}^{(j)}, \delta_{2i}^{(j)})$ are obtained from *Step 1* of ALGORITHMS B1 and B2, for $i = 1, \cdots, n$ and $j = 1, \ldots, J^*$.

3.1.3 Asymptotic Properties

This section studies the asymptotic properties of the proposed estimators and presents the variance estimation. The derivation of the asymptotic properties is outlined in Section 1 of the Supplementary Material.

We use $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_0$ to represent the true values of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, respectively, under the model LRM-Indpt (2.4a). The proposed estimators $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ and the estimator for the conditional probability using these estimators, $\hat{F}_n(t|\mathbf{x})$, have the following asymptotic properties.

Theorem 1 *Under assumptions (A1)–(A2) and conditions (C1)–(C7) in the Appendix, $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\beta}}_n$ have the following properties:*

(i) Strong Consistency*: $||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|| \to 0$, $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| \to 0$ almost surely as $n \to \infty$.*

(ii) Asymptotic Normality*: $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\Big(0, AV(\boldsymbol{\theta}_0)\Big)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N\Big(0, AV(\boldsymbol{\beta}_0)\Big)$, where the asymptotic variances are given by*

$$
\begin{aligned}
&AV(\boldsymbol{\theta}_0) = \Pi^{-1}(\boldsymbol{\theta}_0)\Sigma(\boldsymbol{\theta}_0)\Pi^{-1}(\boldsymbol{\theta}_0), \text{ and} \\
&AV(\boldsymbol{\beta}_0) = \mathrm{E}^{-1}[\mathtt{X}_i\mathtt{X}_i']\mathrm{Var}\Big(\mathtt{X}_i\big(\mathrm{E}[\log L_i|\mathcal{O}_i; \boldsymbol{\theta_0}] - \mathtt{X}_i'\boldsymbol{\beta}_0\big)\Big) E^{-1}[\mathtt{X}_i\mathtt{X}_i'], \qquad (3.6)
\end{aligned}
$$

with $\Pi(\boldsymbol{\theta}) = \mathrm{E}[-\partial l_{o_i}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}]$, $\Sigma(\boldsymbol{\theta}) = \mathrm{Var}(\partial l_{o_i}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Note that the robust sandwich variance estimator of $\hat{\boldsymbol{\theta}}_n$ is $V(\hat{\boldsymbol{\theta}}_n)$, where

$$V(\boldsymbol{\theta}) = \left(-\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right)^{-1} \left(\left[\frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]'\right) \left(-\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right)^{-1}.$$

The variance estimator of $\hat{\boldsymbol{\beta}}_n$ is

$$V(\hat{\boldsymbol{\beta}}_n) = \Big(\sum_{i=1}^{n} \mathtt{X}_i \mathtt{X}_i'\Big)^{-1} \mathrm{Var}\Big(\sum_{i=1}^{n} \mathtt{X}_i (\mathrm{E}[\log L_i | \mathcal{O}_i; \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}] - \mathtt{X}_i' \hat{\boldsymbol{\beta}}_n)\Big) \Big(\sum_{i=1}^{n} \mathtt{X}_i \mathtt{X}_i'\Big)^{-1}.$$

As presented in ALGORITHM B1, we impute possible values for the duration by generating the reporting delay S_i, for $i = 1, \cdots, n$, from the IG distribution using the estimated parameters from ALGORITHM A and generating δ_{1i} and δ_{2i} from the conditional distribution. Each imputed data set yields a least squares estimator for $\boldsymbol{\beta}$. Following Goetghebeur and Ryan (2000), we can estimate the variance of $\hat{\boldsymbol{\beta}}_n$ using a weighted sum of the empirical variance of the imputation estimates and the mean of the imputation variances. The weights are $1 + 1/J^*$ and 1, respectively, where J^* is the number of imputations used in ALGORITHM B. The estimated variance of $\hat{\boldsymbol{\beta}}_n$ is

$$(1 + 1/J^*) \frac{1}{J^* - 1} \sum_{j=1}^{J^*} (\hat{\boldsymbol{\beta}}^{(j)} - \hat{\boldsymbol{\beta}}_n)^2 + \frac{1}{J^*} \sum_{j=1}^{J^*} \hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^{(j)}), \tag{3.7}$$

where $\hat{\boldsymbol{\beta}}^{(j)} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\tilde{\mathbf{Y}}^{(j)}$ is the least squares estimator, and $\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^{(j)})$ is the estimated variance of $\hat{\boldsymbol{\beta}}^{(j)}$ with the jth imputed data set. In our algorithm, we set $J^* = 200$.

Further, the proposed estimator $\hat{F}_n(t|\mathbf{x}) = \tilde{F}_{\epsilon,n}(\log t - \mathbf{x}'\boldsymbol{\beta}; \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{n}})$ has the following asymptotic property.

Theorem 2 *Under assumptions (A1)–(A2) and conditions (C1)–(C7) for the log-*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

likelihood function in (3.3), $\hat{F}_n(t|\mathbf{x})$ *has the following properties, with fixed* $\mathbf{x}$*:*

(i) Strong Consistency*:* $\sup_{t\in[0,\tau]} |\hat{F}_n(t|\mathbf{x}) - F(t|\mathbf{x})| \xrightarrow{p} 0$ *as* $n \to \infty$.

(ii) Weak Convergence*: For* $t \in [0,\tau]$*, as* $n \to \infty$*,* $\sqrt{n}(\hat{F}_n(t|\mathbf{x}) - F(t|\mathbf{x}))$ *converges weakly in* $\ell^\infty([0,\tau])$ *to a tight, mean-zero Gaussian process* $\mathcal{G}$*, with covariance* $Cov(\mathcal{G}(t|\mathbf{x}), \mathcal{G}(s|\mathbf{x}))$ *given by*

$$\begin{cases} \mathrm{E}\Big\{M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta_0}, \boldsymbol{\beta_0}) M(s, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta_0}, \boldsymbol{\beta_0})\Big\} - F(t|\mathbf{x})F(s|\mathbf{x}) & t \neq s, \\ \mathrm{E}\Big\{M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta_0}, \boldsymbol{\beta_0})^2\Big\} - F^2(t|\mathbf{x}) \\ \quad + \mathrm{E}_{\boldsymbol{\theta_0},\boldsymbol{\beta_0}}\left[\frac{\partial M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\theta}}\right]' AV(\boldsymbol{\theta}_0) \mathrm{E}_{\boldsymbol{\theta_0},\boldsymbol{\beta_0}}\left[\frac{\partial M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\theta}}\right] \\ \quad + \mathrm{E}_{\boldsymbol{\theta_0},\boldsymbol{\beta_0}}\left[\frac{\partial M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]' AV(\boldsymbol{\beta}_0) \mathrm{E}_{\boldsymbol{\theta_0},\boldsymbol{\beta_0}}\left[\frac{\partial M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] & t = s, \end{cases} \tag{3.8}$$

where $M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta})$ *is*

$$\iint G\Big(te^{-\mathbf{x}'\boldsymbol{\beta}_1 + \mathbf{X}_i'\boldsymbol{\beta}_1} - L_i^* | B_i, \mathbf{X}_i, \delta_1, \delta_2; \nu, \sigma, \boldsymbol{\gamma}\Big) \phi_1(\delta_1|\mathcal{O}_i; \boldsymbol{\theta}) \phi_2(\delta_2|\mathcal{O}_i; \boldsymbol{\theta}) \mathrm{d}\delta_1 \mathrm{d}\delta_2.$$

Note that $\mathrm{E}\Big\{M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta_0}, \boldsymbol{\beta_0})^2\Big\}$ can be approximated by the average $\sum_{i=1}^n [\sum_{j=1}^{J^*} G(te^{-\mathbf{x}\boldsymbol{\beta} + \mathbf{X}_i\boldsymbol{\beta}} - L_i^* | B_i, \mathbf{X}_i, \delta_{1i}^{(j)}, \delta_{2i}^{(j)}; \hat{\nu}_n, \hat{\sigma}_n, \hat{\boldsymbol{\gamma}}_n)/J^*]^2/n$, with $(\delta_{1i}^{(j)}, \delta_{2i}^{(j)})$ obtained from *Step 1* of Algorithm B1 for $j = 1, \cdots, J^*$. This strategy can be used to approximate $\mathrm{E}_{\boldsymbol{\theta_0},\boldsymbol{\beta_0}}[\partial M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta})/\partial \boldsymbol{\theta}]$ and $\mathrm{E}_{\boldsymbol{\theta_0},\boldsymbol{\beta_0}}[\partial M(t, L_i^*, B_i, \mathbf{X}_i; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}]$.

The proposed estimator of $\boldsymbol{\beta}^\dagger$ in the PLRM-Indpt (2.4b) with kernel smoothing also has the asymptotic properties.

