Published in final edited form as:

Stat Methods Med Res. 2017 October ; 26(5): 2244–2256. doi:10.1177/0962280215596186.

Bayesian accelerated failure time model for space-time dependency in a geographically augmented survival model

Georgiana Onicescu¹, Andrew Lawson¹, Jiajia Zhang², Mulugeta Gebregziabher¹, Kristin Wallace¹, and Jan M Eberth²

¹Department of Public Health Sciences, Medical University of South Carolina, Charleston, USA

²Department of Biostatistics and Epidemiology, Arnold School of Public Health, University of South Carolina, Columbia, USA

Abstract

In this paper, we extend the spatially explicit survival model for small area cancer data by allowing dependency between space and time and using accelerated failure time models. Spatial dependency is modeled directly in the definition of the survival, density, and hazard functions. The models are developed in the context of county level aggregated data. Two cases are considered: the first assumes that the spatial and temporal distributions are independent; the second allows for dependency between the spatial and temporal components. We apply the models to prostate cancer data from the Louisiana SEER cancer registry.

Keywords

Accelerated failure time; Bayesian hierarchical models; spatial analysis; Markov chain Monte Carlo; prostate cancer

1 Introduction

Accelerated failure time (AFT) models assume that the effect of the covariates act multiplicatively on the failure time by linearly modeling the natural logarithm of the survival time.¹ While this is analogous to the classical linear model, for survival data the model is restricted by the limited choices one can assume for the distribution of the survival times. The advantages of AFT models are their flexibility and intuitive interpretability of the covariates effect on the mean survival time.

Modeling the spatial variation using the proportional hazards models is widely used.^{2–5} Henderson et al.² investigated possible spatial variation in survival of adult acute myeloid leukemia patients in northwest England. Banerjee et al.³ modeled the spatial arrangement of the strata in two ways, one using geostatistical approaches, where the exact geographic

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Georgiana Onicescu, Medical University of South Carolina, 135 Cannon Street, Suite 303, Charleston, SC 29425, USA. onicescu@musc.edu.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

locations were used, and another using lattice approaches. Li et al.⁴ extended the ordinary frailty models by allowing random effects accommodating spatial correlations to enter into the baseline hazard function multiplicatively. Hennerfeind et al.⁵ proposed flexible continuous-time geoadditive models. The AFT model provides an alternative when the proportional hazards assumption does not hold. Bayesian spatial random effects AFT models have been used previously in order to investigate geographical variations in survival rates.^{6,7} However, they rely on the use of random effects to account for spatial variation. Zhang et al.⁶ investigated the prostate cancer (PrCa) data of Louisiana from the SEER program and the violation of the proportional hazards assumption suggested that the spatial survival model based on the AFT model is more appropriate for this data set.

Wang et al.⁷ proposed a normal mixture AFT spatial model, which does not require the proportional hazards assumption and allows the multi-model distribution to be modeled. Methodology including the geographical risk in the definition of the survival measures was recently proposed for the case of independent space and time models⁸ in the context of a Weibull survival model using Bayesian approaches. This model found a high county level variability of the PrCa risk, with higher risk regions in the south-east. Simulations showed a very good fit for both the spatio-temporal survival Weibull (STSU-Weibull) and random effects model, with a slightly overall better fit of the STSU-Weibull.

PrCa is a major public health problem, being the most common cancer among men.⁹ Approximately 15.3% of men will be diagnosed with PrCa at some point during their lifetime, based on 2008–2010 data.¹⁰ PrCa survival rates have been shown to cluster spatially and change over time.¹¹ In 2010, the PrCa incidence rates in the US varied at county level from 90.6 to 187.0 out of 100,000 men.¹² Many studies have found that there is regional variation in the PrCa survival.^{13–16} Thus, it is important to account for the geographical nature of the data in order to better estimate survival rates and related measures.

The objective of this paper is to extend the spatially explicit survival methodology by allowing for dependency between the spatial and temporal component and to exemplify our approach using PrCa incidence data from the Louisiana SEER cancer registry. Our outcome is death from any cause, which could also include non-PrCa related deaths such as accidents or other age related diseases or deaths from other cancers.

The all cause mortality endpoint depends only on accurate ascertainment of deaths and when they occur, and therefore eliminates attribution bias (i.e. incorrect labeling of death from other causes as death from PrCa).

The inclusion of the spatial component in the definition of survival and related measures avoids the use of random effects to allow for county level correlations. However, due to computational feasibility, the dependency between space and time is accomplished by the use of spatial parish level random effects. We fit and compare models with both independent and dependent space time components. The results may be instrumental in planning for PrCa health resource allocation that take into account regional health disparities.

2 Spatial model

The traditional approach to spatial model construction is to use functions such as semivariogram or covariance to describe the spatial correlation. However, certain requirements, such as positiveness of the covariance function, make it impractical to use for complex models with large datasets. An alternative for directly specifying the covariance function is to use the kernel convolution model,¹⁷ which is based on the idea that any stationary Gaussian process can be expressed as the convolution of a white noise process $x(s)$ with a specified kernel $k(s)$. Kernel convolution have been used previously for various Bayesian modeling,^{18,19} including the spatially explicit survival modeling.⁸

In order to construct the spatial model, we used a process convolution using a Gaussian kernel function, allowed to vary over the study area. As described in Higdon et al.,¹⁷ the model for the spatial process is determined by specifying the white noise process $x(s)$ and the smoothing kernel $k(s)$.

