Published in final edited form as:

Spat Spatiotemporal Epidemiol. 2019 June ; 29: 139–148. doi:10.1016/j.sste.2018.06.001.

Spatially-explicit Survival Modeling with Discrete Grouping of Cancer Predictors

Georgiana Onicescu*,

Western Michigan University, Department of Statistics, Kalamazoo, MI

Andrew B Lawson,

Medical University of South Carolina, Department of Public Health Sciences, Charleston, SC

Jiajia Zhang,

Arnold School of Public Health, Medical University of South Carolina, Columbia, SC, USA

Mulugeta Gebregziabher,

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

Kristin Wallace, and

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

Jan M. Eberth

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

Abstract

In this paper the spatially explicit survival model is extended by allowing the relation with the explanatory covariates to be spatially adaptive using a threshold conditional autoregressive (CAR) model, further extended to allow the inclusion of multiple threshold levels. The model is applied to prostate cancer survival based on Louisiana SEER registry, which holds individual records linked to vital outcomes and is geocoded at the parish level.

Keywords

bayesian; latent; spatial; prostate cancer

1. Introduction

The use of spatial referencing in cancer studies is gaining prominence due to the increasing availability of spatial data and analytical tools, increased computing power, and the growing

*Corresponding author: georgiana.onicescu@wmich.edu (Georgiana Onicescu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

understanding of the importance of the linkage between geographical regions and health. An important characteristic of spatial data is the tendency of the relationship between covariates and the chosen outcome to vary by geographical location.

Prostate cancer (PrCA) incidence has increased during recent years, possibly linked to environmental exposures. Due to the uneven geographical distribution of environmental carcinogens that may be linked with PrCA, there are variations in PrCA disease occurrence that is detectable in a spatial analysis [1]. Several studies have found large spatial variation in PrCA survival in Louisiana, making this type of cancer suitable for implementation of spatial analysis approaches [2].

An hierarchical Bayesian approach for geographically augmented spatial time to event data have been developed by explicitly modeling the spatial dependency using direct derivations of the density, survival and hazard functions [3] and the model has been applied to the Surveillance, Epidemiology and End Results (SEER) prostate cancer data. An extension of this model using accelerated failure time and spatially dependent space time distributions has also been proposed [4]. Simulation studies showed that these models are comparable and even outperformed the most commonly used spatial random effects models. In addition, one of the main advantage of the spatially explicit model is the increased interpretability of the risk estimates, which range between 0 and 1, versus the random effects, which have limited interpretability due to their unlimited range. Results from the spatially explicit models assuming a Weibull distribution for the survival times [4] do not differ qualitatively from the models using accelerated failure time assumption [3], suggesting robustness of these models under various distributional assumptions.

There has been an increased interest in trying to detect these local patterns within global trends. Geographically weighted regression (GWR) [5] can be used to model relationships that exhibit spatial nonstationarity, assuming that the regression coefficients have a continuous spatial prior distribution. However, it can also be the case that only a small number of coefficients are needed to describe spatially the relationship between the covariate with the outcome. Therefore, it is natural to allow the possibility of a discrete spatial grouping as a response to a latent spatial process. In this context, a spatially structured discrete distribution can be used to address the discrete clustering of coefficients [6].

Specifically, in a two level threshold conditional autoregressive CAR model [6], the values of a spatial process are dichotomized into two levels. This is accomplished by using a CAR process and, for a specific parish of interest, creating an indicator variable signaling whether the CAR component of that parish is greater than or equal to the threshold value. Each variable in the model has coefficients for each threshold grouping. Therefore, the association of a variable with the outcome is spatially dependent. In this paper, the two level threshold CAR is extended to allow for multiple threshold levels in the context of the spatially explicit Spatio-Temporal Survival Unstructured Weibull (STSU- Weibull) model developed by authors [4]. We apply the model using the SEER data [7] using the Louisiana registry during the years 2007–2010.

2. Statistical model

We develop our model using the spatially extended probability distribution function, which is allowed to have both a temporal and spatial component. For the probability density function (pdf), space is defined on the real plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, while the temporal component is defined on the real line $\mathbb{R}$. Hence $f: (\mathbb{R} \times \mathbb{R}) \times \mathbb{R} \rightarrow (0, \infty)$, $f(s, t) = f_1(s) f_2(t|s)$, where $f_1(s)$ is the space density functions, $s = (s_1, s_2)$ being a location on the surface area defined by latitude and longitude and $f_2(t|s)$ is the temporal density function conditional on the spatial threshold grouping. The corresponding cumulative distribution functions, survival and hazard functions can be derived for the temporal component only, spatial component only and for the space and time components combined as described in [4] and [3]. In our notation, the function indices “*time*”, “*space*” and “*space, time*” refers to the domain of integration: temporal, spatial and spatio-temporal respectively.

