Fall 2021

Multiple Frailty Model for Spatially Correlated Interval-Censored

Wanfang Zhang

MULTIPLE FRAILTY MODEL FOR SPATIALLY CORRELATED INTERVAL-CENSORED DATA

By

Wanfang Zhang

Bachelor of Science
Shandong Agricultural University (SAU), 2017

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in

Biostatistics

Arnold School of Public Health

University of South Carolina

2021

Accepted by:

Bo Cai, Director of Thesis

Feifei Xiao, Reader

Alexander C McLain, Reader

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

ABSTRACT

In this paper, we consider the problem of multiple frailty selection for general interval-censored spatial survival data, which often occurs in clinical trials and epidemiological studies. The general interval-censored data is a mixture of left-, right- and interval-censored data. We propose a Bayesian semiparametric approach based on the Cox proportional hazard model, where monotone splines were used for non-parametrical modeling of the cumulative baseline hazards where the variable selection priors were used for frailty selection. A two-stage data augmentation with Poisson latent variables is developed for efficient computation. The approach is evaluated based a simulation study and illustrated using a set of geographically referenced smoking cessation data in Minnesota. The whole procedure is implemented in software R 4.0.4 and WinBUGS.

KEYWORDS

Semiparametric regression, interval-censored data, spatial clustering, variable selection, multiple frailty, I-spline.

TABLE OF CONTENTS

ABSTRACT iii

Keywords iii

Chapter 1. Introduction 1

Chapter 2. The model 5

Chapter 3. Simulation STUDIES 21

Chapter 4. Real data analysis 27

Chapter 5. Discussion 32

References 33

LIST OF TABLES

Table 3.1 Posterior probability of frailty selection and corresponding Bayes factor in favor of homogeneity. 22

Table 3.2 Posterior probability of the eight possible models in terms of frailty selection. 23

Table 3.3 Estimation of regression coefficients βh and spatial parameter $\tau\phi$. 23

Table 4.1 Estimation of proposed model and Weibull model for Minnesota smoke cessation data. 27

LIST OF FIGURES

Figure 3.1 Plot of estimated baseline survival curve based on 100 simulated datasets from true value, proposed model (95% pointwise credible intervals) and Weibull model. 25

Figure 3.2 Maps of posterior means of the spatial parameter ϕi over 46 counties of SC based on proposed model and Weibull model. 26

Figure 4.1 Maps of posterior means of the spatial parameter ϕi over 51 zip code areas of Minnesota based on proposed model and Weibull model. 28

Figure 4.2 Estimated survival curves for the smoking cessation study, using Turnbull method, proposed model, and Weibull model. Event of interest is time to relapse to smoke. 30

LIST OF SYMBOLS

$h_0(t)$	Baseline hazard function
$\Lambda_0(t)$	Cumulative baseline hazard function
x	Predictor
β	Co-efficient
ϕ_i	Spatial information corresponding to ith area
ξ_j	Frailty information corresponding to jth binary variable
$\delta_{1,2,3}$	Censoring indicators for left-, interval- and right- censoring
k_h^{-1}	Precision of jth frailty
π_{0h}	Prior probability of the null hypothesis of homogeneity for the hth frailty
ρ_h	Homogeneity indicator for the hth frailty
γ_l	Set of nonnegative coefficients for I-spline corresponding to lth subset
b_l	A set of basis I-splines corresponding to lth subset
ω	Binary covariates
t	Time
τ_ϕ	Precision hyper parameter of spatial parameter ϕ

CHAPTER 1. INTRODUCTION

In medical studies, failure time may occur when patients have periodical visits. Data are collected at each visit, so the failure time is unobserved and often only known to lie in an interval $(L, R]$. Such data are referred to as interval-censored data. In this article, left and right censoring are special cases of interval censoring, with the beginning of the left interval at zero and the end of the right interval at infinity.[1] This is referred to as general interval-censored data.[2]

The most widely used regression model for time-to-event data is the proportional hazard (PH) model[3] in survival literature. A number of Bayesian semiparametric approaches have been proposed for regression analysis of interval-censored data under the PH model. Yavuz and Lambert[4] have proposed the use of penalized B-splines when modeling the baseline density function; Hanson, Johnson[5], Komarek and Lesaffre[6] studied the accelerated failure time model; Lin and Wang[7] use semiparametric probative model; Sinha et al.[8] observed the hazard to be a piecewise constant function; Lin et al.[9] adopted monotone splines to the baseline cumulative hazard function; Wang et al.[10] proposed a dynamic regression model with time varying coefficients. A comprehensive review of the PH model for interval-censored data before 2011 is given by Zhang and Sun[11]. Most recently, Li et al.[12] proposed a marginal Bayesian semiparametric model of mismeasured multivariate interval-censored data; Gao et al.[13] formulated the joint distribution of multiple right- and interval-censored events with proportional hazard models with random effects; Pak et al.[14] proposed a

semiparametric multi-state frailty model to analyze clustered event-history data which is subjected to interval censoring; Mao et al.[15] proposed semiparametric regression models for competing risk data with potentially time-varying (external) covariates; a great summary of difference between a parametric model and a semiparametric model for the same interval-censored data is given by Pak[16].

Although some R packages and SAS macros have been developed for analysis on interval-censored data, only a handful allow the inclusion of both spatial information and multiple frailty. Zhou et al.[17] developed an efficient method in R packag 'spBayesSurv' for semiparametric regression analysis for interval-censored data combined with spatial information in R; Pan et al.[18] proposed an R package, 'ICBayes', to perform semiparametric regression for interval-censored data combined with the frailty model. Other than the packages above, there are three recently developed packages for regression analysis in R, 'ICsurv'[19], 'icenReg'[20] and 'SmoothHazard'[21]. The 'ICsurv' package provides semiparametric models that use splines for the baseline distribution.[22] The 'icenReg' contains functions for imputation of the censored response variables and diagnostics of both regression effects and baseline distribution. The package 'SmoothHazard' implements algorithms for simultaneously fitting regression models to the three transition intensities of an illness-death model where the transition times to the intermediate state may be interval-censored and all the event times can be right-censored.

The term frailty itself is introduced by Vaupel et al.[23]. A frailty model is a model with random effect for time-to-event data, where the random effect (the frailty) has a multiplicative effect on the baseline hazard function.[24] A comprehensive summary of

the frailty model is given by Hazarika and Mahanta[25]. The clusters may be not only formed based on frailties such as age, sex and the specific clinical center visited, but also on geographical areas. This study focuses on areal data, which is a summary of data from a specific region (Banerjee et al.[26]) as lattice data. Recently, more and more methods have been developed to combine spatially referenced data with the frailty model. Zhou et al.[27] generalized an accelerated failure time in the spatial frailty model for arbitrarily censored data by using the R package 'spBayesSurv'. Hesam et al.[28] proposed a cause-specific hazard spatial frailty model with multivariate conditional autoregressive distribution for frailties. In addition to the PH model, Yiqi et al.[29] adopted a cure rate-proportional odds model with spatial frailties for interval-censored data. Since some variables like age group and sex can be considered as clusters, the frailties can then be added to the model to consider the correlation between individuals in clusters. However, methods rarely take into account the spatial correlation of various areas. They also allow for the variation of predictor effects across clusters under semiparametric settings. This study seeks to extend the model proposed by Pan et al.[30] by incorporating the spatial information into the semiparametric regression with multiple frailty. In addition, the proposed approach allows us to identify the inclusion and exclusion of frailties corresponding to the baseline hazard and the predictors.

The rest of the article is organized as follows: Section 2 formulates our proposed approach, including the modeling of the cumulative hazard function with monotonic splines and potential frailties of binary predictors, a reparameterization for frailties, two-stage Poisson data augmentation, the CAR prior for lattice spatial frailties, and prior specification. Section 3 provides the posterior computation details and model comparison

criteria. Section 4 shows the simulation results and the comparison with several existing approaches, models compared deviance information criterion (DIC). Section 5 presents real spatial smoking cessation data applications. Section 6 concludes with a discussion.

CHAPTER 2. THE MODEL

2.1 DATA AND THE LIKELIHOOD

The most widely used regression model for time-to-event data is the Cox PH model. Let T denote the failure time of interest, the Cox PH model is

$$h(t|x) = h_0(t)exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{1}$$

where $h_0(t)$ denotes an unspecified baseline hazard function and $\boldsymbol{x}$ indicates a vector of covariates. For spatial data, suppose there are I areas of interest and there are n_i patients who stay in the ith area during the research period. Let T be T_{ij} which denote the failure time for the jth subject in the ith area. Let $\boldsymbol{X_{ij}}$ denote the covariate vector and $\boldsymbol{\beta}$ is a vector of regression coefficients. Furthermore, ϕ_i will denote the spatial frailty for the ith area and ξ_i is a vector of frailties for classes in the predictors, which is the exponential of random effects in ith area. The corresponding survival function for T_{ij} given by $\boldsymbol{X_{ij}}$ and ϕ_i can then be expressed as

$$S\left(t\middle|\boldsymbol{X_{ij}}, \phi_i, \xi_i\right) = exp\left\{-\Lambda_0(t)\left[\xi_{i0}\prod_{h=1}^{H}\xi_{ih}^{\omega_{ijh}}\right]exp\left(\boldsymbol{X'_{ij}}\boldsymbol{\beta} + \phi_i\right)\right\} \tag{2}$$

where $\Lambda_0(t) = \int_0^t h_0(s)ds$ is the unspecified cumulative baseline hazard function. In this article, only binary predictors (gender, treatment) were considered as potential frailties, denoted as $\omega_{ijh}(h = 1, \dots, H)$, which more intuitively reflects the initial frailty term. To characterize the heterogeneity among frailties and different areas, we assume the frailties and spatial information are independent. Also, our model is developed under the

assumption that different patients are statistically independent form each other and that each patient's failure time should be distributed identically given the covariates, area and frailties.