We have chosen the Gaussian kernel $k(s) \approx e^{-\frac{1}{2}\|s\|^2}$ ($\|s\|$ being the Euclidian norm), since it induces a covariance matrix which is a function of the squared distance between two spatial locations and gradually dies off with increased distance. Specifically, denoting $s - s'$ the displacement vector between the locations s and s' , the Gaussian kernel induces the covariance¹⁷

$$c(d) = \text{Cov}(z(s), z(s')) \propto e^{-\frac{1}{2}\left\|\frac{d}{\sqrt{2}}\right\|^2}$$

The Gaussian process $z(s)$ can be constructed over a spatial region $\mathbb{S}$ as follows^{17,20}

$$z(s) = \int_{\mathbb{S}} k(u-s)x(u)du, \quad s \in \mathbb{S} \quad (1)$$

Since the above integration cannot be explicitly solved, the integral can be approximated by a finite sum

$$z(s) = \sum_{j=1}^{n_g} x_j k(w_j - s), \quad x_j \sim N(0, \sigma^2) \quad (2)$$

where $w_j, j = 1, \dots, n_g$ are the grid points, n_g is the total number of grid points over the area. Variations of the above formula can be used by restricting the domain of the white noise process.¹⁷ In order to have positiveness of the white noise process, we used the following formula

$$z(s) = \sum_{j=1}^{n_g} e^{x_j} k(w_j - s), \quad x_j \sim N(0, \sigma^2) \quad (3)$$

For an area A we computed

$$z_A = \int_A z(s) ds \approx \sum_{s \in A} \sum_{j=1}^{n_g} e^{x_j} k(w_j - s) \quad (4)$$

Since in the application to our PrCa data example we had access only to the number of deaths and total population of interest in each parish, we used the coordinates of the centroid of each parish and approximated the above formula with the following

$$z_A \approx p_A \sum_{j=1}^{n_g} e^{x_j} k(w_j - c_A) \quad (5)$$

where p_A is the percentage of the number of deaths in partition A and c_A is the centroid of partition A.

For each parish A_l , $l = 1, \dots, L$ the kernel was computed at the differences between the centroid of the parish and each grid point in the whole polygon region. We stored these values in the kernel matrix K_{ij} , $i = 1, \dots, L$, $j = 1, \dots, n_g$ of dimension $L \times n_g$, n_g being the number of gridpoints. z_{A_l} was therefore calculated by multiplying the percentage of deaths for county A with the scalar product between the l^{th} row of the kernel matrix and the exponentiated white noise vector $(e^{x_1}, \dots, e^{x_{n_g}})'$.

The integral over an area of the spatial probability density function (pdf) is defined as the ratio between the county z_A and the total z_{A_T} , as described in Onicescu et al.⁸

$$\int_A f_1(s) ds = \frac{z_A}{z_{A_T}}, \quad z_{A_T} = \sum_{l=1}^L z_{A_l} \quad (6)$$

L being the total number of partitions of the study region.

3 Definitions and notations

The generalized pdf f is allowed to have both a temporal and spatial component. We defined $f: \mathbb{R} \times \mathbb{R} \rightarrow (0, \infty)$, $f(s, t) = f_1(s) f_2(t)$, where $f_1(s)$ is the spatial pdf and $f_2(t)$ is the temporal pdf given the spatial location s defined by latitude and longitude and t being a time.

The integral of the spatial pdf $f_1(s)$ is defined as described in the previous section

$$\int_A f_1(s) ds = \frac{z_A}{z_{A_T}}$$

The temporal pdf is defined in the context of the AFT model. Let n = number of subjects, t_i = survival times, $i = 1, \dots, n$, independent, identically distributed with density function $f(s, t)$, c_i = censoring time, $i = 1, \dots, n$

$$y_i = \min(c_i, t_i), \quad \nu_i = \begin{cases} 1 & \text{if } t_i \leq c_i \\ 0 & \text{if } t_i > c_i \end{cases}$$

Let $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ and $\mathbf{\nu} = (\nu_1, \nu_2, \dots, \nu_n)'$. Let $\mathbf{P}$ be the vector of parameters of interest.

The AFT model can be expressed as

$$\log(y_i) = \mathbf{M}'\boldsymbol{\beta} + \sigma\varepsilon_i, \quad i = 1, \dots, n$$

where $\mathbf{M} = (\mathbf{1}, \mathbf{m}^1, \dots, \mathbf{m}^p)$ is a matrix of covariates including the intercept term, $\mathbf{m}^1, \dots, \mathbf{m}^p$ are the p covariates, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ is the vector of unknown coefficients, σ is the scale parameter, and $\boldsymbol{\varepsilon}$ is the vector of errors, which can have various distributional assumptions. We assumed that $\boldsymbol{\varepsilon}$ had a standard logistic distribution, which implies a log logistic distribution of times given the spatial location. This is a flexible distribution assumption that also leads to closed form of the survival function and related measures. The pdf of $\boldsymbol{\varepsilon}$ is defined as follows

$$f_{\varepsilon}(x) = \frac{e^x}{(1+e^x)^2}$$

This corresponds to the log-logistic temporal pdf with parameters $f_2(y_i) = \frac{\lambda_i \alpha y_i^{\alpha-1}}{(1+\lambda_i y_i^{\alpha})^2}$, $\alpha = \frac{1}{\sigma}$.