Theorem 3 *Under assumptions (A1)–(A2) and conditions (C8)–(C10) in the Appendix,* $\hat{\boldsymbol{\beta}}_n^\dagger$ *has the following properties:*

(i) Strong Consistency*:* $||\hat{\boldsymbol{\beta}}_n^\dagger - \boldsymbol{\beta}_0^\dagger|| \to 0$ *almost surely as* $n \to \infty$.

(ii) Asymptotic Normality*:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^\dagger - \boldsymbol{\beta}_0^\dagger) \xrightarrow{d} N\Big(0, AV(\boldsymbol{\beta}_0^\dagger)\Big)$*, where the asymptotic variances are given by*

$$AV(\boldsymbol{\beta}_0^\dagger) = \sigma_\epsilon^2 \boldsymbol{\Omega}^{-1} \boldsymbol{\Xi} \boldsymbol{\Omega}^{-1}, \tag{3.9}$$

with $\boldsymbol{\Omega} = (\breve{\mathbb{X}}^{\dagger\prime}\breve{\mathbb{X}}^\dagger)^{-1}$, $\boldsymbol{\Xi} = \breve{\mathbb{X}}^{\dagger\prime}(\mathbf{I} - \mathbf{H_d})(\mathbf{I} - \mathbf{H_d})'\breve{\mathbb{X}}^\dagger$*, and* $\sigma_\epsilon^2 = \mathrm{Var}\Big(\mathrm{E}[\log L_i|\mathcal{O}_i] - \mathbf{X}_i^\dagger\boldsymbol{\beta}^\dagger - h_0(\mathbf{X}_i^\ddagger)\Big)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Using ALGORITHM B2, we can also estimate the variance of $\hat{\boldsymbol{\beta}}_n^{\dagger}$ using equation (3.7). Here, $\hat{\boldsymbol{\beta}}^{(j)} = (\breve{\mathbb{X}}^{\dagger\prime}_{\mathbf{d}^{(j)}}\breve{\mathbb{X}}^{\dagger}_{\mathbf{d}^{(j)}})^{-1}\breve{\mathbb{X}}^{\dagger\prime}_{\mathbf{d}^{(j)}}\breve{\breve{\mathbf{Y}}}^{(j)}$, $\hat{\text{Var}}(\hat{\boldsymbol{\beta}}^{(j)}) = \widetilde{\sigma^{(j)}_{\epsilon}{}^2}\boldsymbol{\Omega}^{-1}\boldsymbol{\Xi}\boldsymbol{\Omega}^{-1}$, and $\widetilde{\sigma^{(j)}_{\epsilon}{}^2} = \text{Var}\Big(\tilde{Y}_i^{(j)} - \mathbf{X}_i^{\dagger}\hat{\boldsymbol{\beta}}^{(j)} - h_0(\mathbf{X}_i^{\ddagger})\Big)$.

3.2 Estimation with Spatially Correlated Study Units

The estimation procedures (Section 3.1) for both the conditional distribution and the regression parameters can be extended to accommodate spatial correlation in the data. In addition to assumption (A2) given in Section 3.1, we make the following assumption throughout this section:

> ASSUMPTION (A1*). *Let* $\boldsymbol{\mathcal{X}} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)'$. *Conditional on* $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\mathcal{X}})$, *assume that* $(T_{Ri}, T_{Fi}, B_i, D_i)$ *is independent of* $(T_{Rj}, T_{Fj}, B_j, D_j)$, *for* $i, j = 1, \cdots, n$ *and* $i \neq j$.

Assumption $(A1^*)$ implies that realizations are independent, conditional on the covariates and the random effects. This is plausible, because we use correlated random effects to accommodate the spatial correlation. From model (2.4), the conditional distribution function $F(t|\mathbf{x}) = P(L_i \leq t|\mathbf{X}_i)$ can be written as $\iint F_{\epsilon}\Big(\log t - \big(h(\mathbf{x}, \delta_1, \delta_2)\big)\Big)\phi_1(\delta_1;\psi_1)\phi_2(\delta_2;\psi_2)d\delta_1 d\delta_2$. Here, $\phi_1(\cdot;\psi_1)$ and $\phi_2(\cdot;\psi_2)$ are density functions of the marginal distributions of δ_{1i} and δ_{2i}, respectively, for $i = 1, \cdots, n$, that is, the density functions for $N(0, \psi_1^2)$ and $N(0, \psi_2^2)$, respectively.

We employ the techniques of Section 3.1.1 and estimate $F_{\epsilon}(\zeta)$ by $\tilde{F}_{\epsilon,n}(\zeta;\boldsymbol{\theta}, h(\cdot)) = n^{-1}\sum_{i=1}^{n} P\Big(S_i \leq e^{\zeta + h(\mathbf{X}_i, \delta_1, \delta_2)} - L_i^*|\mathcal{O}_i\Big)$, with $P\Big(S_i \leq e^{\zeta + h(\mathbf{X}_i, \delta_1, \delta_2)} - L_i^*|\mathcal{O}_i\Big)$ being $\iint G\Big(e^{\zeta + h(\mathbf{X}_i, \delta_1, \delta_2)} - L_i^*|B_i, \mathbf{X}_i, \delta_1, \delta_2; \nu, \sigma, \boldsymbol{\gamma}\Big)\phi_1(\delta_1|\mathcal{O}_i;\boldsymbol{\theta})\phi_2(\delta_2|\mathcal{O}_i;\boldsymbol{\theta})\mathrm{d}\delta_1\mathrm{d}\delta_2$, with $\boldsymbol{\theta} = (\nu, \sigma, \boldsymbol{\gamma}, \psi_1, \rho_1, \psi_2, \rho_2)$ and $h(\cdot)$ specified in LRM-Corrltd (2.4c) and PLRM-Corrltd (2.4d). This estimator $\tilde{F}_{\epsilon,n}(\cdot;\boldsymbol{\theta}, h(\cdot))$ can then be used to estimate the conditional distribution. Denote the resulting estimator by $\tilde{F}_n\big(t|\mathbf{x};\boldsymbol{\theta}, h(\cdot)\big)$. Using the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

estimators of $\boldsymbol{\theta}$ and $h(\cdot)$, we can obtain the applicable estimator, denoted by $\hat{F}_n(t|\mathbf{x})$, using $\hat{F}_n(t|\mathbf{x}) = \tilde{F}_n(t|\mathbf{x}; \hat{\boldsymbol{\theta}}, \hat{h}(\cdot))$.

Under assumptions (A1*) and (A2), the log-likelihood function with observed data conditional on $\mathcal{X}$ is $\log L_{obs|\mathcal{X}}(\boldsymbol{\theta}) = \log \iint \left\{ L_{obs|\boldsymbol{\delta_1},\boldsymbol{\delta_2},\mathcal{X}}(\nu, \sigma, \boldsymbol{\gamma}) \right\} d[\boldsymbol{\delta}_1, \boldsymbol{\delta}_2]$, with $L_{obs|\boldsymbol{\delta_1},\boldsymbol{\delta_2},\mathcal{X}} = \prod_{i=1}^{n} \int_0^{\infty} \left\{ L_{obs|S,\delta_{1i},\delta_{2i},\mathbf{X}_i} \right\} d[S_i|\delta_{1i}, \delta_{2i}, \mathbf{X}_i]$ and $L_{obs|S,\delta_{1i},\delta_{2i},\mathbf{X}_i}$ being the product of the two densities $[D_i|B_i, L_i^*, \delta_{1i}, \delta_{2i}, \mathbf{X}_i]$ and $[B_i|S_i, \delta_{1i}, \delta_{2i}, \mathbf{X}_i]$. The MCEM algorithm can be employed to compute the MLE of $\boldsymbol{\theta}$.

We now consider the full data set, which is the observed data (2.1) augmented by $\underline{S} = \{S_i, i = 1, 2, \cdots, n\}$, $\underline{\delta_1} = \{\delta_{1i}, i = 1, 2, \cdots, n\}$, and $\underline{\delta_2} = \{\delta_{2i}, i = 1, 2, \cdots, n\}$. The full-data log-conditional likelihood function is $l_F^*(\boldsymbol{\theta}|\mathcal{O}, \underline{S}, \underline{\delta_1}, \underline{\delta_2}) = l_{F_1}^*(\nu, \sigma, \boldsymbol{\gamma}|\underline{S}, \underline{\delta_1}, \underline{\delta_2}) + l_{F_2}^*(\boldsymbol{\theta}; \underline{S}, \underline{\delta_1}, \underline{\delta_2})$, where $l_{F_1}^*$ is the same as l_{F_1} under the independence assumption, and $l_{F_2}^*(\boldsymbol{\theta}; \underline{S}, \underline{\delta}) = \sum_{i=1}^{n} \log[S_i|\delta_{1i}, \delta_{2i}, \mathbf{X}_i] - 1/2 \log|\Sigma(\psi_1, \rho_1)| - 1/2 \underline{\delta_1}' \Sigma^{-1}(\psi_1, \rho_1) \underline{\delta_1} - 1/2 \log|\Sigma(\psi_2, \rho_2)| - 1/2 \underline{\delta_2}' \Sigma^{-1}(\psi_2, \rho_2) \underline{\delta_2}$. The algorithm is similar to ALGORITHM A in Section 3.1.2; the main difference is the generation of $\underline{\delta_1}^{(j)}$ and $\underline{\delta_2}^{(j)}$ for the *E-step*. We generate the random vector $(\delta_{11}^{(j)}, \cdots, \delta_{1n}^{(j)}, \delta_{21}^{(j)}, \cdots, \delta_{2n}^{(j)})$, for $j = 1, \cdots, J$, from $\boldsymbol{\phi}^{(m+1)}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2|\mathcal{O}; \boldsymbol{\theta}^{(m)})$, which is the conditional distribution of $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ given the observed data with $\boldsymbol{\theta}^{(m)}$, and can be expressed as