Due to the nature of general interval-censored data, the exact failure time T_{ij} for jth subject in ith area cannot be examined exactly. Let (L_{ij}, R_{ij}) be a pair of observation intervals which contain the true unobserved failure time T_{ij}. Take $(0, R_{ij})$ in the case of left-censoring, and (L_{ij}, ∞) in the case of right-censoring. Furthermore, let δ_{ij1}, δ_{ij2}, δ_{ij3} be the left-, interval- and right- censoring indicators for jth subject in ith area within ith cluster. As a consequence, for each subject there is $\delta_{ij1}+ \delta_{ij2}+ \delta_{ij3} = 1$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \Lambda_0)$ be the unknown parameters in our model, then the observed data likelihood is

$$LK_{obs}(\boldsymbol{\theta}) = \prod_{i=1}^{I} \left\{ \iint LK_i(\boldsymbol{\theta}|\phi_i, \xi_i)\pi(\phi_i)\pi(\xi_i)\, d\phi_i d\xi_i \right\}, \tag{3}$$

where $LK_i(\boldsymbol{\theta}|\phi_i, \xi_i) = \prod_{j=1}^{n_i} LK_{ij}\,(\boldsymbol{\theta}|\phi_i, \xi_i)$ is the conditional likelihood contributed by patients in ith area given spatial cluster ϕ_i and frailty ξ_i. In addition, $\pi(\phi_i)$is the probability function of ϕ_i and $\pi(\xi_i)$ is the probability function of ξ_i. Then under the Cox model, the likelihood of an i.d.d. $(\delta_{ij1}, \delta_{ij2}, \delta_{ij3}, L_{ij}, R_{ij}, \boldsymbol{X_{ij}})$ conditional on the spatial frailty ϕ_i and covariate frailty ξ_i is proportional to

$$LK_{ij}(\boldsymbol{\theta}|\phi_i, \xi_i) = \left\{F\left(R_{ij}\middle|\boldsymbol{X_{ij}}, \phi_i, \xi_i\right)\right\}^{\delta_{ij1}} \left\{F\left(R_{ij}\middle|\boldsymbol{X_{ij}}, \phi_i, \xi_i\right) - F\left(L_{ij}\middle|\boldsymbol{X_{ij}}, \phi_i, \xi_i\right)\right\}^{\delta_{ij2}} \left\{1 - F\left(L_{ij}\middle|\boldsymbol{X_{ij}}, \phi_i, \xi_i\right)\right\}^{\delta_{ij3}} \tag{4}$$

where $F(t|\boldsymbol{X}, \phi, \xi) = 1 - S(t|\boldsymbol{X}, \phi, \xi)$ is the conditional cumulative distribution function of T given covariate $\boldsymbol{X}$, area ϕ and frailty ξ.

2.2 PRIORS AND THE FRAILTY SELECTION

For shared frailty model, Dunson and Chen [31] proposed a mixture prior probability distribution in order to allow frailty selection, which can also be useful in a multiple frailty model since the variance of frailty can be both zero and non-zero. Furthermore, for avoiding the MCMC sample of variance being focused on zero alone during iteration, Pan et al. proposed a re-parameterized variance parameter to include frailties with latent variables. This study follows the suggestion from Dunson, Chen, and Pan et al.

Let $\xi_{ih}, h = 0, \dots, H$ represent each frailty in our model. following the suggestion from Dunson and Chen, Gamma density represents frailty when variance is not zero, since Gamma distribution is in an exponential family and is always positive. The assumption is that the frailty density should be:

$$\pi(\xi_{ih}; k_h) = \begin{cases} 1(\xi_{ih} \equiv 1) & if\ k_h = 0 \\ Ga(\xi_{ih}; k_h^{-1}, k_h^{-1}) & if\ k_h > 0, \end{cases} \quad (5)$$

where $\mathrm{Ga}(\xi_{ih}; k_h^{-1}, k_h^{-1})$ denotes the Gamma density with mean 1 and precision k_h^{-1}. When $\xi_{ih} \equiv 1$ and(?) $k_h = 0$ for all i that represents the null hypothesis of (what) there is no heterogeneity between clusters for hth frailty. In other words, variance k_h parameter indexes the degree of between-cluster variability and thus the level of within-cluster dependence. [32] For computing efficiency and simplifying the variance selection procedure, following the suggestion from Dunson and Chen, the mixture prior was chosen to address the problem of excessive numbers of zero

$$\pi(k_h) = 1(k_h = 0)\pi_{0h} + 1(k_h > 0)(1 - \pi_{0h})IG(k_h; a_h, b_h) \quad (6)$$

where $\pi_{0h} = \Pr(H_{0h}: k_h = 0)$ is the prior probability of the null hypothesis of homogeneity for the hth frailty, and $\mathrm{IG}(k_h; a_h, b_h)$ is the inverse- Gamma prior density

of variance under the alternative hypothesis with shape parameter a_h and scale parameter b_h. The form of (6) is a zero-inflated inverse-Gamma density.

Following the suggestion from Pan et al., k_h and ξ_{ih} was re-parameterized with latent variables $\tilde{k}_h$ and $\tilde{\xi}_{ih}$. Therefore, if ρ_h is an indicator then :

$$\begin{cases} H_{0h}: k_h = 0 (i.e. homogeneity) & then\ \rho_h = 1 \\ H_{Ah}: k_h > 0 (i.e. heterogeneity) & then\ \rho_h = 0. \end{cases} \tag{7}$$

Based on the form of (6) , we have $\rho_h \sim \text{Bernoulli}(\pi_{0h})$. Then for simplification and efficiency, under both null hypothesis and alternative hypothesis, k_h and ξ_{ih} $can\ be$ re-parameterized as follows:

$$k_h = (1-\rho_h)\tilde{k}_h, \quad \xi_{ih} = \tilde{\xi}_{ih}^{1-\rho_h}, \quad \text{for } i = 1, \dots I, \ \ h = 1, \dots, H,$$

where $\tilde{k}_h \sim IG(a_h, b_h), \tilde{\xi}_{ih} \sim \text{Ga}(\tilde{k}_h^{-1}, \tilde{k}_h^{-1})$ with mean 1 and variance $\tilde{k}_h$. The most efficient prior density for $(\tilde{k}_h, \tilde{\xi}_h, \rho_h)$ according to equation (5), (6)and (7) is as follows:

$$\pi(\tilde{k}_h, \tilde{\xi}_h, \rho_h) = \text{IG}(\tilde{k}_h;\ a_h, b_h) \left\{ \prod_{i=1}^{I} \text{Ga}\left(\tilde{\xi}_{ih};\ \tilde{k}_h^{-1}, \tilde{k}_h^{-1}\right) \right\} \pi_{0h}^{\rho_h} (1-\pi_{0h})^{1-\rho_h}.$$

2.3 MODELING $\boldsymbol{\Lambda_0(t)}$ WITH MONOTONE SPLINES

For right-censored data, the partial likelihood method under the PH model estimates the regression parameters directly without the need of estimating the cumulative hazard function. However, the partial likelihood does not exist for interval-censored data [33] under the PH model due to the complexity of the data structure, and we need to estimate both $\boldsymbol{\beta}$ and $\Lambda_0(t)$ simultaneously. Following the path of Cai et al.[34], Wang and Dunson[35], and Lin and Wang, $\Lambda_0(t)$ was modeled by a linear combination of monotone I-spline[36] which not only improved the efficiency of our

model by reducing the number of parameters of nonreducing function $\Lambda_0(t)$, but also provided enough flexibility for our model. Specifically, the baseline cumulative function of the model proposed is:

$$\Lambda_0(t) = \sum_{l=1}^{k} \gamma_l b_l(t), \tag{8}$$

where $\{b_l\}$ is a set of basis I-splines which are nondecreasing in range so that $(0,1)$, $\{\gamma_l\}$ is a set of non-negative coefficients and l represents a different section of basis I-splines and $l = (1, \dots, L)$. Nondecreasing and nonnegative are shared qualities of both I-spline and $\Lambda_0(t)$ which make using I-spline to model $\Lambda_0(t)$ attractive. One can use a nonnegative, linear combination of B-spline [37] to model the baseline cumulative hazard function in the PH model as Sharef et al.[38] and Zhang et al.[39] exibited. But a spline is increasing if the coefficients of the linear combination of B-splines are increasing. Thus, an increasing spline can be fit by restricting the coefficients of the linear combination so that they are increasing, again using the I-spline basis is better than using B-spline. This paper refers the reader to De Leeuw [40] for more details about the difference between I-spline and B-spline. When the baseline cumulative hazard the B-spline was evaluated it was found to require too many numerical approximations of integrals, which is also quite inefficient.

The shape of I-spline is determined by the degree of monotone splines and the number of knots. Furthermore, the number of basis splines equals the sum of the number of interior knots and the degree of splines. When degree value is equal to 1, 2, 3, the corresponding basis spline will be linear, quadratic and cubic respectively. Reference was made to Ramsay et al.[36] where more details about I-spline basis function can be found.