In order to introduce the spatial dependency, we define $f_2(y_i | s_i) = \frac{\lambda_i \alpha y_i^{\alpha-1}}{(1+\lambda_i y_i^{\alpha})^2}$ and add the spatial random effects $g(s_i)$ to the linear component of the λ parameter as follows

$$\log(\lambda_i) = \frac{-\mathbf{M}'\boldsymbol{\beta}}{\sigma} + \log(g(s_i))$$

where s_i is the location corresponding to subject i and $g(s_i)$ is a location specific random effect.

The corresponding temporal cumulative distribution functions, as well as survival and hazard functions can be defined as follows

$$F_t(t^*) = \int_0^{t^*} f_2(t) dt \quad (7)$$

where $F_t(t^*)$ is the probability that the time random variable takes a value less or equal to t^*

$$F_s(A_s) = \int_{A_s} f_1(s) ds \quad (8)$$

where $F_s(A_s)$ is the marginal probability over the space A_s

$$\begin{aligned} F_t(s, t^*) &= \int_0^{t^*} f(s, t) dt \\ &= \int_0^{t^*} f_1(s) f_2(t|s) dt \\ &= f_1(s) \int_0^{t^*} f_2(t|s) dt \end{aligned} \quad (9)$$

where $F_t(s, t^*)$ is the probability that the time random variable takes a value less or equal to t^* at the spatial location s

$$\begin{aligned} F_s(A_s, t^*) &= \int_{A_s} f(s, t^*) ds \\ &= \int_{A_s} f_1(s) f_2(t^*|s) ds \\ &\approx \int_{A_s} f_1(s) f_2(t^*|A_s, s \in A_s) ds \\ &= f_2(t^*|A_s, s \in A_s) \int_{A_s} f_1(s) ds \end{aligned} \quad (10)$$

where $F_s(A_s, t^*)$ is the probability that the time random variable takes the value t^* over the domain A_s .

$$\begin{aligned} F_{s,t}(A_s, t^*) &= \int_{A_s} \int_0^{t^*} f(s, t) dt ds \\ &= \int_{A_s} \int_0^{t^*} f_1(s) f_2(t|s) dt ds \\ &\approx \int_{A_s} \int_0^{t^*} f_1(s) f_2(t|A_s, s \in A_s) dt ds \\ &= \int_{A_s} f_1(s) ds \int_0^{t^*} f_2(t|A_s) dt \\ &= F_s(A_s) F_t(t^*|A_s) \end{aligned} \quad (11)$$

where $F_{s,t}(A_s, t^*)$ is the joint probability that the time random variable takes a value less or equal to t^* over the domain A_s .

The corresponding survival and hazard functions are defined as follows

$$S_t(s, t^*) = 1 - F_t(s, t^*) \quad (12)$$

$$\begin{aligned} S_{s,t}(A_s, t^*) &= P(s \in A_s, t > t^*) \\ &= P(s \in A_s) P(t > t^* | A_s) \\ &= \int_{A_s} f_1(s) ds (1 - F_t(t^* | A_s)) \end{aligned} \quad (13)$$

$$h_t(s, t^*) = \frac{f(s, t^*)}{S_t(s, t^*)} \quad (14)$$

$$\begin{aligned} h_{s,t}(A_s, t^*) &= \frac{\int_{A_s} f(s, t^*) ds}{S_{s,t}(A_s, t^*)} \\ &\approx \frac{F_s(A_s, t^*)}{F_s(A_s)(1 - F_t(t^* | A_s))} \end{aligned} \quad (15)$$

For the log-logistic distributed times given the spatial location, the spatial pdf and survival are defined as:

$$\begin{aligned} F_s(A_l, y_i, \mathbf{m}_i) &= \int_{A_l} f(s, y_i, \mathbf{m}_i) \\ &= \int_{A_l} f_1(s) f_2(y_i, \mathbf{m}_i | s) ds \\ &\approx \int_{A_l} f_1(s) f_2(y_i, \mathbf{m}_i | A_s, s \in A_s) ds \\ &= \int_{A_l} \frac{z(s)}{z_{A_T}} \frac{\lambda_i \alpha y_i^{\alpha-1}}{(1 + \lambda_i y_i^\alpha)^2} ds \\ &= \frac{z(s)}{z_{A_T}} \frac{\lambda_i \alpha y_i^{\alpha-1}}{(1 + \lambda_i y_i^\alpha)^2} \end{aligned} \quad (16)$$

where $F_s(A_l, y_i, \mathbf{m}_i)$ and related measures are modified to display the covariate dependence.

$$\begin{aligned} S_{s,t}(A_l, y_i, \mathbf{m}_i) &= \int_{A_l} f_1(s) (1 - F_t(y_i, \mathbf{m}_i | s)) ds \\ &\approx \int_{A_l} f_1(s) (1 - F_t(y_i, \mathbf{m}_i | A_s, s \in A_s)) ds \\ &= \int_{A_l} \frac{z(s)}{z_{A_T}} ds (1 - F_t(y_i, \mathbf{m}_i | A_s)) \\ &= \frac{z_{A_l}}{z_{A_T}} \left(1 - \frac{\lambda_i y_i^\alpha}{1 + \lambda_i y_i^\alpha} \right) \end{aligned} \quad (17)$$