$$\frac{L_{obs|\boldsymbol{\delta_1},\boldsymbol{\delta_2},\mathcal{X}}(\nu^{(m)}, \sigma^{(m)}, \boldsymbol{\gamma}^{(m)}; \boldsymbol{\delta_1}, \boldsymbol{\delta_2}) \boldsymbol{\phi}_1(\boldsymbol{\delta_1}; \psi_1^{(m)}, \rho_1^{(m)}) \boldsymbol{\phi}_2(\boldsymbol{\delta_2}; \psi_2^{(m)}, \rho_2^{(m)})}{\iint L_{obs|\boldsymbol{\delta_1},\boldsymbol{\delta_2},\mathcal{X}}(\nu^{(m)}, \sigma^{(m)}, \boldsymbol{\gamma}^{(m)}; \boldsymbol{\delta_1}, \boldsymbol{\delta_2}) \boldsymbol{\phi}_1(\boldsymbol{\delta_1}; \psi_1^{(m)}, \rho_1^{(m)}) \boldsymbol{\phi}_2(\boldsymbol{\delta_2}; \psi_2^{(m)}, \rho_2^{(m)}) \mathrm{d}\boldsymbol{\delta_1} d\boldsymbol{\delta_2}}, \quad (3.10)$$

where $\boldsymbol{\phi}_1(\cdot, \psi_1, \rho_1), \boldsymbol{\phi}_2(\cdot, \psi_2, \rho_2)$ are the density functions of multivariate normal distributions with the covariance functions specified in (2.3). The Metropolis–Hastings algorithm can be used to generate $(\delta_{11}^{(j)}, \cdots, \delta_{1n}^{(j)}, \delta_{21}^{(j)}, \cdots, \delta_{2n}^{(j)})$.

We use the regression model (2.4) to adjust for spatial correlation. ALGORITHMS B1 and B2 can be applied to estimate the regression function $h(\cdot)$ with models

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

LRM-Corrltd (2.4c) and PLRM-Corrltd (2.4d). The shared random effects can be predicted using the conditional distribution $\phi(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2|\mathcal{O}; \hat{\boldsymbol{\theta}})$, with $\hat{\boldsymbol{\theta}}$ obtained using the MCEM algorithm.

The robust sandwich variance estimator given in Theorem 1 can be used to estimate the variances of $\hat{\boldsymbol{\theta}}_n$. For the variance estimation of the regression parameters in models LRM-Corrltd (2.4c) and PLRM-Corrltd (2.4d), we apply the imputation methods described in Section 3.1.3.

4. Analysis of Alberta Wildfire Data

In this section, we conduct a regression analysis of Alberta wildfire data to demonstrate the proposed approach. The duration of interest is the ISA duration. The goal is to study the association between the ISA duration and a list of risk factors, and then to consider prediction based on this. The time of the initial attack is when the first fire-fighting resource arrives to prevent further extension of the wildfire. It is believed that fires with a longer ISA duration may require more effort to suppress, so this interval is valuable information for fire management agencies. Xiong, Braun, and Hu (2021) estimated the distribution of the ISA duration separately for fires from the upper and lower regions of Alberta, finding that the distribution depends on the region. To quantify the association of the ISA duration with the risk factors, we conduct a regression analysis.

The data set includes data on 603 fires caused by lightning in Alberta in 2004 during the fire season, May to August. There are six risk factors, *region (upper or lower) and fuel type (C1, C2, M2, Others)*, and three weather variables, *temperature, relative humidity, and wind speed*. All of the weather variables are recorded on the reported times of the fires. See Tables S1 and S2 in the Supplementary Material

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

for the summary statistics. The upper region covers Fort McMurray, High Level, Lac La Biche, Peace River, and Slave Lake; the lower region covers Calgary, Edson, Grande Prairie, Rocky Mountain House, and Whitecourt. The three most common fuel types are C1 (Spruce-lichen woodland), C2 (boreal spruce), and M2 (boreal mixedwood–green). The remaining fuel types are categorized into *Other Types*. For detailed characteristics and photographs of these fuel types, see Forestry Canada Fire Danger Group (1992). Fires that were large at the report time occurred more frequently in the upper region, and there were multiple large fires close to each other (see Figure S1 in the Supplementary Material for a map of the fires in the data set).

There appears to be a strong correlation between each pair of weather variables. In addition, *temperature* and *relative humidity* are closely associated with *region*; see the pairwise correlation plot in Figure S2 of the Supplementary Material, for example. Based on these findings, we select potential covariates via forward selection using the Akaike information criterion (AIC). As potential covariates, we consider *region, fuel type* and *wind speed*. The baseline category for *Region* is the *lower region*, and the baseline for *Fuel type* is *other types*.

We consider two cases of random drift ν_i in the longitudinal model (2.2): (1) with a random intercept $\nu_i = \nu e^{\mathbf{X'}_i\boldsymbol{\gamma}+\delta_{1i}}$, for $i = 1, \cdots, n$; and (2) with the mixed effect $\nu_i = \nu e^{\mathbf{X'}_i\boldsymbol{\gamma}+X^*{}_i\delta_{2i}+\delta_{1i}}$, where X^* denotes the factor *Region*. We estimate $\boldsymbol{\theta}$ and all parameters for model (2.2), assuming first that study units are independent, and then that they are spatially correlated. Table 1 presents the estimates for $\boldsymbol{\theta}$. The standard errors are estimated using the sandwich variance estimator, and the significant effects are presented in boldface. For the mixed effects, the estimates of ψ_1 and ψ_2 indicate considerable variation among fires; the variation also depends on the region. The estimated standard errors of the parameters for the spatially corre-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

lated case are smaller than those for the independent case, although the parameter estimates themselves are close.

The estimates of γ can be used to describe the effect of the risk factors on the burnt area process. The analysis that assumes spatial correlation indicates that fires in the upper region or those reported on days with a higher wind speed tend to have a larger drift and may grow faster. Furthermore, fires with M2 fuels tend to have a smaller drift than do fires with C1, C2, or Others. This finding agrees with prior analyses of the differences in burn rates among the fuel types of Alberta (Cumming, 2001; Tremblay, Duchesne, and Cumming, 2018).

—— *Table 1 is here.* ——

We consider LRMs and PLRMs with assumed independent and correlated study units. For the PLRMs, the covariates in the parametric component $\mathbf{X}^{\dagger}$ are the variables *Region* and *Fuel type*, and a smooth function of *Wind speed* is included. We use kernel smoothing and natural cubic splines to estimate the nonparametric term. Tables 2 and 3 present the estimates of the regression parameters in the LRMs and PLRMs with kernel smoothing. We also estimate the LRM parameters using two conventional approaches. The first uses L^* for the regression analysis, and the second treats the fire data as interval-censored, that is, $L_i \in [L_i^*, L_i^*+R_{max}]$, with R_{max} being the longest reporting delay. We set $R_{max} = 12, 48$ hours. In the interval-censored case, we use the R package *smoothsurv* (Komárek and Komárek, 2015) to estimate $\boldsymbol{\beta}$. For the correlated units, we use the regression model given in PLRM-Corrltd (2.4d) with two conventional approaches, where $(\delta_{1i}, \delta_{2i})$ can be obtained as a realization from the last iteration of the MCEM algorithm in Section 3.1.2. The results for both the LRM and the PLRM indicate that *region* is a significant predictor of ISA duration, whereas the two conventional approaches do not reveal this significant

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

association. The positive estimate of β_{region} suggests that fires occurring in the upper region tend to have a longer ISA duration, which is contrary to the estimates based on the observed ISA duration. The wildfire areas in the lower region are closer to Edmonton and Calgary, which are Alberta's two metropolitan areas with populations exceeding one million. These regions have more infrastructure to get fire crews and other resources to suppress fires quickly. Hence, the ISA durations are expected to be shorter. On the other hand, fires in the upper region are harder to get to and are not detected as readily, leading to longer ISA durations. Similar conclusions can be drawn from Table S3 in the Supplementary Material for the estimates of the PLRMs with natural cubic splines.