In general, the random placement of knots may cause a large number of knots and overfitting which results in an inefficiency. In contrast, we should allow a set of selected knots with a fixed number and location in order to avoid a complicated selection method and computational expense. In this paper, following Cai et al.[34], Lin & Wang[7] and Dunson & Chen[31], we take 2 or 3 as degree values and 10-30 equally spaced knots for adequate smoothness and modeling flexibility for the purpose of reduction of computation time. The shrinkage prior for the spline coefficient γ_l serves to prevent overfitting, which is caused by excessive knots and causes the small ones to be shrunk zero.

2.4 DATA AUGMENTATION

The direct computation of the complicated observed likelihood in (3) is not possible since the integral does not have an explicit form. For easily computing the likelihood function, we treat all ϕ_i and ξ_i as unknown parameters:

$$LK(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\xi}) = \prod_{i=1}^{I} \{LK_i(\boldsymbol{\theta}|\phi_i, \xi_i)\pi(\phi_i)\pi(\xi_i)\}, \tag{9}$$

where $\boldsymbol{\phi} = (\phi_1, \dots, \phi_I)$ and $\boldsymbol{\xi} = (\xi_1', \dots, \xi_I')'$.

However, it is still difficult to find standard posterior distribution. To facilitate the posterior computation, we considered a two-step data augmentation by taking advantage of the relationship between the PH model with spline from (8) and the latent nonhomogeneous Poisson process. Specifically, when the time interval is $(0, t]$, we assumed the number of occurrences $\{N(t): t > 0\}$ to be a nonhomogeneous Poisson process with a cumulative intensity function

$$\phi\Lambda_0(t)\left\{\left[\xi_0\prod_{h=1}^{H}\xi_h^{\omega_h}\right]\exp(\boldsymbol{X}'\boldsymbol{\beta}+\phi)\right\},$$

Given area ϕ and frailties ξ, we defined $T=\inf\{t: N(t)>0\}$ as the time of the first occurrence in the Poisson process. This latent process is valid for all t, since

$$P(T>t|\phi,\xi)=P(N(t)=0|\phi,\xi)=\exp\left(-\Lambda_0(t)\left\{\left[\xi_0\prod_{h=1}^{H}\xi_h^{\omega_h}\right]\exp(\boldsymbol{X}'\boldsymbol{\beta}+\phi)\right\}\right).$$

Hence, time points t_1 and t_2 can be defined such that $t_1<t_2$. In following the path of Cai et al.[9],

$$\text{let}\begin{cases} Z=N(t_1)\text{ be the count of occurrences in }(0,t_1];\\ W=[N(t_2)-N(t_1)]\text{ be the count of occurrence in }(t_1,t_2]\end{cases}$$

$$\text{also}\begin{cases} Z=N(t_1)\sim\text{Poi}\left(\Lambda_0(t_1)\left\{\left[\xi_0\prod_{h=1}^{H}\xi_h^{\omega_h}\right]\exp(\boldsymbol{X}'\boldsymbol{\beta}+\phi)\right\}\right);\\ W=[N(t_2)-N(t_1)]\sim\text{Poi}\left([\Lambda_0(t_2)-\Lambda_0(t_1)]\left\{\left[\xi_0\prod_{h=1}^{H}\xi_h^{\omega_h}\right]\exp(\boldsymbol{X}'\boldsymbol{\beta}+\phi)\right\}\right)\end{cases}$$

it should be noticed that Z and W are independent, conditional on area ϕ and frailties ξ, by the properties of nonhomogeneous Poisson's process. Specifically, as mentioned in 2.1, the observed failure time falls into the interval $(L,R]$. For jth subject in ith area it follows that:

(still: $t_{ij1}<t_{ij2}$)

$$\begin{cases}\text{left censoring: } t_{ij1}=R_{ij}\,, t_{ij2}>R_{ij}\text{ and }Z_{ij}>0\\ \text{interval censoring: } t_{ij1}=L_{ij}, t_{ij2}=R_{ij}\text{ and }Z_{ij}=0, W_{ij}>0\\ \text{right censoring: } 0<t_{ij1}<L_{ij}\,, t_{ij2}=L_{ij}\text{ and }Z_{ij}=0, W_{ij}=0\,.\end{cases}$$

It is worth mentioning that for left-censoring data, W_{ij} can take any value, since t_{ij2} is some point greater than the boundary R_{ij} and for $t_{ij2}>R_{ij}$ and belongs to the observed

time period which doesn't need to be estimated at all. Then the augmented data can be formulated for the likelihood of the jth subject in ith area as

$$L_{aug1_{ij}}(\boldsymbol{\theta}|Z_{ij}, W_{ij}, \phi_i, \xi_i)$$

$$= \left[\Lambda_0(t_1)\left\{\left[\xi_0\prod_{h=1}^{H}\xi_h^{\omega_h}\right]\exp(\boldsymbol{X'\beta}+\phi)\right\}\right]^{\delta_{ij}}\exp\left(-\Lambda_0(t_1)\left\{\left[\xi_0\prod_{h=1}^{H}\xi_h^{\omega_h}\right]\exp(\boldsymbol{X'\beta}+\phi)\right\}\right)$$

$$= \mathrm{Poi}(Z_{ij})\mathrm{Poi}(W_{ij})^{\delta_{ij2}+\delta_{ij3}}\{1(Z_{ij}>0)\}^{\delta_{ij1}}\{1(Z_{ij}=0)1(W_{ij}>0)\}^{\delta_{ij2}}\{1(Z_{ij}=0)1(W_{ij}=0)\}^{\delta_{ij3}}, \qquad (10)$$

where $1(\cdot)$ is the indicator function. Integrating Z_{ij} and W_{ij} out of (10) leads back to the conditional likelihood of $LK_{ij}(\boldsymbol{\theta}|\phi_i, \xi_i)$ in (4).

For calculating close-formed posteriors for a monotone spline basis based on the additive property of Poisson distribution and the linearity form of $\Lambda_0(t)$ in (8), data can be augmented by decomposing Z_{ij} as $\sum_l^L Z_{ijl}$ and W_{ij} as $\sum_l^L W_{ijl}$, where Z_{ijl} and W_{ijl} are independent. Furthermore, applying (8):

$$\begin{cases} Z_{ijl}\sim\mathrm{Poi}\left((\gamma_l b_l(t_{ij1})\left\{\left[\xi_0\prod_{h=1}^{H}\xi_{ih}^{\omega_{ijh}}\right]\exp(\boldsymbol{X'\beta}+\phi)\right\}\right); \\ W_{ijl}\sim\mathrm{Poi}\left([\gamma_l b_l(t_{ij2})-\gamma_l b_l(t_{ij1})]\left\{\left[\xi_{i0}\prod_{h=1}^{H}\xi_{ih}^{\omega_{ijh}}\right]\exp(\boldsymbol{X'\beta}+\phi)\right\}\right), \end{cases}$$

$$\text{and } \begin{cases} \text{left censoring } (\delta_{ij1}=1): Z_{ij}=\sum_{l}^{L} Z_{ijl}>0 \\ \text{interval censoring } (\delta_{ij2}=1): Z_{ij}=\sum_{l}^{L} Z_{ijl}=0, W_{ij}=\sum_{l}^{L} W_{ijl}>0 \\ \text{right censoring } (\delta_{ij3}=1): Z_{ij}=\sum_{l}^{L} Z_{ijl}=0, W_{ij}=\sum_{l}^{L} W_{ijl}=0\,. \end{cases}$$

For jth subject in the ith area the likelihood function can be further augmented as:

$$L_{aug2_{ij}}(\boldsymbol{\theta}|Z's, W's, \phi_i, \xi_i) = \left\{\prod_{l=1}^{L} \text{Poi}(Z_{ijl})\text{Poi}(W_{ijl})^{\delta_{ij2}+\delta_{ij3}}\right\}\{1(Z_{ij}>0)\}^{\delta_{ij1}}$$

$$\{1(Z_{ij}=0)1(W_{ij}>0)\}^{\delta_{ij2}}\{1(Z_{ij}=0)1(W_{ij}=0)\}^{\delta_{ij3}}. \qquad (11)$$

Via the introduction of unobserved data or latent variables, the likelihood function in (11) forms the basis of computing the posterior distribution in the next section of this article.