Let A be a spatial area, $s \in A$ and let $t^* \in (0, \infty)$ be a duration time. We define

$$F_{time}(t^*) = \int_0^{t^*} f_2(t) dt, \quad (1)$$

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

$$F_{space}(A) = \int_A f_1(s) ds, \quad (2)$$

where $F_{space}(A)$ is the marginal probability over the space area A ,

$$F_{time}(s, t^*) = f_1(s) \int_0^{t^*} f_2(t | s) dt \approx f_1(s) \int_0^{t^*} f_2(t | G_q, s \in G_q) dt, \quad (3)$$

where $F_{time}(s, t^*)$ is the probability that the time random variable takes a value less or equal to t^* at the spatial location s , and G_q is the spatial grouping corresponding to location s .

$$\begin{aligned}
F_{space}(A, t^*) &= \int_A f(s, t) ds \\
&= \int_A f_1(s) f_2(t^* | s) ds \\
&\approx \int_A f_1(s) f_2(t^* | G_q, s \in G_q) ds \\
&\approx f_2(t^* | G_q, s \in G_q) \int_A f_1(s) ds,
\end{aligned} \tag{4}$$

where $F_{space}(A, t^*)$ is the temporal probability density function at the value t^* over the domain A ,

$$\begin{aligned}
F_{space, time}(A, t^*) &= \int_A \int_0^{t^*} f(s, t) dt ds \\
&= \int_A \int_0^{t^*} f_1(s) f_2(t | s) dt ds \\
&\approx \int_A \int_0^{t^*} f_1(s) f_2(t | G_q, s \in G_q) dt ds \\
&\approx \left(\int_A f_1(s) ds \right) \left(\int_0^{t^*} f_2(t | A, s \in A, A \in G_q) dt \right) \\
&= F_{space}(A) F_{time}(t^* | A),
\end{aligned} \tag{5}$$

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

The corresponding survival functions are defined as follows:

$$S_{time}(s, t^*) = 1 - F_{time}(s, t^*), \tag{6}$$

$$S_{space, time}(A, t^*) = P(s \in A, t > t^*) \approx F_{space}(A)(1 - F_{time}(t^*)), \tag{7}$$

where $S_{time}(s, t^*)$ is the cumulative probability of surviving beyond time t^* for a person residing in location s , $S_{space, time}(A, t^*)$ is the joint cumulative probability of surviving beyond time over the area A . The temporal component is allowed to depend on the spatial threshold grouping.

2.1. Threshold CAR STSU-Weibull

The assumption of similar effects of the covariates over the space region does not always hold true. By including a threshold CAR model, we allow the space to be an effect modifier.

While the GWR assumes a continuous space effect modifier which is not easily interpretable and computationally feasible, in a threshold CAR process we allow the space effect modification to be accounted for using a small number or set of coefficients that are relevant to a given covariate, as a response to a latent continuous process.

The format of the data consists of a time to event variable for each individual, the parish of residence, and individual level covariates. Let n =number of subjects, $t_i, i = 1, \dots, n$, independent, identically distributed survival times with density function $f(s, t)$, $\mathbf{m}_i = (m_i^1, \dots, m_i^p)'$, $i = 1, \dots, n$ a vector of covariates for subject i , p =number of covariates, c_i = censoring time, $i = 1, \dots, n$,

$$y_i = \min(c_i, t_i).$$

We denote by ν_i the death indicator defined as follows:

$$\nu_i = \begin{cases} 1 & \text{if } t_i \leq c_i \\ 0 & \text{if } t_i > c_i. \end{cases}$$

Let $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ and $\boldsymbol{\nu} = (\nu_1, \nu_2, \dots, \nu_n)'$. Let $\mathbf{P}$ be the vector of parameters of interest. Let $A_1, \dots, A_L$ be a partition of the study region and n_l be the number of people in area A_l . The area likelihood is defined as follows:

$$\begin{aligned} L_{Area}(\mathbf{P} \mid n, \mathbf{y}, \boldsymbol{\nu}, \mathbf{m}_i) &= \prod_{l=1}^L \prod_{i=1}^{n_l} [F_{space}(A_l, y_i, \mathbf{m}_i)]^{\nu_i} [S_{space, time}(A_l, y_i, \mathbf{m}_i)]^{1-\nu_i} \quad (8) \\ &= \prod_{l=1}^L \prod_{i=1}^{n_l} [f_2(y_i, \mathbf{m}_i) \int_{A_l} f_1(s) ds]^{\nu_i} [F_{space}(A_l, \mathbf{m}_i)(1 - F_{time}(y_i, \mathbf{m}_i))]^{1-\nu_i} \\ &= \prod_{l=1}^L \prod_{i=1}^{n_l} [f_2(y_i, \mathbf{m}_i) \int_{A_l} f_1(s) ds]^{\nu_i} \left[\int_{A_l} f_1(s) ds (1 - \int_0^{y_i} f_2(t, \mathbf{m}_i) dt) \right]^{1-\nu_i}, \end{aligned}$$