4 Area likelihood

Let $A_1, \dots, A_L$ be a partition of the study region and let n_l be the number of people in area A_l . The area likelihood under the AFT assumption, dependency model is constructed as follows

$$L_A(P|n, \tilde{y}) = \prod_{l=1}^L \prod_{i=1}^{n_l} \left[\frac{\lambda_i \alpha y_i^{\alpha-1}}{(1+\lambda_i y_i^\alpha)} \int_{A_l} f_1(s) ds \right]^{\nu_i} \left[\int_{A_l} f_1(s) ds \left(1 - \frac{\lambda_i y_i^\alpha}{1+\lambda_i y_i^\alpha} \right) \right]^{1-\nu_i}$$

$$\alpha = \frac{1}{\sigma}, \quad \log(\lambda_i) = \frac{-\mathbf{M}'\beta}{\sigma} + \log(g(s_i)) \quad (18)$$

where $g(s_i)$ is a location specific random effect. For computational reasons, we approximated $g(s_i)$ with $g(A_s)$, $s_i \in A_s$, therefore, assuming the same random effect within an area.

5 Computational approaches

5.1 Prior distributions

For each of the coefficients $\beta_0, \dots, \beta_p$, p being the number of covariates, we used Gaussian prior distributions $N(0, \sigma_0), \dots, N(0, \sigma_p)$, respectively, where the hyperparameters $\sigma_0, \dots, \sigma_p$ each were assumed to have a Uniform (0,10) distribution. We used a $N(0, 5^2)$ prior distribution for the log alpha parameter. We used zero mean Gaussian prior distributions $N(0, \Sigma_x)$, with Σ_x diagonal with each variance equal to 5^2 for the white noise effects x_k , $k = 1, \dots, n_g$, n_g being the total number of grid points in the state. Sensitivity analyses suggested that the values of the parameters were not affected by the choice of the prior distributions.

5.2 Software

We have implemented our models using Markov chain Monte Carlo^{21,22} via the Metropolis-Hastings algorithm.

We used Julia version 0.2 software,²³ which is a high level programming language for technical computing. It approaches and often matches the performance of the C programming language and much faster than many other available software. We performed the preliminary data analysis using R²⁴ version 3.1.3 and SAS version 9.2 (SAS Institute Inc. Cary, NC, USA).

6 Goodness of fit and convergence diagnosis

We performed model assessment and chose the best fitting model using the DIC, a Bayesian analog of the Akaike's information criterion. Lower values indicate a better fit of the model. The parameters degrees of freedom p_D reflect the model complexity. They were computed from the posterior variance of the deviance, based on the estimator proposed by Gelman et al.²⁵

Let G be the number of samples used for estimation. Let $\bar{D}$ be the average deviance, $\tilde{D}(\theta_g)$ is the deviance computed at the sample parameter value θ_g , $g = 1, \dots, G$. p_D was estimated as half the variance of the deviance

$$\hat{p}_D = \frac{1}{2} \frac{1}{G-1} \sum_{g=1}^G (\hat{D}(\theta_g) - \bar{D})^2 = \frac{1}{2} \widehat{Var}(D) \quad (19)$$

The DIC was computed as the sum of the average deviance $\bar{D}$ and the effective number of parameters

$$DIC = \bar{D} + p_D \quad (20)$$

Lower values of DIC indicate a better fit of the model. Spiegelhalter et al.²¹ suggests that models within 1–2 units of the best model deserve consideration, while models with 3–7 units or more difference in DIC have considerably less support.

For computational reasons, a single chain was run for each model. Convergence was assessed first informally using visual examination of the trace plots. Furthermore, we constructed two chains from the remaining iterations after discarding the burn-in and used Gelman and Rubin's convergence diagnosis.²⁶

We fitted four model variants to the PrCa models, with both independent and dependent space and time, each unadjusted and adjusted for covariates. The goodness of fit results for these models are listed in Table 1.

7 Application

7.1 Prostate cancer

About 1 man in 7 will be diagnosed with PrCa during his lifetime and about 1 man in 38 will die of PrCa.⁹ There is a significant racial difference in PrCa incidence, mortality, and survival, black men being 60% more likely to be diagnosed with PrCa and 2–3 times more likely to die.^{27,28} There is a substantial geographic variation in PrCa incidence and outcomes likely related to several factors, including differences in socioeconomic status, screening behaviors, and access to specialists.^{29–31}