Under the LRM, the ISA duration is not significantly associated with the explanatory variables *wind speed* and *fuel type*. This may be caused by the correlation between these two independent variables; our simulation confirms this finding. On the other hand, the estimate β_{M2} is significantly different from zero in the PLRMs with spatial correlation. This indicates that the ISA duration is shorter for fires with *M2* fuel than it is for fires with other fuel types. We plot the estimated nonparametric function $h_0(\cdot)$ of *Wind speed* in Figure 3 for models PLRM-Indpt (2.4b) and PLRM-Corrltd (2.4d). For comparison, a reference line at $y = 0$ is overlaid, together with pointwise 95% confidence intervals (CIs) of $\hat{h}_0(\cdot)$, obtained from the 2.5% and 97.5% quantiles of the realizations of $\hat{h}_0$ using a bootstrap procedure (described below) with a resample size of 1000.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Bootstrap Procedure

1: *Steps 1 and 2.* Follow *Steps 1 and 2* of ALGORITHM B2 to generate $(\delta_{1i}^{(j)}, \delta_{2i}^{(j)})$, and $S_i^{(j)}$, for $i = 1, \cdots, n$ and $j = 1, \cdots, J^*$.
2: *Step 3.* Draw a sample of n observations with equal weight and with replacement from $\{(\mathbf{O}_i, S_i^{(j)}, \delta_{1i}^{(j)}, \delta_{2i}^{(j)}), i = 1, \cdots, n; j = 1, \cdots, J^*\}$. Denote this sample $\{(\mathbf{O}_i^{(a)}, S_i^{(j,a)}, \delta_{1i}^{(j,a)}, \delta_{2i}^{(j,a)}), i = 1, \cdots, n; j = 1, \cdots, J^*\}$.
3: *Step 4.* For $j = 1, 2, \cdots, J^*$, obtain the vector $\tilde{\mathbf{Y}}^{(j,a)}$ with the ith element being $\tilde{Y}_i^{(j,a)} = \log(L_i^* + \tilde{S}_i^{(j,a)})$. Follow *Step 3* of ALGORITHM B2 to compute the estimators $\hat{\boldsymbol{\beta}}_n^{\dagger}$ and $\hat{h}_{0n}(\cdot)$ with $\tilde{Y}_i^{(j,a)}$ and $(\mathbf{O}_i, S_i^{(j)}, \delta_{1i}^{(j)}, \delta_{2i}^{(j)})$, for $i = 1, \cdots, n$ and $j = 1, \cdots, J^*$. Denote the resulting estimators $\hat{\boldsymbol{\beta}}_{n,J^*}^{\dagger(a)}$ and $\hat{h}_{0n,J^*}^{(a)}(\cdot)$.
4: *Step 5.* Repeat *Steps 3 and 4* A times to obtain A sets of estimators.

The two nonparametric estimates by kernel smoothing and natural cubic splines under the two models are clearly different from zero. Thus we need to use an unspecified function to capture the true association of the ISA duration with the wind speed. When the wind speed is low, it is not significantly associated with the ISA duration; the association strengthens as the speed increases.

—— *Figure 3 is here.* ——

—— *Tables 2 and 3 are here.* ——

We present the distribution estimates for the models LRM-Indpt (2.4a) and LRM-Corrltd (2.4c) in Figure 2. For fixed values of *Wind Speed*, we evaluate the estimator $\hat{F}_n(t|\mathbf{x})$ in (3.2). The 95% pointwise CIs are also provided. The CIs are calculated using the estimated asymptotic variance given in (3.8) for the analysis in the independent case, and are obtained using the bootstrap variance in the correlated case. In each plot, we overlay estimated distribution curves with two conventional LRM approaches: they use the observed portion of the ISA duration and the interval-censored ISA duration. The distribution curves estimated by these approaches are similar for different subgroups of fires. However, the curves of the proposed estimator show that fires from the upper region are likely to have a longer ISA duration for fixed fuel type and wind speed. Moreover, *M2* fires tend to have a shorter ISA

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

duration. This shows that caution is required when dealing with durations with missing origins.

—— *Figure 2 is here.* ——

Following Xiong et al. (2019), we apply a spatio-temporal extension of Moran's I test to check the validity of models (2.4a), (2.4b), (2.4c), and (2.4d). For each model, we compute one set of the residuals using $e_i^{(j)} = \log(L_i^* + \tilde{S}_i^{(j)}) - \hat{Y}_i$ for $i = 1, \cdots, n$, where $\tilde{S}_i^{(j)}$ is the imputed reporting delay based on *Step 2* of ALGORITHMS B1 and B2, and $\hat{Y}_i$ is the fitted value obtained by each model. To construct the Moran's I test statistic, we define fire neighbors to be any two fires from the same management area, and with the time lag between their report times bounded by a predetermined τ. The plots of Figure 4 compare Moran's I, evaluated at different values of τ, with the LRM and PLRM. For independent units, Figure 4a shows that for the PLRM the Moran's I is covered by the 95% acceptance regions when $\tau \geq 25$. As observed by Xiong et al. (2019), this suggests that the PLRM captures the nonlinear spatial pattern, and appears to be more appropriate than the LRM. For correlated units, Figure 4b shows that for both models, the Moran's I is within the acceptance region. This indicates that spatial correlation is the main cause of the discrepancy between the value of Moran's I and the expected value under the null hypothesis.

—— *Figure 4 is here.* ——

Residual plots also support the Moran's I results. The residuals from the PLRM are near zero for both independent and correlated units; see the scatterplots of the residuals versus *wind speed* and *report time* in Figures S3 and S4, respectively, in the Supplementary Material. These findings suggest that models PLRM-Indpt (2.4b) and PLRM-Corrltd(2.4d) adequately capture the associations between the ISA duration and the risk factors.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

5. Simulation Studies

We carried out simulation studies to examine the finite-sample performance of our approach and to verify the findings of the real-data analysis. In *Simulation A*, we generated longitudinal measurements from the Wiener process model to verify the consistency and efficiency of our approach. In *Simulation B*, the covariates are related by a latent variable, and *Simulation C* assesses the impact of missing covariates.

5.1 Simulation A: Study to Verify Consistency and Efficiency

To mimic the fire data, we simulated a data set with $n = 500$ fires, as follows:

Settings for Simulation A

1: Let $\mathbf{X}_i = (X_{1i}, X_{2i})$ and $X_{1i} \sim Unif(0,1)$, $X_{2i} \sim B(1, 0.6)$, for $i = 1, \cdots, n$.
2: Generate independently n locations $\boldsymbol{\omega}_i$ with each of the two location indices ω_{1i} and ω_{2i} from $Unif(0,1)$.
3: Generate the burnt area process $A_i(t), t \in [0, 2500]$, for $i = 1, \cdots, n$, based on the model $A_i(t) = \nu_i t + \sigma W_i(t)$, where $\nu_i = \nu \exp\{X_{1i}\gamma_1 + X_{2i}(\gamma_2 + \delta_{2i}) + \delta_{1i}\}$. Let $\boldsymbol{\delta}_1 = (\delta_{11}, \cdots, \delta_{1n})'$ and $\boldsymbol{\delta}_2 = (\delta_{21}, \cdots, \delta_{2n})'$. Generate $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ from $MVN(0, \Sigma_1)$ and $MVN(0, \Sigma_2)$, respectively. To simulate different types of study units, we consider two types of covariance matrices Σ_1 and Σ_2:

   Type 1 (independent): $\Sigma_1 = \psi_1^2\mathbf{I}$ and $\Sigma_2 = \psi_2^2\mathbf{I}$ with $\mathbf{I}$ the $n \times n$ identity matrix.

   Type 2 (spatially correlated): The (i,j)th elements of Σ_1 and Σ_2 are given by $(\Sigma_1)_{ij} = \psi_1^2 \exp\{-||\boldsymbol{\omega_i} - \boldsymbol{\omega_j}||/\rho_1\}$ and $(\Sigma_2)_{ij} = \psi_2^2 \exp\{-||\boldsymbol{\omega_i} - \boldsymbol{\omega_j}||/\rho_2\}$, respectively.
4: Let $\mu_B(\mathbf{X}_i) = 0.8 + 2.5X_{1i} - X_{2i}$. Generate the size at the report time $B_i \sim \text{logNormal}(\mu_B(\mathbf{X}_i), 0.6^2)$.
5: Determine the reporting delay as $S_i = \max\{t | t \in [0, 2500], A_i(t) \le B_i\}$.
6: Generate $L_i^* \sim \text{Exp}(\lambda_i)$, where $\lambda_i = \exp\big(0.1\delta_{1i} + X_{1i} - (0.2 + \delta_{2i})X_{2i} - 0.6\big)$.
7: Obtain $L_i = S_i + L_i^*$, $A_i(L_i)$, and $D_i = A_i(L_i) - B_i$.

When $\psi_2 = 0$, only the random intercept term δ_{1i} is included in ν_i. We consider the following four scenarios:

(A1.1) $\psi_1 = 0.25, \psi_2 = 0$, and a Type 1 covariance matrix.

(A1.2) $\psi_1 = 0.25, \psi_2 = 0.15$, and a Type 1 covariance matrix.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

(A2.1) $\psi_1 = 0.25, \psi_2 = 0, \rho_1 = 0.3, \rho_2 = 0$, and a Type 2 covariance matrix.

(A2.2) $\psi_1 = 0.25, \psi_2 = 0.15, \rho_1 = 0.3, \rho_2 = 0.1$, and a Type 2 covariance matrix.

We set $\nu = 3.5, \sigma = 0.8, \gamma_1 = 0.2$, and $\gamma_2 = 0.15$ in all the scenarios. For each generated data set, we consider the models LRM-Indpt (2.4a) and LRM-Corrltd (2.4c), and evaluate the estimators for independent and correlated study units. We calculate the estimated standard errors (SEs) of the estimates of $\boldsymbol{\theta}$ using the robust sandwich variance estimator, and estimate the SE of the estimates of $\boldsymbol{\beta}$ using (3.7). We also evaluate the least-squares estimator of $\boldsymbol{\beta}$ using L and the observed L^*, as well as the interval-censored observations with $L_i \in [L_i^*, L_i^* + R_{max}]$, where R_{max} is the third quantile of S for each generated data set.

In each scenario, we repeated the simulation 400 times. The sample means of the estimates under a correctly specified ν_i are close to the true parameter values. For example, the true specification of ν_i in Scenario (A1.2) is $\nu_i = \nu \exp\{\delta_{1i} + X_{1i}\gamma_1 + X_{2i}\gamma_2\}$. The sample means of the estimates from our approaches for both independent and correlated study units are close to the true parameter values. This provides empirical evidence for the consistency of our estimator. On the other hand, the estimate of $\boldsymbol{\beta}$ from the naïve approach using L_i^* has a large bias, and is in the opposite direction to that based on the true ISA duration and that from our approach. Thus, approaches that uses only the observed or interval-censored ISA duration could be misleading. See Tables S4–S7 in the Supplementary Material for the parameter estimates based on the 400 simulation repetitions in the four scenarios.

We also examine the performance of the SE estimators. The sample means of the estimated SEs under the true settings agree well with the empirical SEs, that is, the sample SEs. For correlated data, both the sample SEs and the sample means

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

of the estimated SEs under the independence assumption are larger than the values that account for the correlation. This suggests that failing to accommodate for correlation may lead to inefficient estimators.

In Scenario (A2.2), we estimated the distribution curves by our approach with correlation, and obtained the estimates by the empirical distribution function of the residuals from the regression model with the true (complete) ISA duration, the observed ISA duration, and interval-censored ISA duration. The sample means of the conditional distribution estimates using our approach are close to the empirical distribution function estimates using the complete ISA duration. We observe that fires with $X_2 = 0$ are likely to have a longer ISA duration. In sharp contrast, the estimates from the two conventional approaches show the opposite behavior. This suggests that a reporting delay affects the association of the ISA duration with the covariates, and should be accounted for. We also evaluated the distribution estimator by our approach under the independence assumption in the scenarios (A2.2) with correlated units. The estimated conditional distribution curve by our approach shows similar behavior, but with a wider 95% confidence interval than the one with correlation. This indicates that considering the correlation yields a more efficient distribution estimator. The estimated distribution curves are presented in Figures S5 and S6 of the Supplementary Material.

5.2 Simulation B: Study to Examine the Impact of Correlated Covariates

In this study, we generated covariates X_1 and X_2 that depend on a latent variable M. We first simulated $M_i \sim B(1, 0.7)$, for $i = 1, \cdots, n$, and then generated $X_{i1} \sim Unif(0, a_1 + M_i b_1)$ and $X_{i2} \sim B(1, \frac{\exp(a_2+M_i b_2)}{1+\exp(a_2+M_i b_2)})$, for $i = 1, 2, \cdots, n$. We used

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

$(a_1, b_1, a_2, b_2) = (0.9, 0.1, 4, -2)$ to simulate X_1 and X_2, with a Pearson correlation coefficient of -0.65. Here, ν_i's includes only the random intercept term, and the covariance matrix is Type 2. All the other variables are as in *Simulation A*. We evaluated the parameter estimators using the simulated data.

The sample means of the estimates for $\boldsymbol{\theta}$ under the correct specification of ν_i based on 400 repetitions with the model including both X_1 and X_2 as covariates are close to the true values. In addition, the estimates of γ_1 and γ_2 suggest that X_1 and X_2 are both significantly associated with the longitudinal process. However, $\boldsymbol{\beta}$ appears to be underestimated. Only X_2 is identified as a significant predictor for the ISA duration. This finding corroborates the observation from the real-data analysis that the ISA duration is not significantly associated with wind speed and fuel type, even though these factors are associated with the burnt area process, and were therefore expected to be important for the ISA duration. The simulation outcomes are presented in Table S8 of the Supplementary Material. We also examined the parameter estimates for the model including only X_1 or X_2, which are shown in Tables S9 and S10 of the Supplementary Material. These results show that X_1 and X_2 are significant predictors when only one of them is included in the model. This suggests that the aforementioned under-estimation of $\boldsymbol{\beta}$ may be caused by the strong correlation of X_1 and X_2. Further investigation of the variable selection or a stratified data analysis would be desirable.

5.3 Simulation C: Study to Examine the Impact of Missing Covariates

This simulation examines the impact of missing covariates under two scenarios: (C.1) a missing covariate in the longitudinal process and the regression model, and (C.2) a missing covariate in the spatial correlation function.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

For Scenario (C.1), we generated the longitudinal measures from model (2.2), with $\nu_i = \nu \exp\{X_{i1}\gamma_1 + X_{i2}\gamma_2 + X_{i3}\gamma_3\}$ and $X_{i3} \sim B(1, 0.2)$. We chose $\gamma_3 = -0.4$, and generated other variables in the same way as in *Simulation A*. We analyzed the data by assuming study units are spatially correlated with the full set of covariates, that is, X_1, X_2, and X_3, and the partial set of covariates, that is, X_1 and X_2. Table S11 in the Supplementary Material presents a summary of the estimates. When only X_1 and X_2 are included in the model, the estimates for γ_1 and γ_2 are still close to the true values. We can also identify X_1 and X_2 as significant predictors for the duration from the estimates for β_1 and β_2. These results suggest that the missing covariate does not change the effects of the other covariates on the longitudinal process and the duration.

In Scenario (C.2), we generated spatially correlated units with a covariate-dependent correlation function. We first simulated $Z_i \sim B(1, 0.8)$ and considered the longitudinal model with $\nu_i = \nu \exp\{X_{i1}\gamma_1 + X_{i2}\gamma_2\}$, that is, $\psi_2 = 0, \rho_2 = 0$. Here, $\boldsymbol{\delta}_1 = (\delta_{11}, \cdots, \delta_{1n})$ were generated from $MVN(0, \Sigma_1^*)$, with the (i, j)th element in Σ_1^* being $\psi_1^2 exp\{-|\boldsymbol{\omega_i} - \boldsymbol{\omega_j}|/(\rho(Z_i)\rho(Z_j))$ and $logit[\rho(Z_i)] = \log(49) - 3Z_i$. We chose $\psi_1 = 0.25$, and analyzed the correlated data with the spatial correlation structure given in (2.3) and with the true spatial correlation function Σ_1^*. To make a comparison, we also approximate the value of ρ_1 in the misspecified correlation function given in (2.3) by $\rho_1 = \mathrm{E}[\rho(Z_i)\rho(Z_j)]$. Table S12 in the Supplementary Material summarizes the estimates with the simulated data. With the true spatial correlation function, we assume the coefficients associated with $\rho(\cdot)$ are known, and only estimate ψ_1. Although the misspecified correlation structure leads to biased estimates for the parameters ψ_1 and ρ_1, we observe that the sample means of the estimates for ν and $\boldsymbol{\gamma}$ are close to the true values. The estimates for β with this

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

misspecified structure are also similar to those with the true correlation function, and close to the estimates with the complete ISA duration. These results suggest that the reported analysis of the real data can still be meaningful and interpretable, even if an important covariate is missing from the correlation structure.

6. Conclusion

We have considered a regression analysis of event durations with missing origins under the semiparametric AFT model. This is an extension of the work by Xiong, Braun, and Hu (2021) to the regression setting. Motivated by records of wildfires, we extended the procedure to account for spatial correlation. We used our approach to estimate the distribution of the ISA duration, given a set of covariates for wildfires caused by lightning, and validated our results using a simulation study. Both our real-data analysis and our simulation studies indicate that an inference based on conventional approaches could give misleading results. This confirms the importance of dealing with the missing time origins. The proof of the asymptotic properties of the estimators with spatially correlated units remains an open problem, but our simulation results appear to support the asymptotic validity of our approach.