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

The probability distribution function for the temporal random variable T for the i^{th} observation, $i = 1, \dots, n$, follows the Weibull(μ, λ_i) distribution. The Weibull distribution is widely used for time to event data due to its flexibility in modeling either a decreasing, constant or increasing hazard rate, if its shape parameter μ is less, equal or greater than 1 respectively. One of the reasons for the popularity of the Weibull distribution is that it includes other useful distributions as special cases or close approximations. For example, when $\mu=1$, the Weibull distribution becomes the exponential distribution and the hazard rate remains constant as time increases. In addition, the Weibull distribution allows the inclusion

of covariates, usually linked to its scale parameter. The scale parameter λ_i of the Weibull distribution is known as the characteristic life and can be modeled in terms of covariates. Let $A_1, \dots, A_L$ be a spatial partition within the study region. Using the Weibull temporal density and assuming independent space and time, the area likelihood is defined as follows:

$$L_{Area}(\mathbf{P} \mid n, \mathbf{y}, \nu) \quad (9)$$

$$= \prod_{l=1}^L \prod_{i=1}^{n_l} [\mu \lambda_i (y_i)^{\mu-1} e^{-(y_i)^{\mu} \lambda_i} \int_{A_l} f_1(s) ds]^{\nu_i} \left[\int_{A_l} f_1(s) ds (1 - \int_0^{y_i} \mu \lambda_i t^{\mu-1} e^{-t^{\mu} \lambda_i} dt) \right]^{1-\nu_i}.$$

The covariates were linked to the $\log(\lambda_i)$ parameter. Each covariate was considered to have a fixed number of groups G and therefore $G-1$ thresholds. The choice for the fixed number of groups is important and in some situations it needs to be estimated. [6] used $G=2$ but this was restrictive since models with higher number of groups can have better fit. In this work we compared models with various number of groups and for computational simplicity we considered each covariate to have the same threshold. The intercept was also allowed to be spatially dependent.

$$\log(\lambda_i) = \beta_{0,g}^{i,j} + \beta_{1,g}^{i,j} m_i^1 + \dots + \beta_{p,g}^{i,j} m_i^p$$

where $M=(1, m^1, \dots, m^p)$ is a matrix of covariates including the intercept term, $m^1, \dots, m^p$ are the p covariates and the coefficients are defined as follows for $G=4$:

$$\beta_{h,g}^{i,j} = \begin{cases} \beta_{h1} & \text{if subject } i \in \text{parish } j, \text{ parish } j \in \text{group } 1 \\ \beta_{h2} & \text{if subject } i \in \text{parish } j, \text{ parish } j \in \text{group } 2 \\ \beta_{h3} & \text{if subject } i \in \text{parish } j, \text{ parish } j \in \text{group } 3 \\ \beta_{h4} & \text{if subject } i \in \text{parish } j, \text{ parish } j \in \text{group } 4 \end{cases}, h = 1, \dots, p.$$

The thresholds that define the four groups are the quantiles $q_1, \dots, q_{G-1}$ a continuous spatially referenced process u updated within each step of the Metropolis Hastings (MH) algorithm. In order to establish to which group a parish belongs to, a group indicator vector is also computed at each step of the MH algorithm, $I_j, j=1, \dots, n_p$, n_p being the number of parishes, defined as follows for a model with $G=4$ groups:

$$I_j = \begin{cases} 1 & \text{if } u_j < q_1 \\ 2 & \text{if } q_1 \leq u_j < q_2 \\ 3 & \text{if } q_2 \leq u_j < q_3 \\ 4 & \text{if } u_j \geq q_3 \end{cases}, j = 1, \dots, n_p$$

2.2. Spatial model

Modeling the spatial dependence structure is of fundamental importance to all spatially referenced data. There are a wide variety of choices that can be used to specify the spatial model. One approach is to assume a geostatistical model whereby the spatial component is assumed to follow a Gaussian process with spatial dependence defined by a covariance function, usually assumed to be second order stationary, with the covariance between any two locations depending on the distance between them [8]. Spatial modeling using the covariance function is computationally restrictive due to the necessity for inversion of a possibly large dimension positive definite matrix.

An alternative to directly specifying the covariance function is to assume a process convolution model [9], 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 $k(s)$. The advantage of convolution based models lies in their computational simplification. In addition, they always induce valid covariance functions and, due to their nonparametric nature, have considerable flexibility versus a fully parametric approach.

Kernel convolutions have been widely used in the spatial literature [10, 11]. Special cases such as Gaussian component mixture (GSM) have also been proposed [12]. In order to construct the spatial model, a process convolution was used, with a Gaussian kernel function allowed to vary over the study area. As described in [9], the model for the spatial process is determined by specifying the white noise process $x(s)$ and the smoothing kernel $k(s)$. The Gaussian kernel $k(s) \propto \exp(-\frac{1}{2}\|s\|^2)$, ($\|s\|$ being the Euclidian norm) was chosen since it induces a covariance matrix which is a function of the squared distance and gradually decreases with increased distance. Specifically, denoting $d = s - s'$ the displacement vector between the locations s and s' , the Gaussian kernel induces the covariance $c(d) = \text{Cov}(z(s), z(s')) \propto \exp(-\frac{1}{2}\|\frac{d}{\sqrt{2}}\|^2)$, [9].

The Gaussian process $z(s)$ can be constructed over a spatial region $\mathbb{S}$ as follows [9, 13]:

$$z(s) = \int_{\mathbb{S}} k(v - s)x(v)dv, \quad (10)$$

for $s \in \mathbb{S}$. Since the above integration cannot be explicitly solved, the integral can be approximated by a finite sum:

$$z(s) = \sum_{j=1}^{n_g} x_j k(w_j - s), \quad x_j \sim N(0, \sigma^2), \quad (11)$$

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 [9]. In order to have positiveness of the white noise process, the following formulas was used:

$$z(s) = \sum_{j=1}^{n_g} e^{x_j k(w_j - s)}, x_j \sim N(0, \sigma^2), \quad (12)$$

$$z_A = \int_A z(s) ds \approx \sum_{s \in A} \sum_{j=1}^{n_g} e^{x_j k(w_j - s)}. \quad (13)$$

Since only the number of all cause deaths and total population of interest (PrCA cases) were available in each parish, the coordinates of the centroid of each parish was used and the above formula was approximated with the following:

$$z_A \approx p_A \sum_{j=1}^{n_g} e^{x_j k(w_j - c_A)}, \quad (14)$$

where c_A is the centroid of parish A and p_A is the percentage of deaths in area A, computed as the ratio of the number of all cause deaths among the PrCA cases and the total number of PrCA cases, using the SEER data.