7.2 Data

We examined cancer registry data from the SEER Louisiana registry for the years 2007 through 2010 as they represent the four most recent years of data available and provide a large sample size. SEER registry data has been used previously in the development of spatial survival methods.^{3,32–34} For stage definition, we used SEER historic stage, which was available for our selected years and was defined as having two categories: localized/regional and distant. Localized refers to a cancer confined to the prostate with no penetration of the

capsule. Regional indicates a cancer that involves the regional lymph nodes and/or penetration of the prostatic capsule with or without direct extension beyond the limits of the prostatic capsule into the surrounding organs or tissues. Distant means that the cancer has spread to parts of the body remote from the primary tumor. For PrCa, the SEER staging system combines localized and regional cases into one stage group. We further selected only observations with complete dates available and excluded subjects with survival time of zero. For our time to event outcome, we used the time to death from any causes since the prostate only cancer deaths were too infrequent. The model was adjusted for variables that were available and thought a-priori to be associated with the vital outcome. The individual-level patient covariates that were used in this study include: race (Black versus White and Other), marital status (grouped into married and single/separated/divorced/widowed), stage (Surveillance, Epidemiology, and End Results (SEER) historic stage A grouped into localized/regional versus distant), grade (grouped 1 and 2 versus 3 and 4), and age (age at diagnosis in years). Values labeled “unknown” were considered missing and excluded from the analysis. As county level covariates, we used the mean family income and the number of accredited cancer centers. Data for the county specific covariate “mean family income” has been downloaded from the US Census American Community Survey 2011, five year estimates.³⁵ The number of the Commission on Cancer accredited cancer centers has been obtained from the American College of Surgeons website.³⁶

7.3 Results

Demographic characteristics for the subjects in our data are described in Onicescu et al.⁸ The data used for the analyses consisted of 11943 observations. The range of age of diagnosis was 34–98 years, with the mean being 66. Mean survival time was 22.60 months (sd=13.58), ranging from 1 to 47. The majority of the subjects were white (67%) versus 32.52% black and only 0.48% other races. The majority of the cancer stages were localized/regional (95.94%) with only 4.06% distant. Of all subjects, 92.82% were alive by the end of the study period, while 7.83% died of any cause. Most cancers were Grade 2 (44.65%) and 3 (54.18%) with only 0.87% Grade 1 and 0.30% Grade 4. Due to the small number in grade categories 1 and 4, for estimation we grouped the grade variable and compared grades 3 and 4 versus grades 1 and 2. Regarding marital status, 73.08% of the subjects were married, 12.77% were single and 14.15% were separated, divorced or widowed.

A fixed number of grid points (1000) were generated over the rectangle enclosing the Louisiana polygon area. The total number of gridpoints was chosen in order to have enough grid points in each county but at the same time to keep the computational programming time at a feasible level. Out of these, we further selected the $n_g = 571$ grid points that were inside the study region. The number of grid points in areas ranged from 2 to 31 grid points per county, with a median of 7.

Four models were considered. Model 1 (STSU-AFT unadjusted) is unadjusted for covariates and assumes independent space and time. Model 2 (STSU-AFT adjusted) is adjusted for race, marital status, stage, grade, age at diagnosis, mean family income, and number of accredited cancer centers and assumes independent space and time. Model 3 (STSD-AFT unadjusted) is unadjusted, assumes dependent space and time. Model 4 (STSD-AFT

adjusted) is adjusted for race, marital status, stage, grade, age at diagnosis, mean family income, and number of accredited cancer centers and assumes dependent space and time. Table 1 shows the DIC and the parameters degrees of freedom for the four models considered. All the adjusted models had lower DICs compared to the unadjusted models. The best model with the lowest DIC was Model 4 (STSD-AFT adjusted).

Higher coefficients for the covariates indicate higher survival associated with increased values of the covariate. The posterior means and 95% credible intervals (CI) of the coefficients from our best fitting model are shown in Table 2. After controlling for other covariates in the model, black race versus white or other races was associated with lower survival, with 0.36 decrease (CI=0.5 decrease, 0.22 decrease) in log survival times. Being married was associated with higher survival, with 0.32 increase (CI=0.19 to 0.46) in the log survival times. Distant stage versus localized/regional was associated with lower survival, with 1.76 (CI=(1.97 decrease, 1.55 decrease)) in log survival times. Grades 3 and 4 versus 1 and 2 were associated with lower survival, with 0.19 decrease (95% CI=0.32 decrease, 0.045 decrease). Higher age at diagnosis was associated with lower survival times. Age at diagnosis was standardized for computing simplifications and was used mainly for adjusting purposes.

Figure 1 shows the spatial parish probabilities, with higher probabilities indicating higher risk of death. We notice that there are some parishes with higher risk in the south-eastern part of the state.

Figure 2 shows a map of parish level random effects, which show a relatively random pattern. Since the random effects act multiplicatively on the scale parameter λ of the distribution of times, higher values of random effects indicate lower survival times.

Figure 3 shows the temporal only survival curve $S_{\lambda}(t^*) = 1 - F_{\lambda}(t^*)$ for reference categories (white and other races, single/separated/divorced or widowed, localized/regional stage, grades 1 or 2, mean age of diagnosis and mean family income and no cancer centers) assuming a multiplicative random effect of 1. The survival plot has a slightly curved shaped at earliest times and has an almost linear trend afterwards.

8 Simulation comparison

In order to evaluate the proposed methodology, we carried out a simulation study whereby we made comparisons between the spatially explicit AFT model and a model with a standard contextual spatial random effect structure. For computational simplicity, in simulations we used the STSU-AFT model and we compared it with a random effects model having unstructured and structured random effects.

8.1 Simulation using the STSU-AFT model

We generated 50 datasets. Covariates were simulated using various distributions. We simulated two dichotomous variables $m_i^1 \sim \text{Bernoulli}(0.5)$ and $m_i^2 \sim \text{Bernoulli}(0.5)$ and one continuous normally distributed variable $m_i^3 \sim \text{Normal}(0, 1)$. The variables m_i^4 and m_i^5 were each simulated as continuous spatially correlated variables using a stationary isotropic

covariance model, the corresponding covariance function depending only on the distance between the two points. This is implemented in R in the RMGauss function in the RandomFields package. The variable m_i^4 was generated using the RMGauss function with a mean trend of 0 and variance 4, while the m_i^5 variable was generated with a mean 0, variance 10, and scale parameter 2. Censoring times were simulated from $U(0,100)$.

The times were computed in accordance to the AFT model specification: $\log(y_i) = \mathbf{M}'\boldsymbol{\beta} + \sigma\epsilon_i$, $i = 1, \dots, n$, where $\boldsymbol{\beta} = (\beta_0, \dots, \beta_5)$ is the vector of unknown coefficients and

$\mathbf{M} = (\mathbf{1}, \mathbf{m}_i) = (1, m_i^1, \dots, m_i^5)'$ is the corresponding matrix of covariates including the intercept. For simplicity, we assumed $\sigma=1$. The coefficients $(\beta_0, \dots, \beta_5)$ were set to 1, considered to be the truth value. For simulations, we used the same STSU-AFT model and prior distributions described in the previous sections.

8.2 Simulation using AFT model with correlated and uncorrelated random effects

We used the same simulated data as the one used for the STSU-AFT model. The random effects AFT model is described as follows: Let n be the total number of subjects, n_j be the number of subjects in county j and t_{ij} be the survival time for subject i in county j ,

$\mathbf{M} = (\mathbf{1}, \mathbf{m}_i) = (1, m_i^1, \dots, m_i^5)'$ is the corresponding matrix of covariates including the intercept, c_{ij} =censoring time, $i = 1, \dots, n$ and $j = 1, \dots, 64$, $y_{ij} = \min(c_{ij}, t_{ij})$,

$$\nu_{ij} = \begin{cases} 1 & \text{if } t_{ij} \leq c_{ij} \\ 0 & \text{if } t_{ij} > c_{ij} \end{cases}$$

In order to be able to make the comparison for the random effects model, we used the same data as the one used in the STSU-AFT model simulation. The times were computed in accordance to the AFT random effects model specification as follows: $\log(\lambda_{ij}) = \mathbf{M}'\boldsymbol{\beta} + \mathbf{W}_{j1} + \mathbf{W}_{j2}$, $i = 1, \dots, n$, where $\boldsymbol{\beta} = (\beta_0, \dots, \beta_5)$ is the vector of unknown coefficients, $\mathbf{W}_{j1}$ are the uncorrelated spatial random effects and $\mathbf{W}_{j2}$ are the correlated spatial random effects. We assumed the same prior distributions for the coefficients as the ones used in the STSU-AFT model. For the random effects we have assumed that $\mathbf{W}_{j1}$ follows an independent Normal distribution, $\mathbf{W}_{j1} \sim N(0, \nu_1^2)$, where ν_1^2 denotes the variance of the uncorrelated spatial random effect. For the correlated spatial random effect $\mathbf{W}_{j2}$, we assumed a conditional

autoregressive model. Specifically, $\mathbf{W}_{j2} \sim \text{Normal} \left(\sum_{k \in \delta_j} \frac{\mathbf{w}_{k2}}{n_{\delta_j}}, \frac{\nu_2^2}{n_{\delta_j}} \right)$, where ν_2^2 is the variance and ρ_j is the set of neighbors corresponding to parish j .

The coefficients $(\beta_0, \dots, \beta_5)$ were set to 1, considered to be the truth value.

8.3 Simulation results and comparison

Table 3 shows the simulation results from the STSU-AFT and the random effects models with both correlated and uncorrelated heterogeneity. We display the mean and 95% CI of the estimated coefficients over the 50 iterations, as well as the mean square error for each coefficient, calculated as the sum of the square difference between the estimated coefficient

at each iteration and the truth value, here set to 1. All the coefficients in both models were close to the truth value, suggesting a great fit for both models. All the coefficients except the intercept were closer to the truth value in the STSU-AFT model versus the random effects model. The mean square error values were slightly smaller for the continuous variables in both the STSU-AFT and random effects models, suggesting a better estimate for the continuous versus the categorical variables.

9 Discussion

In this paper, we proposed an extension to the spatially explicit survival model in the context of a Bayesian AFT model by allowing spatial dependency between space and time. Although the space and time survival and related functions are constructed explicitly, we used random effects in order to account for space time dependency due to computational simplicity and feasibility.

We considered and compared both the dependent and non dependent case and our results suggested that the dependent model is preferred, having a lower DIC than the independent case model. It was estimated that black race, distant stage, higher grades, and higher age at diagnosis were associated with lower survival, while being married was associated with higher survival times. We also adjusted for county level covariates, such as county level mean family income and number of cancer centers but these variables did not show an association with the outcome.

Both the dependent and independent models arrived at the same qualitative conclusions about the effect of the covariates. In addition, similar results have been estimated using the Weibull model,⁸ suggesting that the geographically augmented survival model is not highly dependent on the model or distributional assumptions.

With respect to the generality of our proposed methodology, it can be applied for other geographical areas or other diseases. However, the estimated county level risk probabilities are not comparable with other regions, since their values depend on the number of parishes in a region.