Our approach not only evaluates the association of the ISA duration with the risk factors, but also provides a distribution estimator. Given the covariates at the report time, our distribution estimator can be used to predict the full ISA duration. In the presence of spatial correlation, it can straightforwardly be extended to estimate the joint density, which will improve the deployment of wildfire suppression resources. In addition, although we assume the two random effects are independent of each other, our approach can be readily adapted by specifying the correlation between the two random effects.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Several future investigations are worthwhile. First, our approach finds no association between the ISA duration and the covariates *wind speed* and *fuel type*, which are believed to play important roles in the development of a fire. This unexpected finding may be due to their strong association with the covariate *region*. A stratified data analysis may help to reveal the true effects of *wind speed* and *fuel type*. Although the application presented here does not involve high-dimensional data, it would be of both scientific and statistical interest to account for high-dimensional variable selection when applying this approach to other situations. Second, the environmental factors are often time-varying. Thus, we could extend the model to dynamically predict the ISA duration. For example, we may consider linear transformation models, and adapt our approach to approximate the start time and the inherent missing segment of the time-varying covariate. Further, the estimation based on spatially correlated data can be computationally intensive; feasible methods based on the composite likelihood (e.g. Paik and Ying, 2012) may reduce this burden. It would also be worth making the current spatial covariance functions depend on directional variables. Following Neto et al. (2014), we can extend the proposed procedure to incorporate directional variables in the spatial covariance function. Finally, our approach is applicable to many practical situations with missing time origins. Examples include predicting the time from HIV infection to an AIDS diagnosis (Doksum and Normand, 1995), and studying the gap times of the labor process for pregnant women with an unknown start time of the dilation process (Ma and Sundaram, 2018). A recent example is COVID-19 with asymptomatic infection and transmission. One can estimate the distribution of the incubation period with our approach by using the longitudinal viral load measures.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Supplementary Material

The derivation of the asymptotic properties in Section 3.1.3, and additional tables and figures for the real-data analysis in Section 4 and the simulation studies in Section 5, together with sample code are provided in the online Supplementary Material.

Acknowledgments

We are grateful to N. McLoughlin of Alberta Wildfire Management Branch for providing the wildland fire data. We thank G.A. Whitmore for his many constructive suggestions, and the AE and two referees for their careful reviews of a previous version of the manuscript. The research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and a CRT grant from the Canadian Statistical Sciences Institute (CANSSI).

References

Banerjee, S., M. M. Wall, and B. P. Carlin (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics 4*(1), 123–142.

Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika 66*, 429–436.

Chhikara, R. S. and J. L. Folks (1989). *The Inverse Gaussian Distribution: Theory, Methodology and Applications*, Volume 162. New York, USA: Marcel Dekker, Inc.

Cumming, S. (2001). Forest type and wildfire in the alberta boreal mixedwood: what do fires burn? *Ecological applications 11*(1), 97–110.

Doksum, K. A. and S.-L. T. Normand (1995). Gaussian models for degradation processes-part i: Methods for the analysis of biomarker data. *Lifetime Data Analysis 1*(2), 131–144.

Forestry Canada Fire Danger Group (1992). *Development and structure of the Canadian forest fire behavior prediction system*, Volume 3. Forestry Canada.

Furgal, A. K., A. Sen, and J. M. Taylor (2019). Review and comparison of computational approaches for joint longitudinal and time-to-event models. *International Statistical Review 87*(2), 393–418.

Goetghebeur, E. and L. Ryan (2000). Semiparametric regression analysis of interval-censored data. *Biometrics 56*(4), 1139–1144.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Komárek, A. and M. A. Komárek (2015). Package 'smoothsurv'.

Lee, M.-L. T. and G. A. Whitmore (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science 21*(4), 501–513.

Li, Y. and L. Ryan (2002). Modeling spatial survival data using semiparametric frailty models. *Biometrics 58*(2), 287–297.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *The Annals of Statistics 27*(2), 415–438.

Ma, L. and R. Sundaram (2018). Analysis of gap times based on panel count data with informative observation times and unknown start time. *Journal of the American Statistical Association 113*(521), 294–305.

Martell, D. L. (2007). Forest fire management. In *Handbook of Operations Research in Natural Resources*, pp. 489–509. Springer.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Morin, A. A. (2014). *A spatial analysis of forest fire survival and a marked cluster process for simulating fire load.* The University of Western Ontario, London, Ontario, Canada: MSc thesis.

Motarjem, K., M. Mohammadzadeh, and A. Abyar (2017). Geostatistical survival model with gaussian random effect. *Statistical Papers*, 1–23.

Neto, J. H. V., A. M. Schmidt, and P. Guttorp (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 103–122.

Ning, J., J. Qin, and Y. Shen (2011). Buckley-james-type estimator with right-censored and length-biased data. *Biometrics 67*(4), 1369–1378.

Paik, J. and Z. Ying (2012). A composite likelihood approach for spatially correlated survival data. *Computational statistics & data analysis 56*(1), 209–216.

Papageorgiou, G., K. Mauff, A. Tomer, and D. Rizopoulos (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application*.

Qin, J., C. You, Q. Lin, T. Hu, S. Yu, and X.-H. Zhou (2020). Estimation of incubation period distribution of covid-19 using disease onset forward time: a novel cross-sectional and forward follow-up study. *Science advances 6*(33), eabc1202.

Tremblay, P.-O., T. Duchesne, and S. G. Cumming (2018). Survival analysis and classification methods for forest fire size. *PloS one 13*(1).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

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

Wu, L., W. Liu, and X. J. Hu (2010). Joint inference on hiv viral dynamics and immune suppression in presence of measurement errors. *Biometrics 66*(2), 327–335.

Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 330–339.

Xiong, Y., D. Bingham, W. J. Braun, and X. J. Hu (2019). Moran's i statistic-based nonparametric test with spatio-temporal observations. *Journal of Nonparametric Statistics 31*(1), 244–267.

Xiong, Y., W. J. Braun, and X. J. Hu (2021). Estimating duration distribution aided by auxiliary longitudinal measures in presence of missing time origin. *Lifetime Data Analysis*. published online.

Department of Statistics and Actuarial Science, Simon Fraser University
Biostatistics Program, Fred Hutchinson Cancer Center
E-mail: (yi_xiong@sfu.ca)
Department of Computer Science, Mathematics, Physics and Statistics, University of British Columbia-Okanagan
E-mail: (john.braun@ubc.ca)
Département de mathématiques et de statistique, Université Laval
E-mail: (thierry.duchesne@mat.ulaval.ca)
Department of Statistics and Actuarial Science, Simon Fraser University
E-mail: (joanh@stat.sfu.ca)

Table 1: Estimates of the parameters in the longitudinal model (2.2)

		ν	σ	ψ_1^2	ρ_1	ψ_2^2	ρ_2	$\gamma^{\S}_{Region:upper}$	$\gamma_{WindSpeed}$	$\gamma^{\S}_{Fuel:C1}$	$\gamma^{\S}_{Fuel:C2}$	$\gamma^{\S}_{Fuel:M2}$
	Independent Units											
Random	Est.	0.036	1.272	0.020				0.225	**2.878**	0.528	0.252	-0.370
Intercept	SE	0.024	0.026	0.011				0.173	0.521	0.399	0.309	0.434
	AIC	2887.758										
Mixed	Est.	0.044	1.270	**0.008**		**0.011**		**0.378**	**3.231**	0.300	0.417	-0.025
Effect	SE	0.022	0.026	0.003		0.005		0.148	0.507	0.435	0.331	0.462
	AIC	1343.669										
	Correlated Units											
Random	Est.	0.041	1.272	**0.015**	**0.017**			**0.294**	**2.830**	0.536	0.065	0.306
Intercept	SE	0.025	0.026	0.007	0.008			0.141	0.422	0.274	0.298	0.424
	AIC	1353.652										
Mixed	Est.	0.043	1.272	**0.019**	**2.927**	**0.010**	**0.011**	**0.786**	**3.844**	-0.006	0.138	**-0.751**
Effect	SE	0.024	0.026	0.006	0.038	0.005	0.006	0.133	0.479	0.426	0.316	0.346
	AIC	1329.088										

$\gamma^{\S}_{Region:upper}$: coefficient for categorical variable *Region* and baseline category is *lower region*.
$\gamma^{\S}_{Fuel:C1}, \gamma^{\S}_{Fuel:C2}, \gamma^{\S}_{Fuel:M2}$: coefficients for categorical variable *Fuel type* and baseline category is *other fuel types*.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Table 2: Estimates of the parameters in the models LRM-Indpt (2.4a) and LRM-Crrltd (2.4c)

			β_0	$\beta^{\S}_{Region:upper}$	$\beta_{WindSpeed}$	$\beta^{\S}_{Fuel:C1}$	$\beta^{\S}_{Fuel:C2}$	$\beta^{\S}_{Fuel:M2}$	β_{δ_1}	β_{δ_2}
		Independent Units								
Observed		Est.	0.014	-0.097	-0.405	-0.499	-0.341	-0.186		
duration		SE	0.281	0.147	0.521	0.372	0.267	0.327		
	$R_{max} = 12$	Est.	2.013	0.092	-0.437	-0.030	-0.001	0.012		
Interval-	$R_{max} = 12$	SE	0.199	0.109	0.447	0.261	0.181	0.235		
-censored	$R_{max} = 48$	Est.	3.446	0.041	0.211	0.026	-0.023	-0.260		
	$R_{max} = 48$	SE	0.101	0.063	0.187	0.131	0.092	0.243		
Proposed	Random	Est.	0.981	**0.341**	-0.230	-0.231	-0.097	-0.446		
	intercept	SE	0.323	0.175	0.600	0.425	0.308	0.377		
	Mixed	Est.	0.967	**0.292**	-0.287	-0.183	-0.046	-0.406		
	Effect	SE	0.325	0.124	0.612	0.434	0.306	0.379		
		Correlated Units								
Observed		Est.	0.012	-0.112	-0.344	-0.520	-0.349	-0.166	-0.978	-0.514
duration		SE	0.281	0.147	0.523	0.372	0.267	0.327	0.734	0.709
	$R_{max} = 12$	Est.	2.026	0.087	-0.465	-0.037	0.001	0.007	-0.214	0.066
Interval-	$R_{max} = 12$	SE	0.201	0.105	0.433	0.254	0.184	0.237	0.574	0.532
-censored	$R_{max} = 48$	Est.	3.445	0.043	0.208	0.030	-0.022	-0.261	0.033	-0.168
	$R_{max} = 48$	SE	0.102	0.060	0.191	0.125	0.088	0.238	0.330	0.305
Proposed	Random	Est.	0.987	**0.382**	-0.294	-0.238	-0.076	-0.446	-0.486	
	intercept	SE	0.322	0.172	0.602	0.446	0.332	0.383	0.575	
	Mixed	Est.	0.880	**0.387**	-0.248	-0.087	0.053	-0.256	**-0.862**	-0.603
	Effect	SE	0.319	0.179	0.600	0.417	0.317	0.376	0.339	0.371

$\beta^{\S}_{Region:upper}$: regression coefficient for categorical variable *Region* and baseline category is *lower region*.
$\beta^{\S}_{Fuel:C1}, \beta^{\S}_{Fuel:C2}, \beta^{\S}_{Fuel:M2}$: regression coefficients for categorical variable *Fuel type* and baseline category is *other fuel types*.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118

Table 3: Estimates of the parameters in the models PLRM-Indpt (2.4b) and PLRM-Crrltd (2.4d), with h_0 estimated using kernel smoothing

		β_0	$\beta^{\dagger}_{Region:Northeast}$	$\beta^{\dagger}_{Fuel:C1}$	$\beta^{\dagger}_{Fuel:C2}$	$\beta^{\dagger}_{Fuel:M2}$	$\beta^{\dagger}_{\delta_1}$	$\beta^{\dagger}_{\delta_2}$
	Model (2.4b), *Independent Units*							
Random	Est.	0.929	**0.382**	-0.231	-0.097	-0.435		
Intercept	SE	0.325	0.193	0.471	0.341	0.416		
Mixed	Est.	0.902	**0.393**	-0.182	-0.046	-0.393		
Effect	SE	0.327	0.192	0.479	0.339	0.418		
	Model (2.4d), *Correlated Units*							
Random	Est.	0.920	**0.383**	-0.238	-0.076	-0.432	-0.477	
Intercept	SE	0.328	0.188	0.468	0.352	0.402	0.534	
Mixed	Est.	0.924	**0.386**	-0.172	-0.071	-0.409	**-0.848**	-0.540
Effect	SE	0.329	0.189	0.475	0.335	0.382	0.433	0.557

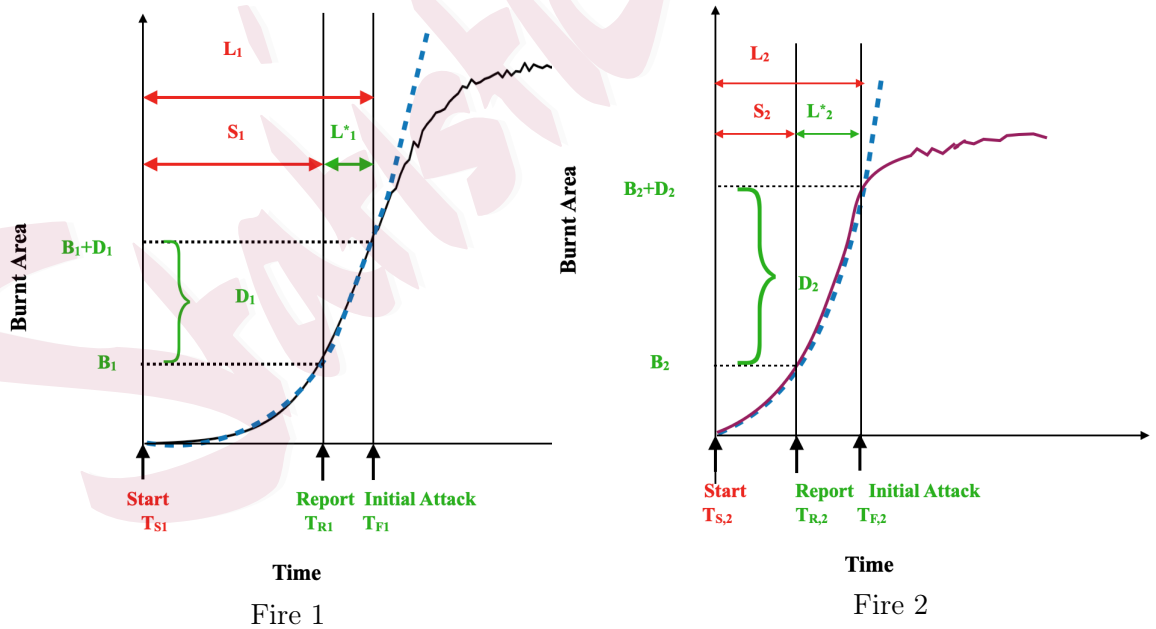
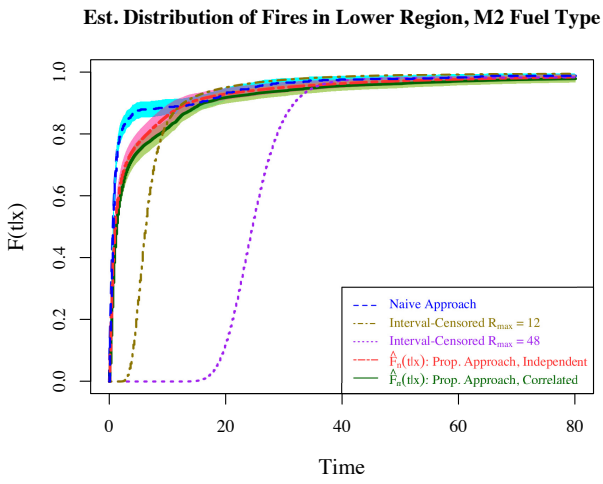
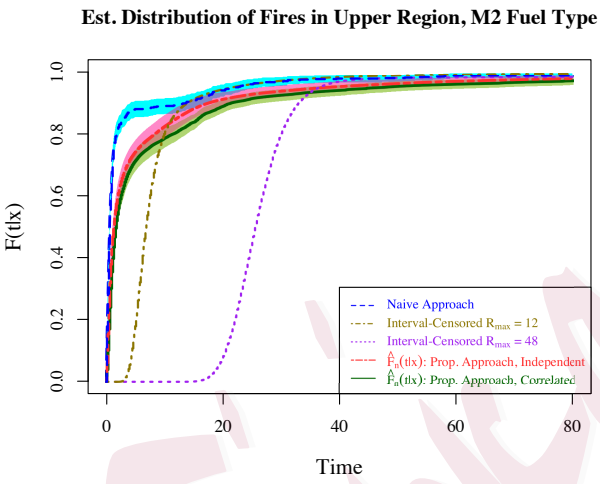


Figure 1: Hypothetical description of the progression through fire management phases for two fires with different covariate values

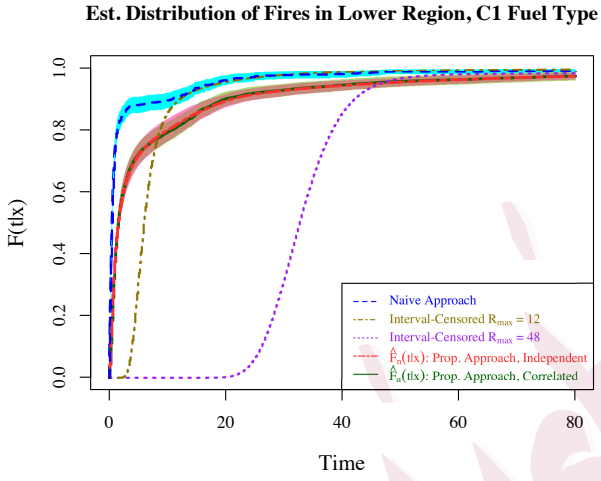
Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118