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 area. These values were stored in the kernel matrix K_{lj} , $j = 1, \dots, L$ and $j = 1, \dots, n_g$ of dimension $L \times n_g$. z_{A_l} was therefore calculated by multiplying the number of deaths for parish A with the scalar product between the l^{th} row of the kernel matrix and the exponentiated white noise vector $(e^{x_1}, \dots, e^{x_{n_g}})'$.

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 parish but at the same time to keep the computational programming time at a feasible level. Out of the total grid points, 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.

The integral over an area of the spatial pdf function is defined as the ratio between z_A and z_{A_T} , which is the risk estimated in area A. Higher values for the risk estimates indicate a higher risk of death in a particular parish, compared to the entire region of interest. Since the risk estimates are ratios between the spatial process in a parish and the spatial process in the entire region, a value of 0 means that there is no risk of death in the parish, all the risk being distributed in the remaining parishes. An estimated risk of 1 means that the risk of death in a

particular region is the same as the risk in the entire region, with no risk being distributed to the other parishes.

$$z_{A_T} = \sum_{l=1}^L z_{A_l} \frac{z_{A_l}}{z_{A_T}} = \int_A f_1(s) ds, \quad (15)$$

L being the total number of partitions of the study region. Therefore the area likelihood becomes:

$$L_A(\mu, \lambda \mid n, \mathbf{y}, \nu) = \prod_{l=1}^L \prod_{i=1}^{n_l} [\mu \lambda_i (y_i)^\mu - 1] e^{-(y_i)^\mu \lambda_i \frac{z_{A_l}}{z_{A_T}}} \left[\frac{z_{A_l}}{z_{A_T}} (1 - \int_0^{y_i} \mu \lambda_i t^{\mu-1} e^{-t^\mu \lambda_i} dt) \right]^{1-\nu_i}. \quad (16)$$

3. Prior distributions

Since the models are implemented in the Bayesian framework, prior distributions were assigned for each parameter. Let G be the number of spatial groups and p the number of covariates. The intercept and each covariate are allowed to have different levels of coefficients corresponding to each group level. For each of the coefficients $\beta_{0r}, \dots, \beta_{pr}, r = 1 \dots G$, Gaussian prior distributions $N(0, \sigma_{0r}^2), \dots, N(0, \sigma_{pr}^2)$ were used respectively, where the hyperparameters $\sigma_{0r} \dots \sigma_{pr}$ each were assumed to have a Uniform (0,10) distribution. The correlated spatial component u was assumed to have a conditional autoregressive (CAR)

model. Specifically, conditionally, $u_j \mid u_{-j} \sim N(\sum_{k \in \rho_j} \frac{u_k}{n_{\rho_j}}, \frac{\psi^2}{n_{\rho_j}})$, where ψ^2 is the variance, ρ_j is

the set of neighbors corresponding to parish j and u_{-j} is the $(n-1)$ dimensional vector without the j^{th} coordinate of u . A Uniform(-0.01,0.01) prior distribution was assumed for the logarithm of the standard deviation $\log(\psi)$. Visual displays of the trace plots did not show that the values of the parameters were constrained above or below by the choices of the prior distribution parameters, suggesting that the posterior parameter values were not sensitive to these choices.

4. Computational approach and software

The models were implemented using Markov chain Monte Carlo (MCMC) [14, 15] via the Metropolis-Hastings algorithm [16]. Julia version 0.2 software [17] was used for the MCMC algorithm, which is a high level programming language for technical computing. Preliminary data analysis was performed using R [18] and SAS version 9.4 software. SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

In summary, the model is specified as follows:

$$L_{Area}(\mu, \lambda \mid n, \mathbf{y}, \nu) \quad (17)$$

$$= \prod_{l=1}^L \prod_{i=1}^{n_l} [\mu \lambda_i (y_i)^{\mu-1} e^{-(y_i)^\mu \lambda_i \frac{z_{A_l}}{z_{A_T}}}]^{\nu_i} \left[\frac{z_{A_l}}{z_{A_T}} (1 - \int_0^{y_i} \mu \lambda_i t^{\mu-1} e^{-t^\mu \lambda_i} dt) \right]^{1-\nu_i}.$$

$$z_A \approx p_A \sum_{j=1}^{n_g} e^{x_j k(w_j - c_A)}, z_{A_T} = \sum_{l=1}^L z_{A_l}, \quad (18)$$

where p_A is the percentage of deaths in area A, w_j are the grid points, c_A is the centroid of parish A, n_g being the total number of gridpoints,