Data on the number of cancer centers was currently available for 2013 and therefore a potential limitation of our study is that the actual number of facilities may have differed in the years 2007 through 2010. However, since earning accreditation is based on a rigorous peer review process that is valid for three years usually followed by a reapplication for accreditation, we do not expect the number of accredited cancer centers to largely vary in time.

It is worthwhile pointing out that the model can be extended in a number of ways. In the same AFT context, the model can be used with other distributional assumptions of the survival times. In addition, smoothing terms rather than linear terms could be allowed for covariate modeling. In our model specification, we considered the joint space time pdf as the product between the spatial pdf and the temporal pdf given the space location. Alternatively, the joint space time pdf can be specified as the product between the temporal pdf and the

spatial pdf given a time. Finally, other ways of specifying the dependency structure can be considered.

Acknowledgments

Funding

This work was supported by the National Institutes of Health [R03 Grant No. CA176702-01A1].

References

1. Kalbfleisch, JD., Prentice, RL. The statistical analysis of failure time data. New York, NY: John Wiley; 1980.
2. Henderson R, Shimakura S, Gorst D. Modeling spatial variation in leukaemia survival data. J Am Statist Assoc. 2002; 27:965–972.
3. Banerjee S, Wall MM, Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. Biostatistics. 2003; 4:123–142. [PubMed: 12925334]
4. Li Y, Ryan L. Modeling spatial survival data using semiparametric frailty models. Biometrics. 2002; 58:287–297. [PubMed: 12071401]
5. Hennerfeind A, Brezger A, Fahrmeir L. Geoaddivitive survival models. J Am Statist Assoc. 2006; 101:1065–1075.
6. Zhang J, Lawson AB. Bayesian parametric accelerated failure time spatial model and its application to prostate cancer. J Appl Stat. 2011; 38:591–603. [PubMed: 21475617]
7. Wang S, Zhang J, Lawson AB. A Bayesian normal mixture accelerated failure time spatial model and its application to prostate cancer. Stat Methods Med Res. Epub ahead of print 2012.
8. Onicescu G, Lawson AB, Zhang J, et al. Spatially-explicit survival modeling for small area cancer data. 2015 Manuscript submitted for publication.
9. American Cancer Society. [accessed 15 February 2015] <http://www.cancer.org>
10. National Cancer Institute. [accessed 17 June 2014] <http://seer.cancer.gov/statfacts/html/prost.html>
11. Hsu CE, Mas FS, Miller JA, et al. A spatial-temporal approach to surveillance of prostate cancer disparities in population subgroups. J Natl Med Assoc. 2007; 99:5–7. 72–80.
12. Centers for Disease Control and Prevention. [accessed 17 June 2014] <http://www.cdc.gov/cancer/prostate/statistics/state.htm>
13. Farrow DC, Samet JM, Hunt WC. Regional variation in survival following the diagnosis of cancer. J Clin Epidemiol. 1996; 49:843–847. [PubMed: 8699202]
14. Dickman PW, Gibberd RW, Hakulinen T. Estimating potential savings in cancer deaths by eliminating regional and social class variation in cancer survival in the Nordic countries. J Epidemiol Community Health. 1997; 51:289–298. [PubMed: 9229059]
15. Post PN, Damhuis RA, van der Meyden AP. Variation in survival of patients with prostate cancer in Europe since 1978. EURO CARE Working Group. Eur J Cancer. 1998; 34:2226–2231. [PubMed: 10070291]
16. Yu XQ, O'Connell DL, Gibberd RW, et al. Estimating regional variation in cancer survival: a tool for improving cancer care. Cancer Causes Control. 2004; 15:611–618. [PubMed: 15280640]
17. Higdon, D. Space and space-time modeling using process convolutions. In: Anderson, C. Barnett, V. Chatwin, P., El-Shaarawi, A., editors. Quantitative methods for current environmental issues. London: Springer; 2002. p. 37–56.
18. Calder CA. A dynamic process convolution approach to modeling ambient PM_{2.5} and PM₁₀ concentration levels. Environmetrics. 2007; 19:39–48.
19. Calder, CA., Holloman, C., Higdon, D. Exploring spacetime structure in ozone concentration using a dynamic process convolution model. In: Gatsonis, C. Kass, RE. Carriquiry, A., et al., editors. Case Studies in Bayesian Biostatistics. Vol. 6. New York: Springer-Verlag; 2002. p. 165–176.
20. Calder, CA., Cressie, N. Some topics in convolution-based spatial modeling. Proceedings of the 56th Session of the International Statistics Institute; August 2007;