2.5 SPATIAL FRAILTIES

The Banerjee et al.[26] model shows a spatial arrangement into two general settings: the geostatistical approach, which contains the exact geographic locations of subjects, and the lattice approach, which only uses the reginal summary data to find the relevance of each area. The Conditional Autoregressive (CAR) model developed by Besag [41] is a common method in the lattice approach. Following Besag and Kooperberg[42], let $\boldsymbol{\phi} = (\phi_1, \dots, \phi_I)$ be the spatial random vector which has density

$$p(\boldsymbol{\phi}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\phi}' Q \boldsymbol{\phi}\right\}, \qquad (12)$$

where Q is an $I \times I$ positive definite symmetric matrix. By using the Brook expansion [43] in the Markov random field approach we can get $p(\boldsymbol{\phi})$ via I full conditional

distribution or local characteristics $p(\phi_i|\phi_{-i})$, where ϕ_{-i} represents all spatial frailties other than ϕ_i. Then

$$\phi_i|\phi_{-i}\sim N\left(\sum_g \zeta_{ig}\phi_g,\ k_i\right), \text{for } i = 1, \dots, I, \tag{13}$$

where $\zeta_{ii} = 0$, $\zeta_{ig} = -(Q_{ig}/Q_{ii})1_{(i\sim g)}$, $1_{()}$ is an indicator function, and $k_i = 1/\ Q_{ii}$. Q_{ig} represents a connection between area i and g, and Q_{ii} is set to 0 and $Q_{ig} = 1$ if area i and g are neighbors. The symmetric of Q requires $\zeta_{ig}k_g = \zeta_{gi}k_i$. Following Besag and Kooperberg [42], $\boldsymbol{\phi}$ in (12) is replaced with $\boldsymbol{\phi} + \mu$,where μ is an arbitrary real I vector with a corresponding adjustment to (13). Then,

$$\boldsymbol{\phi}'Q\boldsymbol{\phi} \equiv \sum_i Q_{i+}\phi_i^2 - \sum_{i<g} Q_{ig}\left(\phi_i - \phi_g\right)^2, \tag{14}$$

where $Q_{i+} = \sum_{g=1}^{I} Q_{ig}$. Besag and Kooperberg [42] propose a boundary in the parameter space where $Q_{i+} = 0$. In the limiting forms of standard Gaussian conditional autoregressions, for vector $\mathbf{1}$ and vector $\mathbf{0}$**,** we have $Q\mathbf{1} = \mathbf{0}$ even though Q is well defined [42]. The form is called intrinsic conditional autoregressive model. (12) and (13) are still valid, but the positive definiteness of Q must now be replaced by positive semi-definiteness, therefore, the variance matrix Q^{-1} no longer exists. Then, based on (14) we can reform (12) as follows:

$$p(\boldsymbol{\phi}) \propto \exp\left\{-\frac{1}{2}\sum_{i<g} Q_{ig}\left(\phi_i - \phi_g\right)^2\right\}. \tag{15}$$

Furthermore, studies such as Banerjee et al. [26], Zhou et al.[27] and Hesam et al.[28], as well as others, prefer to specify $\zeta_{ig} = \frac{1}{u_i}1_{(i\sim g)}$ like weighted ζ_{ig}, where u_i is the number

of neighbors for the ith area. Then $Q = \tau_\phi W$, where $W_{ii} = u_i$, τ_ϕ is a precision parameter and $W_{ig} = -1_{(i\sim g)}$. Now, (15) can be reformed as

$$p(\boldsymbol{\phi}) \propto \exp\left\{-\frac{\tau_\phi}{2}\sum_{i<g}\left(\phi_i - \phi_g\right)^2 1_{(i\sim g)}\right\}. \tag{16}$$

Then, (13) can be rewritten as:

$$\phi_i|\phi_{-i} \sim N\left(\bar{\phi}_{\partial_i}, \frac{1}{u_i\tau_\phi}\right), \text{for } i = 1, \dots, I, \tag{17}$$

where $\bar{\phi}_{\partial_i}$ is the average number of the neighbors where ϕ_{-i} is adjacent to ϕ_i. Equation (17) is to be used as prior for ϕ_i.

We also need to involve precision parameter τ_ϕ in the joint density of spatial frailty since the prior function (17) contains τ_ϕ. The density of $\boldsymbol{\phi}$ is as follow:

$$p(\boldsymbol{\phi}) \propto \tau_\phi^{\frac{I-B}{2}} \exp\left\{-\frac{\tau_\phi}{2}\sum_{i<g}\left(\phi_i - \phi_g\right)^2 1_{(i\sim g)}\right\},$$

where B is termed an island which has no connection to other areas$B = 1$ in the data used.[44]

2.6 POSTERIOR COMPUTATION

Gibbs sampling was adopted for posterior computations. The full conditional distributions for all unknown parameters can then be derived by combining the augmented data likelihood (11) and priors. The steps of derivation and the specifications of the initial values of the unknown parameters are below:

(i) Let $Z_{ij} = 0$ and $W_{ij} = 0$ for all i and $j, Z_{ijl} = 0$ and $W_{ijl} = 0$ for all i, j and l. If $\delta_{ij1} = 1$ (i.e. left-censoring), then sample

$$Z_{ij}\sim\text{Poi}\left(\Lambda_0(R_{ij})\left\{\left[\xi_0\prod_{h=1}^{H}\xi_{ih}^{\omega_{ijh}}\right]\exp\left(\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\phi_i\right)\right\}\right)1(Z_{ij}>0),$$

$$(Z_{ij1},\dots,Z_{ijL}|\,Z_{ij})\sim\text{Multinomial}(Z_{ij},\boldsymbol{p}_{ij}),\qquad \text{with } \boldsymbol{p}_{ij}=(p_{ij1},\dots,p_{ijL}),$$

$$\text{and } p_{ijl}=\frac{\gamma_l b_l(R_{ij})}{\sum_{l'}^{L}\gamma_{l'}b_{l'}(R_{ij})},\quad l=1,\dots,L \text{ and } l' \text{means all the } l.$$

If $\delta_{ij2}=1$ (i.e. interval-censoring), then sample

$$W_{ij}\sim\text{Poi}\left([\Lambda_0(R_{ij})-\Lambda_0(L_{ij})]\left\{\left[\xi_{i0}\prod_{h=1}^{H}\xi_{ih}^{\omega_{ijh}}\right]\exp\left(\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\phi_i\right)\right\}\right)1(W_{ij}>0),$$

$$(W_{ij1},\dots,W_{ijL}|\,W_{ij})\sim\text{Multinomial}(W_{ij},\boldsymbol{q}_{ij}),\qquad \text{with } \boldsymbol{q}_{ij}=(q_{ij1},\dots,q_{ijL}),$$

$$\text{and } q_{ijl}=\frac{\gamma_l[b_l(R_{ij})-b_l(L_{ij})]}{\sum_{l'}^{L}\gamma_{l'}[b_{l'}(R_{ij})-b_{l'}(L_{ij})]},\quad l=1,\dots,L \text{ and } l' \text{means all the } l.$$

(ii) For regression coefficient $\beta_r,\ r=1,\dots,p$, we assume a normal prior $N(0,\sigma_0^2)$ for β_r. The adaptive rejection Metropolis sampling (ARMS)[45] and the adaptive rejection sampling (ARS)[46] are used to sample from the posterior distribution for each β_r, since the posterior of each β_r is log-concave and the posterior of each β_r is not conjugate. The full conditional distribution of β_r is:

$$P(\beta_r|Z's, W's, \phi_i, \xi_i, \beta_{-r})$$

$$\propto \exp\left[\sum_{i=1}^{I}\sum_{j=1}^{n_i}\exp\left\{X_{ijr}\beta_r + \phi_i + log(\xi_0) + \sum_{h=1}^{H}\omega_{ijh}\, log(\xi_h)\right\}\right.$$

$$-\sum_{i=1}^{I}\sum_{j=1}^{n_i}\exp\left\{\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \phi_i + log(\xi_0)\right.$$

$$\left.\left.+\sum_{h=1}^{H}\omega_{ijh}\, log(\xi_h)\right\}\left\{\Lambda_0\left(R_{ij}\right)\left(\delta_{ij1}+\delta_{ij2}\right) + \Lambda_0\left(L_{ij}\right)\delta_{ij3}\right\}\right]\cdot\pi(\beta_r),$$

(iii) For sample $\gamma_l, l = 1, \ldots, L$, when an independent exponential prior $\exp(\eta)$ for γ_ls and a Gamma hyperprior $\mathrm{Ga}(a_\eta, b_\eta)$ for η are assigned. This prior specification leads to conjugate forms for each of the conditional posterior distributions of γ_ls and η and penalize large values of the coefficients γ_ls and functions to shrink the coefficients of those unnecessary spline bases towards 0. Sample γ_l from Gamma distribution $\mathrm{Ga}(a_{\gamma_l}, b_{\gamma_l})$, where

$$a_{\gamma_l} = 1 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Z_{ijl}\delta_{ij1} + W_{ijl}\delta_{ij2}\right)$$

$$b_{\gamma_l} = \eta + \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left\{\left[\xi_{i0}\prod_{h=1}^{H}\xi_{ih}^{\omega_{ijh}}\right]\exp\left(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \phi_i\right)\left\{b_l\left(R_{ij}\right)\left(\delta_{ij1}+\delta_{ij2}\right)\right.\right.$$

$$\left.\left.+\, b_l\left(L_{ij}\right)\delta_{ij3}\right\}\right\}.$$

(iv) Sample η from $\mathrm{Ga}(a_\eta + L, b_\eta + \sum_{l=1}^{L}\gamma_l)$.

(v) Sample $\phi_i, i = 1, \dots, I$, the posterior for each ϕ_i is not conjugate. The Metropolis-Hastings (MH) algorithm was used for sampling. The full conditional distribution is:

$$P(\phi_i | Z's, W's, \boldsymbol{\theta}, \xi_i, \phi_{-i})$$

$$\propto \exp\left[\sum_{j=1}^{n_i}\sum_{i=1}^{I}\phi_i\left(Z_{ij}\delta_{ij1} + W_{ij}\delta_{ij2}\right)\right.$$

$$-\sum_{j=1}^{n_i}\sum_{i=1}^{I}\exp\left(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \phi_i + log(\xi_{i0})\right.$$

$$\left.\left.+\sum_{h=1}^{H}\omega_{ijh}\, log(\xi_h)\right)\left\{\Lambda_0\left(R_{ij}\right)\left(\delta_{ij1} + \delta_{ij2}\right) + \Lambda_0\left(L_{ij}\right)\delta_{ij3}\right\}\right] \cdot P(\phi_i | \phi_{-i}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \Lambda_0)$ and $P(\phi_i | \phi_{-i})$ denote the prior in (17).