$k(s) \propto \exp -\frac{1}{2}\|s\|^2$, ($\|s\|$ being the Euclidian norm) is the Gaussian kernel,

$$\log(\lambda_i) = \beta_{0,g}^{i,j} + \beta_{1,g}^{i,j} m_i^1 + \dots + \beta_{p,g}^{i,j} m_i^p$$

where $M=(1, m^1, \dots, m^p)$ is a matrix of covariates including the intercept term, $m^1, \dots, m^p$ are the p covariates and the coefficients are defined as follows for $G=4$:

$$\beta_{h,g}^{i,j} = \begin{cases} \beta_{h1} & \text{if subject } i \in \text{parish } j, \text{ parish } j \in \text{group } 1 \\ \beta_{h2} & \text{if subject } i \in \text{parish } j, \text{ parish } j \in \text{group } 2 \\ \beta_{h3} & \text{if subject } i \in \text{parish } j, \text{ parish } j \in \text{group } 3 \\ \beta_{h4} & \text{if subject } i \in \text{parish } j, \text{ parish } j \in \text{group } 4 \end{cases}, h = 1, \dots, p.$$

$$\beta_{0r} \sim N(0, \sigma_{0r}^2), \dots, \beta_{pr} \sim N(0, \sigma_{pr}^2), r = 1, \dots, 4$$

$$\sigma_{0r} \dots \sigma_{pr} \sim \text{Uniform}(0, 10),$$

$$u_j \mid u_{-j} \sim N\left(\sum_{k \in \rho_j} \frac{u_k}{n \rho_j}, \frac{\psi^2}{n \rho_j}\right), \text{ where } \psi^2 \text{ is the variance, } \rho_j \text{ is the set of neighbors corresponding}$$

to parish j , and $\mathbf{u}_{-j}$ is the $(n-1)$ dimensional vector without the j^{th} coordinate of $\mathbf{u}$,

$$\log(\psi) \sim \text{Uniform}(-0.01, 0.01)$$

$$x_k \sim N(0, 5^2), k = 1, \dots, n_g,$$

$$\log(\mu) \sim N(0, 0.1^2).$$

5. Model selection

The model fit was assessed using the DIC [14], a Bayesian generalization of the Akaike Information criteria (AIC). We have used the DIC as a measure of model comparisons due to computational feasibility. WAIC is an alternative model comparison measure, however, a cost of using WAIC is that it relies on a partition of the data into n pieces, which is not so easy to do in some structured-data settings such as time series, spatial, and network data (Gelman et al., 2014). DIC do not make this partition explicitly and therefore was feasible to implement in our already complex model. Cross-validation is like WAIC in that it requires data to be divided into disjoint, ideally conditionally independent, pieces and can be computationally expensive. WAIC has been shown to give results similar to the DIC (Gelman et al., 2014).

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. [19, 20].

Let Q be the number of samples used for estimation. Let $\bar{D}$ be the average deviance, $\hat{D}(\theta_q)$ is the deviance computed at the sample parameter value θ_q , $q = 1, \dots, Q$.

p_D was estimated as half the variance of the deviance:

$$\widehat{p_D} = \frac{1}{2} \frac{1}{Q-1} \sum_{q=1}^Q (\hat{D}(\theta_q) - \bar{D})^2 = \frac{1}{2} \widehat{\text{Var}}(D). \quad (19)$$

Compared to other proposals for estimating the parameters' degrees of freedom, this measure has the advantage of always being positive and it also gives the correct estimate for large sample sizes [20]. The DIC was computed as the sum of the average deviance $\bar{D}$ and the effective number of parameters:

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

Lower values of DIC indicate a better fit of the model. [14] 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 simplicity, a single chain was run for each model. Convergence was assessed first informally using visual examination of the trace plots. Furthermore, two chains were constructed from the remaining

iterations after discarding the burn-in and used Gelman and Rubin's convergence diagnosis [21] to formally assess convergence.

6. Application

6.1. Prostate cancer

Prostate cancer (PrCa) is a significant challenge to public health, being the second leading cause of cancer death in American men, behind only lung cancer. It is estimated that 164,690 new cases of prostate cancer will be diagnosed and 29,430 will die of prostate cancer in 2018 [22]. The average annual prostate cancer incidence rate between 2010 and 2014 was 114.9 per 100,000, age adjusted to the 2000 US standard population. The average annual prostate cancer death rate between 2011 and 2015 was 19.5 per 100,000, age adjusted to the 2000 US standard population [22]. As with most cancer, there is no single risk factor identified to be responsible for PrCa occurrence. Risk factors include but are not limited to age, race, family history/genetics and geographical location [22]. There are large geographical variations in the survival rate of prostate cancer [2]. Geographic studies of prostate cancer data can provide useful information for disease control and prevention, by identifying regions with high death risk in the prostate cancer population that may need public health interventions.

6.2. Data

Cancer registry data from the SEER Louisiana registry was used, selecting the years 2007 through 2010. We started with year 2007 in order to avoid possible population movement and recording issues related to Hurricane Katrina. SEER registry data has been used previously in the development of spatial survival methods [23, 24, 25, 26] and has been used for the development of spatially explicit survival modeling for small area cancer data [4]. The data included prostate cancer cases. For stage of diagnosis, SEER historic stage was used, which was available for the 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.