21. Spiegelhalter DJ, Best NG, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc Ser B*. 2002; 64:583–639.
22. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*. 1970; 57:97–109.
23. Bezanson, J., Karpinski, S., Shah, VB., et al. [accessed 17 June 2015] Julia: A Fast Dynamic Language for Technical Computing. 2012. CoRR abs/1209.5145. <http://arxiv.org/abs/1209.5145>
24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2015. <http://www.R-project.org/>
25. Gelman, A., Carlin, JB., Stern, HS., et al. Bayesian data analysis. London: CRC Press; 2004.
26. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statist Sci*. 1992; 7:457–511.
27. Wu I. Disparities in prostate cancer in African American men: What primary care physicians can do. *Cleve Clin J Med*. 2012; 79:313–20. [PubMed: 22550073]
28. Siegel R, Ma J, Zou Z, et al. Cancer statistics, 2014. *CA Cancer J Clin*. 2014; 64:9–29. [PubMed: 24399786]
29. Onega T, Duell EJ, Shi X, et al. Influence of place of residence in access to specialized cancer care for African Americans. *J Rural Health*. 2010; 26:12–19. [PubMed: 20105263]
30. Jemal A, Ward E, Wu X, et al. Geographic patterns of prostate cancer mortality and variations in access to medical care in the United States. *Cancer Epidemiol Biomarkers Prev*. 2005; 14:590–595. [PubMed: 15767335]
31. Gregorio DI, Kulldorff M, Sheehan TJ, et al. Geographic distribution of prostate cancer incidence in the era of PSA testing, Connecticut, 1984 to 1998. *Urology*. 2004; 63:78–82. [PubMed: 14751353]
32. Diva U, Banerjee S, Dey DK. Modelling spatially correlated survival data for individuals with multiple cancers. *Stat Modell*. 2007; 7:191–213.
33. Carlin, BP., Banerjee, S. Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In: Bernardo, JB., Dawid, AP., Berger, JO., West, M., editors. *Bayesian Statistics*. New York, NY: Oxford University Press; 2003. p. 7
34. Banerjee S, Dey DK. Semiparametric proportional odds models for spatially correlated survival data. *Lifetime Data Anal*. 2005; 11:175–191. [PubMed: 15938545]
35. American Community Survey. [accessed 30 December 2013] 2013. <http://www.census.gov/acs/>
36. American College of Surgeons. [accessed 30 December 2013] 2013. <http://www.facs.org/>

Figure 1.

Map of model based estimated spatial risk probabilities from the STSD-AFT adjusted for race, marital status, stage, grade, age at diagnosis, mean family income, and number of accredited cancer centers.

Figure 2.

Map of model based estimated random effects from the STSD-AFT adjusted for race, marital status, stage, grade, age at diagnosis, mean family income, and number of accredited cancer centers.

Figure 3.

Estimated model based temporal only survival curves $S_i(t^*) = 1 - F_i(t^*)$ for a county with random effect value $g(A_S)=1$.

Table 1

Goodness of fit measures.

	DIC	p_D
Model 1 (STSU-AFT unadjusted ^a)	124647.6	8.11
Model 2 (STSU-AFT adjusted ^b)	124027.0	20.49
Model 3 (STSD-AFT unadjusted ^c)	124747.1	8.80
Model 4 (STSD-AFT adjusted ^d)	123826.5	16.80

^aModel 1 is the spatial temporal survival unstructured AFT model, unadjusted.

^bModel 2 is the spatial temporal survival unstructured AFT model, adjusted for race, marital status, stage, grade, age at diagnosis, mean family income and number of accredited cancer centers.

^cModel 3 is the spatial temporal survival AFT model with dependent space and time, unadjusted.

^dModel 4 is the spatial temporal survival AFT model with dependent space and time, adjusted for race, marital status, stage, grade, age at diagnosis, mean family income and number of accredited cancer centers.

Table 2

Parameter estimates and 95% credible intervals (CI), AFT-STSD adjusted model.

	Estimate	95% CI
Intercept	5.67	(5.46, 5.89) ^a
Race		
White and Other races	Ref.	
Blacks	-0.36	(-0.50, -0.22) ^a
Marital status		
Single/Separated/Divorced/Widowed	Ref.	
Married	0.32	(0.19, 0.46) ^a
Stage		
Localized/regional	Ref.	
Distant	-1.76	(-1.97, -1.55) ^a
Grade		
1 and 2	Ref.	
3 and 4	-0.19	(-0.32, -0.045) ^a
Age at diagnosis (standardized)	-0.61	(-0.68, -0.54) ^a
Mean family income (standardized)	-0.0069	(-0.070, 0.057)
No. cancer centers	-0.00025	(-0.021, 0.029)

^aThe credible interval does not include zero.

Table 3

Mean coefficients estimates obtained from simulated data when truth coefficients are 1.

	STSU-AFT ^a	UH+CH random effects ^b	STSU-AFT ^a	UH+CH random effects ^b
	Mean (95% CI)	Mean (95% CI)	MSE	MSE
β_0	0.9700 (0.8271, 1.1141)	1.0097 (0.8671, 1.1528)	0.2579	0.2179
β_1	1.0100 (0.8647, 1.1543)	0.9934 (0.8657, 1.1219)	0.1746	0.1669
β_2	1.0117 (0.8672, 1.1568)	0.9877 (0.8582, 1.1192)	0.3323	0.3169
β_3	1.0012 (0.9255, 1.0770)	0.9866 (0.9148, 1.0564)	0.0767	0.0838
β_4	1.0074 (0.9470, 1.0683)&	0.9847 (0.9107, 1.0608)	0.0445	0.0544
β_5	0.9947 (0.9292, 1.0605)&	0.9896 (0.9164, 1.0602)	0.0391	0.0524

^aSTSU-AFT is the Spatial Temporal Survival Uncorrelated AFT model.

^bUH+CH random effects is the contextual AFT model with unstructured (UH) and spatially correlated (CH) random effects.