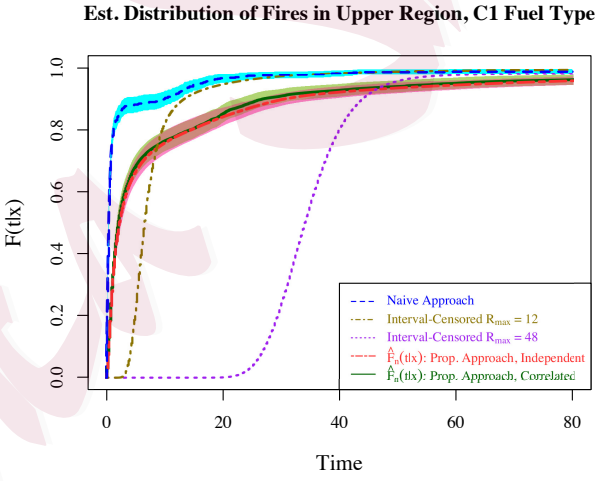
(a.1) Lower region, M2 type



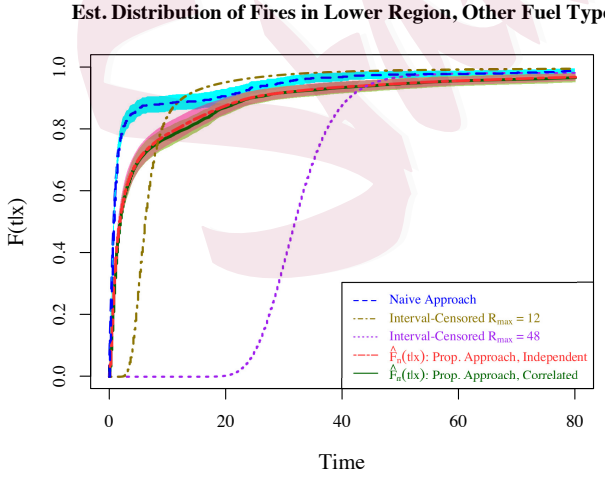
(a.2) Upper region, M2 type



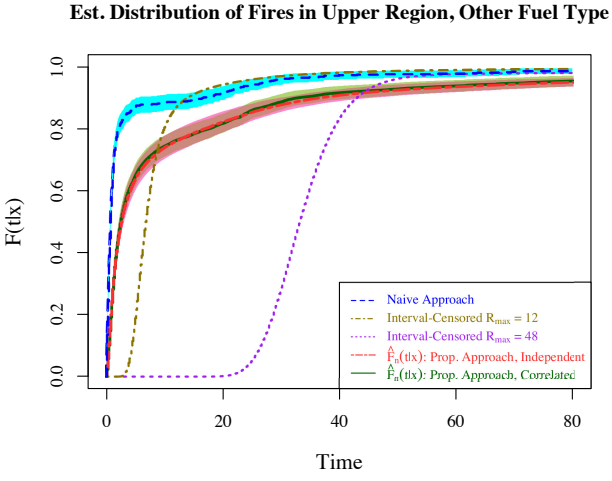
(b.1) Lower region, C1 type



(b.2) Upper region, C1 type



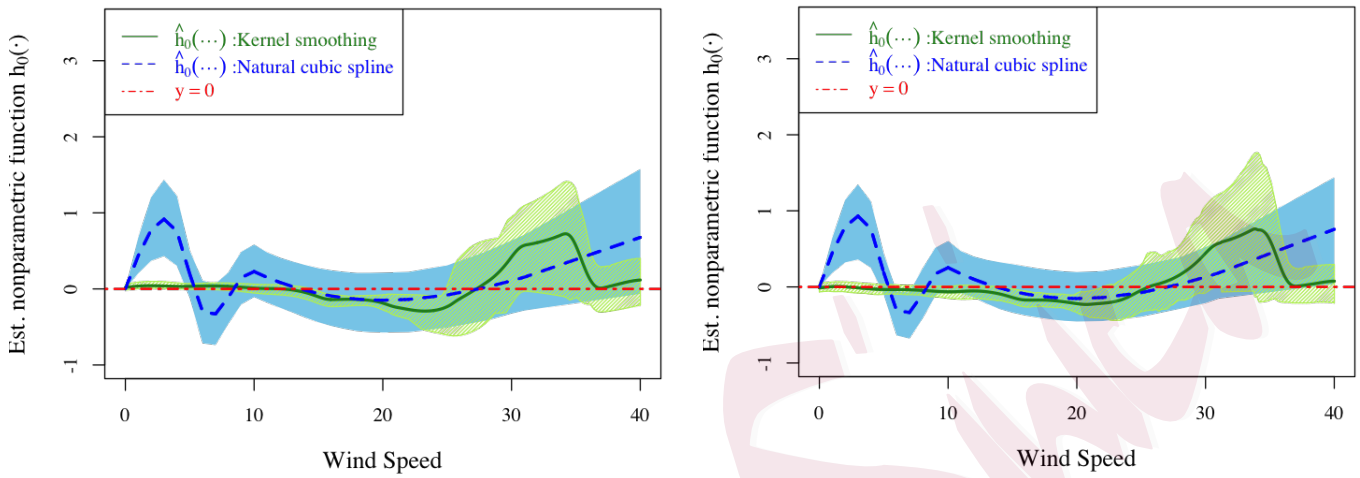
(c.1) Lower region, Other type



(c.2) Upper region, Other type

Figure 2: Estimated $F(t|\mathbf{x})$ with wind speed $= 7.75\,\mathrm{km/h}$.

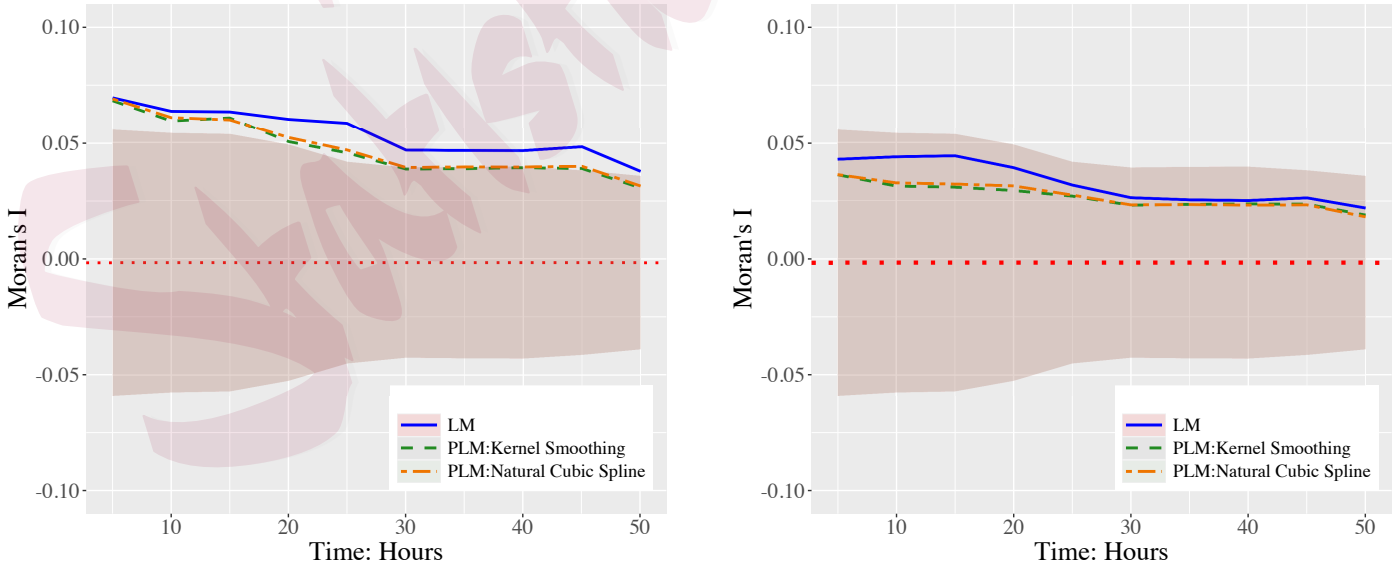
Statistica Sinica: Preprint
doi:10.5705/ss.202021.0118



(a) Model PLRM-Indpt (2.4b)

(b) Model PLRM-Crrltd (2.4d)

Figure 3: Plots of estimated nonparametric function $h_0(\cdot)$ of *wind speed*. The shaded area represents the 95% pointwise confidence intervals, which are approximated by the 2.5% and 97.5% quantiles of the realizations of $\hat{h}_0(\cdot)$ using a bootstrap with resample size 1000.



(a) Models LRM-Indpt (2.4a) and PLRM-Indpt (2.4b), with independent units

(b) Models LRM-Crrltd (2.4c) and PLRM-Crrltd (2.4d), with correlated units

Figure 4: Model diagnosis by Moran's I using residuals from different regression models.