The time to death from any causes was used as the outcome of interest since the prostate only cancer deaths were too infrequent. The number of any cause deaths in our prostate cancer patients data in Louisiana parishes ranged from 0 to 81, with a median of 5. The model was adjusted for variables that were available and thought a-priori to be associated with the vital outcome. The individual-level patient data that were used in this study include: race (African-American versus Whites and Other), marital status at diagnosis (grouped into married (including both married with certificate or domestic partner of same sex or opposite sex or unregistered) versus single and separated/divorced/widowed), stage at diagnosis (SEER historic stage A grouped into localized/regional versus distant), grade (grouped into grades 1 and 2 versus 3 and 4) and age (age at diagnosis in years). For computational

efficiency, the variable age was standardized by subtracting the mean and dividing by its standard deviation. Values labeled "unknown" were considered missing and excluded from the analysis in the adjusted models. The data selection process is summarized as follows: 1) We downloaded the male genital cancers with 336795 observations; 2) We selected only the Louisiana registry, the resulting data contained 37555 observations; 3) we selected only years 2007–2010, the resulting data had 15135 observations; 4) we selected only PrCA, the resulting data contained 14622 observations; 5) we excluded one subject with in situ historic stage, the resulting data had 14621 observations; We excluded the observations with survival time zero, the resulting data has 14184 observations; 6) We selected only the observations with complete dates available, including days (for more information see SEER description on the calculation of the survival times [27]), the resulting data had 13835 observations; 7) We excluded observations with missing values for the covariates of interest, the final resulting data had 11943 observations.

6.3. Results

Demographic characteristics of the subjects are as follows: The range of age of diagnosis was 34 to 98 years, with the mean being 66. Mean time to death or censoring was 22.60 months ($sd=13.58$), ranging from 1 to 47. The majority of the subjects were white (67%) while 32.52% were 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 the grade variable was grouped into grades 3 and 4 versus grades 1 and 2 categories. Regarding marital status, 73.08% of the subjects were married, 12.77% were single and 14.15% were separated, divorced or widowed.

Figure 1 displays the percentage of all cause deaths in our PrCA population.

6.3.1. Model selection and goodness of fit assessment—The STSU-Weibull threshold CAR model is adjusted for the following covariates: race, marital status, stage, grade and age at diagnosis. In table 1, the DIC and p_D values are displayed. The results show that the threshold CAR model with three thresholds had the lowest DIC. The model was run for 40000 iterations, the first 28000 being discarded as burn-in.

6.3.2. Estimation—Table 2 displays the posterior means and 95% credible intervals (CI) of the coefficients, higher coefficients indicating lower survival times. Black race, distant stage and higher age at diagnosis were associated with lower survival in all groups categories, while being married or in a domestic relationship was associated with higher survival. Higher grade (3 and 4 versus 1 and 2) was associated with lower survival only in the third and fourth groups.

Figure 2 displays the spatial probabilities $\frac{z_A}{Z_{A_T}}$, with higher probabilities indicating higher risk of death. Some parishes with higher risk are in the South and North-Western part of the state, while parishes in the North-East have lower risk.

Figure 3 displays the map of the estimated group levels, computed as the most frequently occurring value. Most counties in the first group are located in the Northern part of the state, counties in the second group are in the central part of the state, counties in the third group are located in the South-Central part of the state, while counties in the fourth group can be found in the South-Eastern part of the state.

6.3.3. Discussion—In this work, a Bayesian methodology was constructed for directly modeling the spatial dependency in the specification of the survival, density and hazard functions and allowing the relationship between the predictors and outcome to vary spatially with an extended threshold CAR model to define the spatial grouping. The model was applied to analyze the SEER PrCa data in Louisiana during the years 2007 through 2010. In our model, certain areas are grouped and have different regression parameters versus others, therefore space being regarded as an effect modifier. An alternative approach would be the use of a Geographically Weighted Regression (GWR) model, which is a random slopes model with different coefficients in each area. However, this approach is not always computationally feasible and parsimonious due to the large number of parishes.

In our study the geographical locations were supplied as parishes rather than exact latitude and longitude addresses, which resulted in some limitations. Although point process data give the closest link to the conceptual framework, such data are rarely available routinely due to confidentiality reasons. As is the case with the majority of the disease mapping models, using aggregated data versus point process data may introduce some ecological bias which is not easily quantified due to the absence of point process data for comparison. Also, although the use of threshold CAR process is an approximation of a latent continuous process, it still allows the modeling of space effect modification, which standard spatial models often ignore. Our estimated all death risk probabilities are computed as ratios between the spatial process in a parish and the spatial process in the region, as described in the spatial model section. Therefore, these values can be only used for comparisons to one another since they are relative to the number of parishes within a state.

Due to computational feasibility, in this work we used a maximum of four groups and we assumed the same number of groups for each variable. For smaller data or less complex models, it may be feasible to allow the number of groups to be greater than four, or even construct models that allow the estimation of the number of groups.

The estimates for some variables (such as Grade) were qualitatively different in various groups, showing associations in groups 3 and 4 (zero belonging to the credible interval), while no association was observed in groups 1 and 2, this emphasizing the need of using threshold groupings that can allow the discovery of different group specific patterns and association. For other variables, such as race, marital status, stage, and age of diagnosis, the results were qualitatively similar, these variables showing a similar association in all groups.