(vi) For frailty selection, we need to update $(\boldsymbol{\rho}, \boldsymbol{k}, \boldsymbol{\xi})$ through the following sub-steps:

- Sample $\rho_h, h = 0, \dots, H$, from Bernoulli$(\tilde{\pi}_h)$, where

$$\tilde{\pi}_h = \frac{\pi_{0h}}{\pi_{0h} + (1 - \pi_{0h})C},$$

  where $C = L(\Lambda_0, \boldsymbol{\beta}, \xi_h = \tilde{\xi}_h, \xi_{(-h)}) / L(\Lambda_0, \boldsymbol{\beta}, \xi_h \equiv 1, \xi_{(-h)})$ and $\xi_{(-h)}$ denotes all frailties except the hth frailty.

- Sample $\tilde{k}_h, h = 0, \dots, H$, using ARMS from its full conditional distribution proportional to

$$\tilde{k}_h | \cdot \propto \exp\left(-\tilde{k}_h^{-1}\left[b_h + \sum_{i=1}^{I}\tilde{\xi}_{ih} - \sum_{i=1}^{I}\log\left(\tilde{\xi}_{ih}\right)\right]\right) \cdot \left\{\frac{(\tilde{k}_h^{-1})^{\tilde{k}_h^{-1}}}{\Gamma(\tilde{k}_h^{-1})}\right\}^I \cdot (\tilde{k}_h^{-1})^{a_h+1}$$

- Sample $\tilde{\xi}_{i0}, i = 1, \dots, I$, from

$$\mathrm{Ga}\Bigg(\tilde{k}_0 + (1-\rho_0)\sum_{j=1}^{n_i}\sum_{i=1}^{I}\left(Z_{ij}\delta_{ij1} + W_{ij}\delta_{ij2}\right),$$

$$\tilde{k}_0^{-1} + (1-\rho_0)\sum_{j=1}^{n_i}\sum_{i=1}^{I}\exp\left(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + \phi_i\right)\left[\prod_{h=1}^{H}\xi_{ih}^{\omega_{ijh}}\right]\left\{\Lambda_0\left(R_{ij}\right)\left(\delta_{ij1} + \delta_{ij2}\right)\right.$$

$$\left. + \Lambda_0\left(L_{ij}\right)\delta_{ij3}\right\}\Bigg).$$

- For each $h = 1, \dots, H$, sample $\tilde{\xi}_{ih}, i = 1, \dots, I$, from

$$\mathrm{Ga}\Bigg(\tilde{k}_0 + (1-\rho_0)\sum_{j=1}^{n_i}\sum_{i=1}^{I}\omega_{ijh}\left(Z_{ij}\delta_{ij1} + W_{ij}\delta_{ij2}\right),$$

$$\tilde{k}_h^{-1} + (1-\rho_h)\sum_{j=1}^{n_i}\sum_{i=1}^{I}\omega_{ijh}\exp(\boldsymbol{X}_i'\boldsymbol{\beta} + \phi_i)\xi_{i0}\prod_{h|-h}\xi_{i(-h)}^{\omega_{ij(-h)}}\left\{\Lambda_0\left(R_{ij}\right)\left(\delta_{ij1} + \delta_{ij2}\right)\right.$$

$$\left. + \Lambda_0\left(L_{ij}\right)\delta_{ij3}\right\}\Bigg).$$

(vii) For the spatial precision parameter τ_ϕ, a Gamma prior $\mathrm{Ga}(a_\tau, b_\tau)$ was assumed, which leads to conjugate posterior. The parameter τ_ϕ is derived from $\mathrm{Ga}(\frac{I-B}{2} + a_\tau, b_\tau + \frac{1}{2}\sum_{i<g}(\phi_i - \phi_g)^2)1_{(i\sim g)}$.

2.7 MODEL COMPARISON

To test the performance of the proposed model, it has been compared to a Weibull proportional hazards model with a spatial parameter and multiple frailties. The Weibull model is very flexible and also has theoretical justification in many applications, which also allows the use of CAR distribution as prior for spatial information. For comparing

the competing models and selecting the best one that fits the data in both simulation and real data analysis, the following two Bayesian model selection criteria were considered, they are the Bayes factor and the deviance information criterion (DIC)[47].

Based on the Gibbs sampler, the marginal posterior probability of homogeneity for each frailty was estimated by getting the mean of probability of homogeneity from all iterations (S):

$$\hat{\pi}_h = \frac{1}{S}\sum_{s=1}^{S} \tilde{\pi}_h^{(s)}, h = 1, \dots, H.$$

Then, assuming the prior probability for both homogeneity and heterogeneity are 0.5, the Bayes factor of homogeneity based on marginal posterior probability is:

$$BF_h = \frac{\Pr(\rho_h = 1|\boldsymbol{L}, \boldsymbol{R}, \boldsymbol{x})}{\Pr(\rho_h = 0|\boldsymbol{L}, \boldsymbol{R}, \boldsymbol{x})}, h = 1, \dots, H.$$

The smaller the Bayes factor (<1) shows better evidence of heterogeneity.

Spiegelhalter et al.[47] provides evidence that DIC is a suitable measure of model complexity even in hierarchical settings, and thus, DIC is considered as a sensible generalization of the expected Akaike information criterion to hierarchical settings. The model, with the smallest value of DIC, is commonly taken as the preferred model to describe the data set given. DIC can be readily computed in Markov chain Monte Carlo analysis.

CHAPTER 3. SIMULATION STUDIES

A simulation study was conducted to evaluate the proposed method, using 100 replications of simulated data with sample size N=460. Specifically, we utilize map of South Carolina counties. The 46 counties were considered as separate spatial areas containing 10 patients each. Data was generated with two variables, where x_{ij1} and x_{ij2} follows Bernoulli distribution with a probability of 0.5.

$$\{\ x_{ij1} \sim Bernoulli(0.5), \qquad x_{ij2} \sim Bernoulli(0.5)\ \}$$

For each dataset, the survival probabilities were generated from a PH model with frailties and spatial parameters:

$$S\left(t \middle| x_{ij1}, x_{ij2}, \phi_i, \xi_i\right) = \exp\left\{-\Lambda_0(t)\xi_{i0}\xi_{i1}^{x_{ij1}}\xi_{i2}^{x_{ij2}}\exp\left(x_{ij1}\beta_1 + x_{ij2}\beta_2 + \phi_i\right)\right\}, \tag{18}$$

where we set $\Lambda_0(t) = \log(1+t)$, $\xi_{i0} \equiv 1$(homogeneity in baseline), $\xi_{i1} \sim \text{Ga}(2,2)$ with mean of 1, variance of 0.5 and $\xi_{i2} \sim \text{Ga}(2,2)$ with mean of 1, variance of 0.5, $\beta_1 = \beta_2 = 1$, $\tau_\phi = 4$. The simulation scenario of proposed model are compared with a typical Weibull PH model. Each subject is assumed to have a random number of observations, determined by 1 plus a Poisson random variable with mean 2. The observation times were gathered by generating gap times between adjacent observation times from independent exponential distributions with mean 1. The observed interval containing the true failure time was determined by the two adjacent observation times (from zero to infinite). To generate spatial parameter ϕ_i, we sample ϕ_i^* first from multivariate normal

$N\left(0, \left(\tau_\phi W^*\right)^{-1}\right)$ with $W^* = W + \text{diag}(0.0001, I)$, followed by centering ϕ_i^* to get the spatial parameter of ϕ_i.

To construct monotone splines, each simulated data set was determined by using degree=2 and 25 equally spaced knots between 0 and the maximum value of the finite endpoints of all the observed intervals for. For prior, $\pi_{0h} = 0.5$, $a_h = b_h = 0.01$ where $h = 0, 1, 2$ was chosen. For hyper-parameters, $\sigma_0 = 10$, $a_\tau = b_\tau = 0.01$was employed. The range of ARMS was greater than negative ten and smaller than ten. Finally, 13000 Monte Carlo samples were generated for each dataset, with the first 3000 burn-in. The simulation results for each model are summarized over the 100 datasets.

Table 3.1 Posterior probability of frailty selection and corresponding Bayes factor in favor of homogeneity.

	Proposed model		Weibull model	
	Estimate	Bayes Factor	Estimate	Bayes Factor
$\tilde{\pi}_0 = \Pr(\rho_0 = 1 \mid \text{data})$	0.9443	17	0.9390	15
$\tilde{\pi}_1 = \Pr(\rho_1 = 1 \mid \text{data})$	0.0209	1/47	0.0192	1/51
$\tilde{\pi}_2 = \Pr(\rho_2 = 1 \mid \text{data})$	0.0337	1/29	0.0134	1/74

Based on a classification scheme for the Bayes factor, as proposed by Jeffreys and reformed by Wagenmakers et al.[48] and Christian P. et al.[49], there was very strong evidence for H_0 with a Bayes factor that falls between 30 and 100. Additionally, very strong evidence for H_0 with a Bayes factor greater or equal to 30 was observed, along with substantial evidence of H_0 with a Bayes factor between 3 and 10. Given the result shown in Table 3.1 it was summarized that under the proposed model the Bayes factor of homogeneity for intercept is 17 and is 15 under Weibull model. This is strong evident that there is no heterogeneity in baseline risk. The Bayes factor for heterogeneity of frailty for covariate x_1 is 47, indicating a very strong evidence that there is a

heterogeneous impact from x_1 on the outcome. The Bayes factor for heterogeneity of frailty for covariate x_2 is 29, which indicates a very strong evidence that there is heterogeneity impact from x_2 on the outcome.

Table 3.2 Posterior probability of the eight possible models in terms of frailty selection.

(ρ_0, ρ_1, ρ_2)	(1,0,0)	(0,0,0)	(1,0,1)	(1,1,0)	(1,1,1)	(0,0,1)	(0,1,1)	(0,1,0)
Proposed model	0.9110	0.0442	0.0185	0.0091	0.0057	0.0054	0.0041	0.0020
Weibull model	0.9102	0.0535	0.0137	0.0089	0.0062	0.0034	0.0024	0.0017

The summarized outcome of frailty selection term ρ_h is shown in Table 3.2. There are 8 possible combinations of three selection terms (ρ_0, ρ_1, ρ_2). Both the proposed model and the Weibull model illustrate the potential structure of the simulated datasets. It's obvious that the highest probability of combinations is the true structure (ρ_0, ρ_1, ρ_2)=(1,0,0) with a 91.10% probability in the proposed model and a 91.02% probability in the Weibull model. The results were closely compared to two models of each potential combination. Another essential term for detecting heterogeneity across clusters is k_h which is estimated from $\tilde{k}_h$ (not shown in table). From the proposed model it was determined that (k_0, k_1, k_2)=(0.094,0.472,0.455), which is as close to the true value as (0,0.5,0.5). The estimated r from the Weibull model is reported as 0.62.

Table 3.3 Estimation of regression coefficients β_h and spatial parameter τ_ϕ.

		Proposed model				Weibull model			
	TRUE	Estimate	SSD	ESE	95CP	Estimate	SSD	ESE	95CP
β_1	1	1.041	0.232	0.267	0.95	1.031	0.248	0.255	0.94
β_2	1	1.033	0.241	0.256	0.94	1.037	0.231	0.249	0.94
τ_ϕ	4	3.764	0.542	1.204	1.00	4.095	0.537	1.144	1.00
DIC		742				745			

Table 3.3 summarizes the estimated results of the proposed model versus the Weibull model. For each parameter, the point estimated is the average of the 100

posterior means from 100 datasets, the empirical standard error (ESE) is the average of the 100 estimated standard errors, the sample standard deviation (SSD) is the sample standard deviation of the 100 posterior means, and the 95% coverage probability (95CP) is the percent of the 100 credible intervals of each parameter that contains the true parameter value. The proposed model accomplished its purpose as anticipated. Most of the parameters have coverage probabilities close to the nominal level of 0.95. Both the Weibull model and the proposed model perform well in predicting τ_ϕ and β_1, β_2. The DIC value from Weibull model (745) is very close to the value from proposed model (742).

Figure 3.1 shows the estimated baseline survival functions from true value, in both the proposed and Weibull models. Baseline survival function was collected based on 100 equally spaced points ranging from 0.1 to 10 by 0.1. The shadowed area is the 95% pointwise credible interval for estimated baseline survival functions under the proposed model. The standard error of each $S_0(t_i)$ for calculating the 95 credible intervals were obtained from each t_i time point repeated by 100 datasets. Both models match the pattern of true values but the model proposed herein works better than Weibull model.

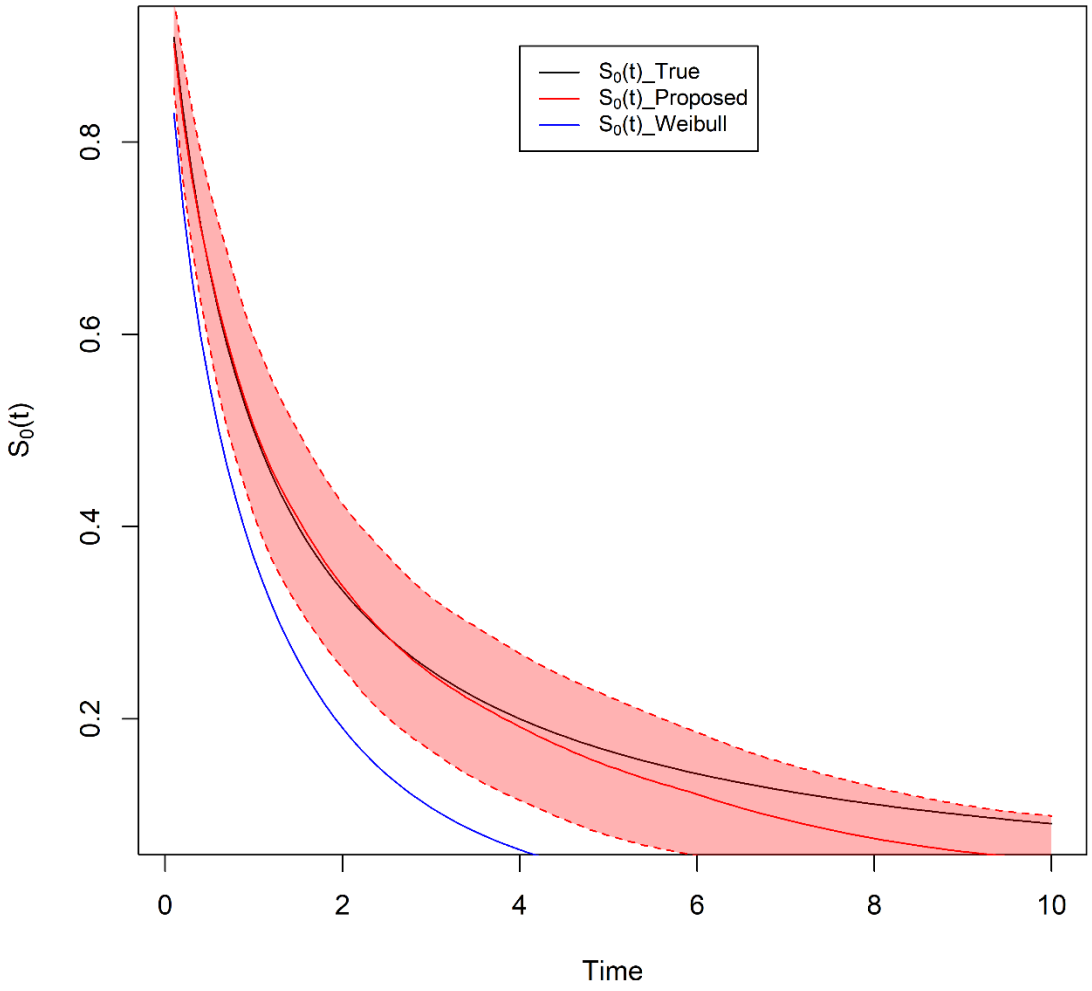

Figure 3.1 Plot of estimated baseline survival curve based on 100 simulated datasets from true value for both the proposed model (95% pointwise credible intervals) and the Weibull model.

The posterior means of spatial parameter ϕ_i for each South Carolina county is plotted in Figure 3.2 for both the proposed and Weibull models. The patterns of ϕ_i are slightly different between the two models and two plots in Figure 3.2 share the same grayscale. For the proposed model, the higher ϕ_i is clustered in the South-East areas of SC; for the Weibull model, the higher ϕ_i scan be seen in the Northern part of the state.

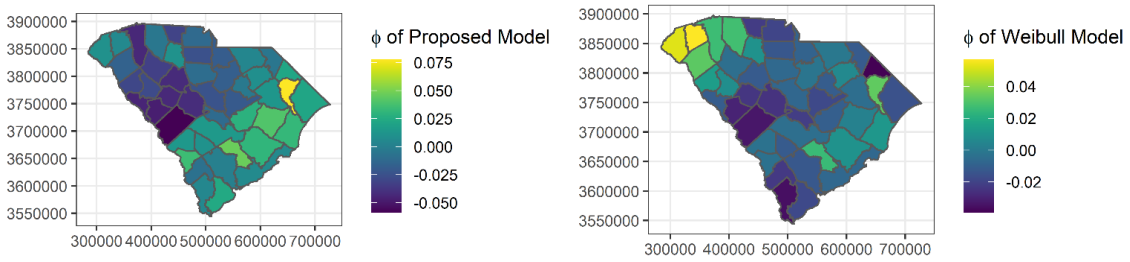

Figure 3.2 Maps of posterior means of the spatial parameter ϕ_i over 46 counties of SC based on the proposed model and the Weibull model.

CHAPTER 4. REAL DATA ANALYSIS

Table 4.1 Estimation of the proposed model and the Weibull model for Minnesota smoke cessation data.

	Proposed model		Weibull model	
	Estimate	95% CI	Estimate	95% CI
Gender (male=0)	0.417	(-0.309, 1.112)	0.526	(-0.280, 1.144)
Treatment (UC=0)	-0.625	(-1.310, -0.144)	-0.513	(-1.101, -0.082)
DIC	427		475	

The proposed model was applied to the smoking cessation dataset in Minnesota which includes 223 participants with geographical information of 51 zip codes. This data is a subset from a lung health study carried out by Murray et al.[50] whose intention was to test the effect of intermittent smoking on pulmonary functions. All patients involve in the analysis have zip codes and records of gender and treatment. The maximum time of record is 5.5 years. 70.85% of the patients have no record of relapse until the last visit. 29.15% of patients have no precise time of relapse but have two observed time points referring to the observed interval where the true time to relapse falls. There is no left-censoring involved in this dataset. Only two binary variables were kept. The first was gender and the second was treatment, since this model only works well within binary predictors. 60.99% of the patients involved in the study were male. There were 2 kinds of treatments: first, 169 subjects were in smoking intervention (SI) groups and secondly, 54 are in usual care (US) groups, male and US have been treated as reference group, no interaction term add in model.

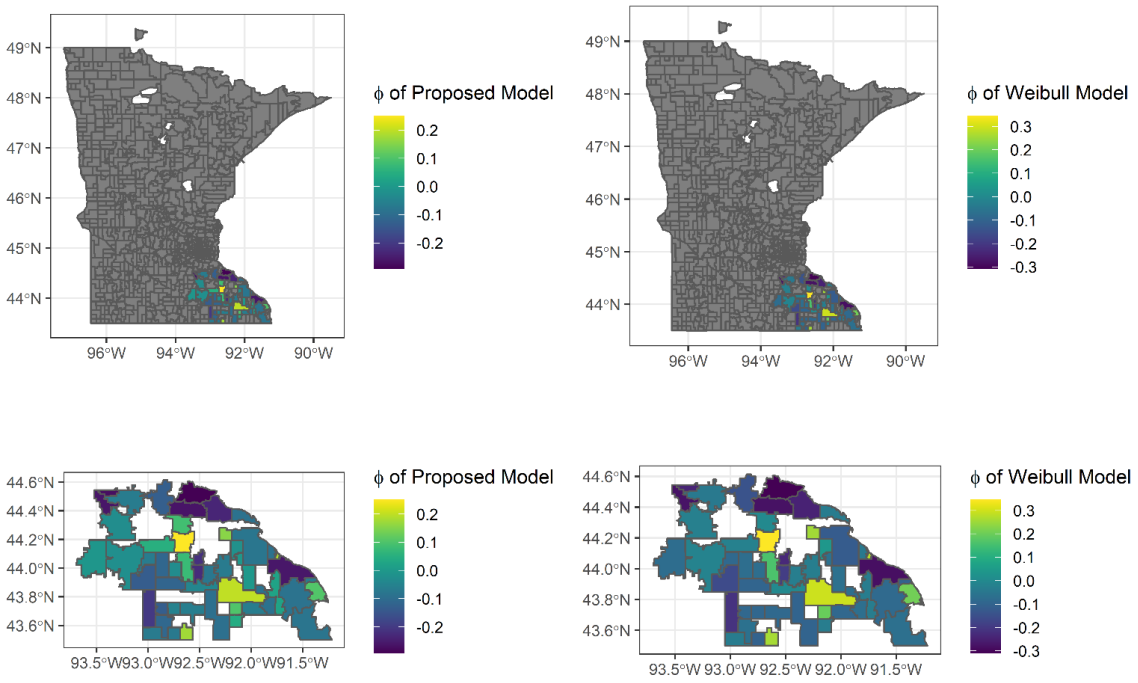

Figure 4.1 Maps of posterior means of the spatial parameter ϕ_i over the 51 zip code areas of Minnesota based on the proposed model and the Weibull model.

The multiple frailty model $\lambda\left(t\middle|x_{ij1}, x_{ij2}, \phi_i, \xi_i\right) = \lambda_0(t)\xi_{i0}\xi_{i1}^{x_{ij1}}\xi_{i2}^{x_{ij2}}\exp\left(x_{ij1}\beta_1 + x_{ij2}\beta_2 + \phi_i\right)$ fits this real data. The proposed model uses $\pi_{0h} = 0.5$, $a_h = b_h = 0.01$ as a value for prior,where $h = 0, 1, 2$. For constructing monotone spline, degree=2 and 40 equally spaced knots between 0 and the maximum value of the finite endpoints of all the observed intervals were utilized for the Minnesota smoking cessation dataset. 13000 Monte Carlo samples were generated for the Minnesota dataset, with the first 3000 as a burn-in. The regression estimation results are presented in Table 4.1. The estimated regression coefficients from both models have the same direction but different magnitudes. For the results from the proposed model, it was observed that the treatment had a significant impact on reducing the risk of relapse. Both models indicate the

significant impact of treatment and non-significant impact of sex on the risk of smoking relapse. In proposed model, a 60.28 % chance of female patients quit smoking faster compared to male and a 34.86% chance of patients involved in intervention (SI) groups have no relapse compared to usual care (US) groups. In Weibull model, a 62.85% chance of female patients quit smoking faster compared to male and a 37.45% chance of patients involved in intervention (SI) groups have no relapse compared to usual care (US) groups. The DIC value from proposed model (427) is smaller than that from Weibull model (475). The highest probability of combinations (ρ_0, ρ_1, ρ_2)=(0,1,1) is 87.72% in the proposed model. Bayes factor of heterogeneity for intercept is 12 means strong evidence for heterogeneity of zip code area-wise variation in baseline hazard function. The Bayes factor for homogeneity of frailty for treatment is 9 and 10 for gender, indicating positive evidence of homogeneity in both treatment and gender across zip code areas.

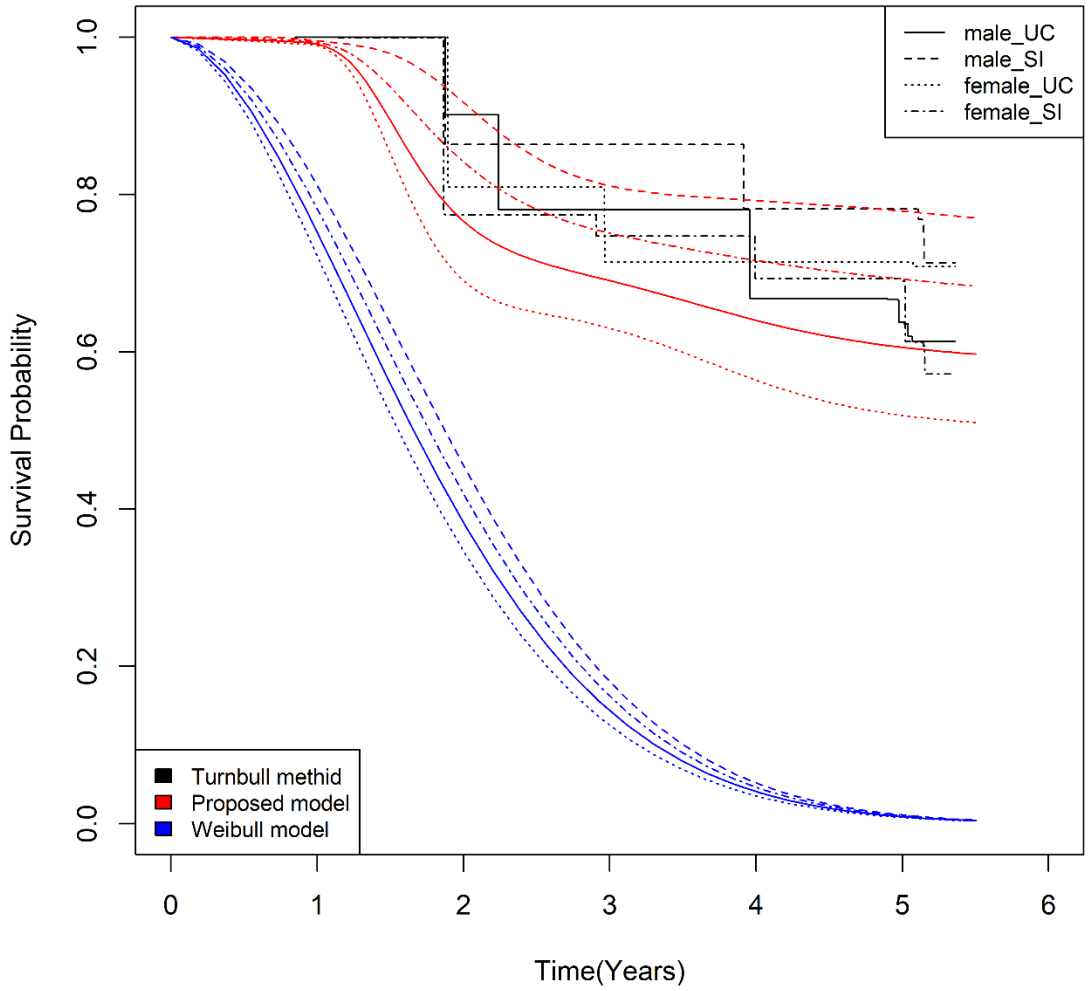

Figure 4.2 Estimated survival curves for the smoking cessation study, using the Turnbull method, the proposed model, and the Weibull model. The event of interest is the time between cessation and relapse.

In Figure 4.1, posterior means of the spatial parameter ϕ_i over 51 zip code areas of Minnesota was plotted based on the proposed model and the Weibull model. There are only a few zip code areas that have records of the subjects, most subjects are clustered in southeast Minnesota. The map based on the proposed model shows lower values for the ϕ_i in the middle regions, which indicates lower risks of relapse in those regions and an

extremely higher risk of relapse in the north and east areas from both models. The patterns of spatial information were similar between two models. In Figure 4.2, the estimated survival functions were plotted based on the proposed model, comparing the nonparametric maximum likelihood estimates (NPMLEs) using the Turnbull method in R. Turnbull proposes a nonparametric maximum likelihood estimator of survival function for interval-censored data. From the plot, the estimated lines from both the proposed and Turnbull models were similar but distant from the estimated Weibull model lines. The proposed model can be said to perform better with covariate gender and treatment than the Weibull model.

CHAPTER 5. DISCUSSION

In this paper, an efficient semiparametric method was developed under the PH multiple frailty model to deal with spatially clustered interval-censored data. The advantage of this method is that not only were the regression parameters and essential predictors estimated, but also the baseline hazard function, indicating that the nonparametric formulation of a monotonic spline provides more flexible approximation for baseline hazard functions compared to parametric assumptions.

Conversely, there are disadvantages to the proposed method in that it is only friendly to binary outcomes. There are indeed 5 variables which had a potential impact on reducing smoking relapse risk: gender, treatment, duration as a smoker (years), average number of cigarettes per day over the last 10 years, zip code area space that can be used in predicting the result. Since only 2 variables are binary, 3 potential predictors must be eliminated. Based on the study and research on this topic, the conclusion formed was that the causation of this issue is focused on ARMS (Adaptive Rejection Metropolis Sampling) function in R. Future research will endeavor to control for continuous variables in ARMS.

REFERENCES

1. Finkelstein DM and Wolfe RA. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* 1985; 41: 933–945.

2. Groeneboom, P., Wellner, J.A., 1992. Information Bounds and Nonparametric Maximum Likelihood Estimation. In: *DMV Seminar Band*, vol. 19. Birkhäuser, Basel.

3. Cox, D.R., 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B. Methodol*. 34, 187–220 (with discussion).

4. Yavuz AC and Lambert P. Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine* 2011; 30: 75–90.

5. Timothy Hanson & Wesley O Johnson (2004) A Bayesian Semiparametric AFT Model for Interval-Censored Data, *Journal of Computational and Graphical Statistics*, 13:2, 341-361, DOI: 10.1198/1061860043489.

6. Komarek A, Lesaffre E. Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution. *Statistica Sinica* 2007; 17:549-569.

7. Lin, X., Wang, L., 2010. A semiparametric probit model for case 2 interval-censored failure time data. S *Statistics in Medicine* 29, 972–981.

8. Sinha D, Chen M-H and Ghosh SK. Bayesian analysis and model selection for interval-censored survival data. *Biometrics* 1999; 55: 585–590.

9. Lin X, Cai B, Wang L and Zhang Z. A Bayesian proportional hazards model for general interval-censored data. *Lifetime Data Anal*. July 2014. DOI: 10.1007/s10985-014-9305-9.

10. Wang X, Chen MH, Yan J (2013) Bayesian dynamic regression models for interval censored survival data with application to children dental health. *Lifetime Data Anal* 19:297–316

11. Zhang Z and Sun J. Interval censoring. *Statistical Methods in Medical Research* 2010; 19: 53–70.

12. Li Li, Alejandro Jara, María José García-Zattera & Timothy E. Hanson (2018): Marginal Bayesian Semiparametric Modeling of Mismeasured Multivariate Interval-Censored Data, *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2018.1476240.

13. Fei Gao, Donglin Zeng, David Couper & D. Y. Lin (2018): Semiparametric Regression Analysis of Multiple Right- and Interval-Censored Events, *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2018.1482756.

14. Pak, D., Li, C., & Todem, D. (2018). Semiparametric analysis of correlated and interval-censored event-history data. *Statistical Methods in Medical Research*.

15. Mao, L., Lin, D., & Zeng, D. (2017). Semiparametric regression analysis of interval-censored competing risks data. *Biometrics*, 73 3, 857-865.

16. Pak, D. (2018). Analysis of complex life-history data and variable selection in survival analysis under interval censoring (Order No. 10931298). Available from ProQuest Dissertations & Theses Global. (2100930323).

17. Zhou, H., Hanson, T. & Zhang, J., 2018. spBayesSurv: Fitting Bayesian Spatial Survival Models Using R. *Journal of Statistical Software,* doi: 10.18637/jss.v000.i00.

18. Pan, C., Cai, B & Wang, L. (2013). Models and Software Development for Interval-Censored Data.

19. McMahan CS, Wang L (2014). ICsurv: A Package for Semiparametric Regression Analysis of Interval-Censored Data. R package version 1.0, URL [https://CRAN.R-project.org/ package=ICsurv].

20. Anderson-Bergman, C. (2017). icenReg: Regression Models for Interval Censored Data in R. *Journal of Statistical Software*, 81(12), 1 - 23. doi: [http://dx.doi.org/10.18637/jss.v081.i12].

21. Touraine, C., Gerds, T., & Joly, P. (2017). SmoothHazard: An R Package for Fitting Regression Models to Interval-Censored Observations of Illness-Death Models. *Journal of Statistical Software*, 79(7), 1 - 22. doi: [http://dx.doi.org/10.18637/jss.v079.i07].

22. Wang L, McMahan C, Hudgens M, Qureshi Z (2016). A Flexible, Computationally Efficient Method for Fitting the Proportional Hazards Model to Interval-Censored Data. *Biometrics*, 72(1), 222–231. doi:10.1111/biom.12389.

23. Vaupel, J.W., Manton, K.G. & Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* (1979) 16: 439. [https://doi.org/10.2307/2061224]

24. Andreas Wienke, 2003. Frailty models, *MPIDR Working Papers* WP-2003-032, Max Planck Institute for Demographic Research, Rostock, Germany.

25. Hazarika, J., & Mahanta, K.K. (2016). Studies on Survival Analysis using Frailty Models under Bayesian Mechanism and its Further Scope. *IOSR Journal of Mathematics* DOI: 10.9790/5728-1204051521.

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

27. Zhou, H., Hanson, T. & Zhang, J. Generalized accelerated failure time spatial frailty model for arbitrarily censored data. *Lifetime Data Anal* (2017) 23: 495. DOI: [https://doi.org/10.1007/s10985-016-9361-4]

28. Hesam, S., Mahmoudi, M., Foroushani, AR, Yaseri, M. & Mansournia, MA. (2018). A cause-specific hazard spatial frailty model for competing risks data. *Spatial Statistics* Volume 26,2018,Pages 101-124.

29. Yiqi, B., Cancho, V. G., Louzada, F., & Suzuki, A. K. (2017). Cure rate proportional odds models with spatial frailties for interval-censored data. *Communications for Statistical Applications and Methods*, 24(6), 605-625. [https://doi.org/10.29220/CSAM.2017.24.6.605]

30. Pan, C., Cai, B., & Wang, L. (2017). Multiple frailty model for clustered interval-censored data with frailty selection. *Statistical Methods in Medical Research*, 26(3), 1308–1322.

31. Dunson, D., & Chen, Z. (2004). Selecting Factors Predictive of Heterogeneity in Multivariate Event Time Data. *Biometrics*, 60(2), 352-358. Retrieved from [http://www.jstor.org/stable/3695762]

32. Glidden DV and Vittinghoff E. Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, 2004; 23: 369–388.

33. Sun, J., 2006. *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer.

34. Cai B, Lin X and Wang L. Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics & Data Analysis* 2011; 55: 2644–2651.

35. Wang L and Dunson DB. Semiparametric Bayes proportional odds models for current status data with under-reporting. *Biometrics* 2011; 67: 1111–1118.

36. Ramsay JO. Monotone regression splines in action. *Statistical Science* 1988; 3: 425–441.

37. Carl de Boor. 2001. *A Practical Guide to Splines*. Revised Edition. New York: Springer-Verlag.

38. Sharef, E.; Strawderman, RL.; Ruppert, D.; Cowen, M. & Halasyamani, L. Bayesian adaptive B-spline estimation in proportional hazards frailty models. *Electronic Journal of Statistics*. 4 (2010), 606--642. doi:10.1214/10-EJS566. [https://projecteuclid.org/euclid.ejs/1278439436]

39. Zhang, Y., Hua, L., Huang, J., 2010. A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics*. 37 (2), 338–354.

40. De Leeuw, Jan. (2017). Computing and Fitting Monotone Splines. 10.13140/RG.2.2.36758.96327. [http://gifi.stat.ucla.edu/splines/splines.html]

41. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*. 36(2):192–236

42. Besag, J., Kooperberg, C., 1995. On conditional and intrinsic autoregressions. *Biometrika* 82, 733–746.

43. Brook, D., 1964. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* 51 (3/4), 481-483.

44. Hodges, J.S., Carlin, B.P., Fan, Q., 2003. On the precision of the conditionally autoregressive prior in spatial models. *Biometrics* 59, 317–322.

45. Gilks WR, Best N and Tan K. Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of Applied Statistics* 1995; 44: 455–472.

46. Gilks, W.R., Wild, P., 1992. Adaptive rejection sampling for Gibbs sampling. *Journal of Applied Statistics* 41, 337–348.

47. Spiegelhalter, DJ, Best, NG, Carlin, BP, and van der Linde, A (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B (Statistical Methodology)*. 64, 583-639.

48. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432. [https://doi.org/10.1037/a0022790]

49. Christian P. Robert, Nicolas Chopin, Judith Rousseau "Harold Jeffreys's Theory of Probability Revisited," *Statistical Science*. 24(2), 141-172, (May 2009)

50. Murray RP, Anthonisen NR, Connett JE, Wise RA, Lindgren PG, Greene PG, Nides MA. Effects of multiple attempts to quit smoking and relapses to smoking on pulmonary function. Lung Health Study Research Group. *Journal of Clinical Epidemiology*. 1998 Dec;51(12):1317-26. doi: 10.1016/s0895-4356(98)00120-6. PMID: 10086826.

51. Therneau, Terry & Lumley, Thomas. (2016). Survival: Survival Analysis.

52. Komárek, A., Lesaffre, E., and Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, 14, 726–745.

53. Turnbull BW. The Empirical distribution function with arbitrarily grouped, censored and truncated Data. *Journal of the Royal Statistical Society: Series B* 1976; 38: 290–295.