The width of the 95% credible intervals for the stage coefficients are not very wide, indicating that the coefficients were well estimated, even for some variables with a smaller percentage of observations in one of the categories, such as the stage variable. For some variables a fewer number of groups may be needed, especially if there are no qualitative differences between the estimates, while other variables may need a larger number of groups. While we have included the same number of groups for all variables due to computational simplicity, an extension of our model would be to allow the inclusion of different number of groups for each covariate, and to allow a different threshold level for each covariate. We expect that allowing different number of groups and even allow the number of groups to be estimated for each covariate could help reduce the model complexity, by reducing the number of groups for the variables in which the results do not differ qualitatively.

Prostate cancer survival data has been previously analyzed using random effects to account for spatial dependency [28, 2] and these models found a large spatial variation of PrCA. Similarly to the findings in our models, it was found that higher age of diagnosis, black race and distant stage were associated with lower survival. No effect was found in this paper related to marital status (as grouped by single, married and other). Grade was not included as a covariate. However, the association of the covariates was not allowed to differ spatially. Spatially explicit models were used to provide more interpretable measures of spatial all cause risk and these models have been found to be comparable, and in some cases outperform the random effects models [3, 4]. Threshold CAR models have been developed [6], which showed reasonably strong evidence in support of discrete spatial grouping in these cancer registry data. The model presented in this paper extended the threshold CAR model by allowing more than two groups and also using a spatially explicit model.

The model can be extended to allow a variety of classes of clusters (ordinal/nominal category prior distributions), and to allow other prior distributions such as Potts, singular multinomial or tessellation coloring priors [29, 30]. Further extensions of our analysis could be to use reversible jump MCMC to allow for estimation of the number of groups for each predictor.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This work was supported by the National Institutes of Health, [grant number xxxx].

References

1. Jarup L, Best N, Toledano MB, Wakefield J, Elliott P. Geographical epidemiology of prostate cancer in Great Britain. *Int J Cancer*. 10(97(5)):695–699.2002;
2. Zhang, Jiajia; Lawson, Andrew B. Bayesian parametric accelerated failure time spatial model and its application to prostate cancer. *J Appl Stat*. 8(2):591–603.2011; DOI: 10.1080/02664760903521476

3. Onicescu G, Lawson AB, Zhang J, Gebregziabher M, Wallace K, Eberth JM. Bayesian accelerated failure time model for space-time dependency in a geographically augmented survival model. *Stat Methods Med Res.* 26(5):2244–2256.2015; DOI: 10.1177/0962280215596186 [PubMed: 26220537]
4. Onicescu G, Lawson AB, Zhang J, Gebregziabher M, Wallace K, Eberth JM. Spatially-explicit survival modeling for small area cancer data. *J Appl Stat.* 45(3):568–585.2017; DOI: 10.1080/02664763.2017.1288200 [PubMed: 30906096]
5. Fotheringham, AS, Brunsdon, C, Charlton, ME. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* Chichester: Wiley; 2002.
6. Lawson, Andrew B; Choi, J; Zhang, J. Prior choice in discrete latent modeling of spatially referenced cancer survival. *Stat Methods Med Res.* 23(2):183–200.2014; DOI: 10.1177/0962280212447148 [PubMed: 22556109]
7. National Cancer Institute. SEER Research Data 1973–2010: Surveillance Epidemiology and End Results (SEER) Program Research Data (1973–2010). DCCPS, Surveillance Research Program, Surveillance Systems Branch; 2010. released April 2013, based on the November 2012 submission. Available at: www.seer.cancer.gov
8. Gelfand, Alan E, Diggle, Peter J, Fuentes, Montserrat; Guttorp, Peter. *Handbook of Spatial statistics.* CRC Press; 2010.
9. Higdon, D. *Quantitative Methods for Current Environmental Issues.* Springer; London: 2002. Space and space-time modeling using process convolutions.
10. D'Hondt O, Lopez-Martinez C, Ferro-Famil L, Pottier E. Spatially nonstationary anisotropic texture analysis in SAR images. *IEEE Transactions on Geoscience and Remote Sensing.* 45:3905–3918.2007;
11. Calder, Catherine A. A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics.* 19(1):39–48.2008; DOI: 10.1002/env.852
12. Moraga, Paula; Lawson, Andrew B. Gaussian component mixtures and CAR models in Bayesian disease mapping. *Comput Stat Data Anal.* 56:1417–1433.2012;
13. Calder, CA, Holloman, C, Higdon, D. Exploring space-time structure in ozone concentration using a dynamic process convolution model. In: Kass, et al., editors. *Bayesian Case Studies VI.* Springer; New York: 2002.
14. Spiegelhalter, David J; Best, Nicola G; van der Linde, Angelika. Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc Series B.* 64:583–639.2002;
15. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* 57:97–109.1970;
16. Chib, Siddhartha; Greenberg, Edward. Understanding the Metropolis-Hastings algorithm. *The American Statistician.* 49:327–335.1995;
17. Bezanson, Jeff; Karpinski, Stefan; Shah, Viral B; Edelman, Alan. *Julia: A fast dynamic language for technical computing.* 2012
18. R. R version 3.0.1. R Foundation for Statistical Computing; Vienna, Austria: 2013. <http://www.R-project.org>
19. Gelman, A, Carlin, JB, Stern, HS, Rubin, D. *Bayesian data analysis.* CRC Press; 2004.
20. Gelman, Andrew; Hwang, Jessica; Vehtari, Aki. Understanding predictive information criteria for bayesian models. *Statistics and Computing.* 24(6):997–1016.2014; DOI: 10.1007/s11222-013-9416-2
21. Gelman, Andrew; Rubin, Donald B. Inference from iterative simulation using multiple sequences, statistical science. *Statistical Science.* 7(4):457–511.1992;
22. American Cancer Society. [Accessed February 13, 2018] 2016. Available at: <http://www.cancer.org>
23. Banerjee, Sudipto; Wall, Melanie M; Carlin, Bradley P. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics.* 4:123–142.2003; [PubMed: 12925334]
24. Carlin, Bradley P, Banerjee, Sudipto. Hierarchical multivariate CAR models for spatio-temporally correlated survival data, volume 4 of *Bayesian Statistics 7.* Bernardo, JB, Dawid, AP, Berger, JO, West, M, editors. Oxford University Press; New York: 2003.

25. Diva, Ulysses; Dey, Dipak K; Banerjee, Sudipto. Parametric models for spatially correlated survival data for individuals with multiple cancers. *Stat Med.* 27(12):2127–2144.2008; DOI: 10.1002/sim.3141 [PubMed: 18167633]
26. Banerjee S, Dey DK. Semiparametric proportional odds models for spatially correlated survival data. *Lifetime Data Anal.* 11(2):175–191.2005; [PubMed: 15938545]
27. National Cancer Institute. [Accessed December 22, 2015] Surveillance, epidemiology and end results program. months survived based on complete dates. 2013. Available at: <http://www.census.gov/acs/>
28. 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.* 2012
29. Guillot, Gilles. On the inference of spatial structure from population genetics data. *Bioinformatics.* 25(14):1796–1801.2009; DOI: 10.1093/bioinformatics/btp267 [PubMed: 19574291]
30. Guillot, Gilles; Estoup, Arnaud; Cosson, Jean Francois. A spatial statistical model for landscape genetics. *Genetics.* 170(3):1261–1280.2005; [PubMed: 15520263]

Highlights

- An extended threshold CAR model was developed to define the spatial grouping in the context of Bayesian spatially explicit survival model
- The model was applied to analyze the SEER data prostate cancer data in Louisiana
- The estimates suggest a spatial variability, with some parishes with higher risk are in the South and North-Western part of the state, while parishes in the North-East have lower risk.

Figure 1.

Figure 2.

Map of model based estimated spatial risk probabilities from the STSU-Weibull threshold CAR model with four groups adjusted for race, marital status, stage, grade and age at diagnosis.

Figure 3.
Map of estimated group indicator from the STSU-Weibull threshold CAR model with four groups

Table 1

Model comparison, STSU-Weibull threshold CAR

	DIC	p_D
Model 1 ^a	122709.2	35.28
Model 2 ^b	122457.3	38.84
Model 3 ^c	122432.1	72.90

^a 1 threshold, 2 groups^b 2 thresholds, 3 groups^c 3 thresholds, 4 groups

Table 2

Parameter estimates and 95% credible intervals (CI), STSU-Weibull threshold CAR for a model with G=4 groups.

	Stratum	Estimate (sd)	95 % CI
Intercept	1	-6.94 (6.73)	(-7.24,-6.68) *
	2	-7.074 (6.79)	(-7.37,-6.58) *
	3	-7.11 (6.76)	(-7.50,-6.48) *
	4	-7.00 (6.80)	(-7.34,-6.50) *
Race			
White and Other races	Ref.		
Blacks	1	0.38 (2.79)	(0.12,0.54) *
	2	0.45 (2.97)	(0.23,0.79) *
	3	0.37 (2.87)	(0.021,0.71) *
	4	0.37 (2.82)	(0.11,0.63) *
Marital status			
Single/Separated/Divorced/Widowed	Ref.		
Married/Domestic partner	1	-0.51 (2.92)	(-0.85,-0.15) *
	2	-0.35 (2.86)	(-0.55,-0.12) *
	3	-0.39 (2.83)	(-0.67,-0.19) *
	4	-0.55 (3.12)	(-0.79,-0.23) *
Stage			
Localized/Regional	Ref.		
Distant	1	1.96 (4.59)	(1.47,2.22) *
	2	1.98 (4.57)	(1.72,2.24) *
	3	1.78 (4.42)	(1.37,2.35) *
	4	1.87 (4.50)	(1.59,2.27) *
Grade			
1 and 2	Ref.		
3 and 4	1	0.17 (2.16)	(-0.082,0.46)
	2	0.14 (2.24)	(-0.046,0.39)
	3	0.33 (2.74)	(0.066,0.54) *
	4	0.30 (2.59)	(0.032,0.66) *
Age at diagnosis (standardized)	1	0.69 (3.35)	(0.55,0.83) *
	2	0.71 (3.43)	(0.59,0.83) *

	Stratum	Estimate (sd)	95 % CI
	3	0.72 (3.39)	(0.59,0.86) *
	4	0.70 (3.34)	(0.58,0.84) *

* indicates that the credible interval does not include zero.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript