BayesSPsurv: An R Package to Estimate Bayesian (Spatial) Split-Population Survival Models

by Brandon Bolte, Nicolás Schmidt, Sergio Béjar, Nguyen Huynh, Bumba Mukherjee

Abstract Survival data often include a fraction of units that are susceptible to an event of interest as well as a fraction of "immune" units. In many applications, spatial clustering in unobserved risk factors across nearby units can also affect their survival rates and odds of becoming immune. To address these methodological challenges, this article introduces our **BayesSPsurv** R-package, which fits parametric Bayesian Spatial split-population survival (cure) models that can account for spatial autocorrelation in both subpopulations of the user's time-to-event data. Spatial autocorrelation is modeled with spatially weighted frailties, which are estimated using a conditionally autoregressive prior. The user can also fit parametric cure models with or without nonspatial i.i.d. frailties, and each model can incorporate time-varying covariates. **BayesSPsurv** also includes various functions to conduct pre-estimation spatial autocorrelation tests, visualize results, and assess model performance, all of which are illustrated using data on post-civil war peace survival.

1 Introduction

Conventional survival models have been applied to analyze time-to-event data across several academic disciplines, but these models rely on two core assumptions that are not always tenable. The first is that all units, including right-censored observations, will eventually experience the event of interest. In many applications, however, some fraction of subjects that are "immune" or "cured" may never experience the event (Maller and Zhou, 1996; Peng and Taylor, 2014; Beger et al., 2017). A clinical study of obesity on human death rates can employ a standard survival model because all subjects will eventually die, but a study on vaccine effectiveness will likely assume that some fraction of the treated population will become immune while others may respond differently and remain uncured. To account for both subpopulations, scholars have developed a class of split-population (SP) survival models that probabilistically separate the immune fraction from the units that are susceptible to the event of interest and then estimate the conditional hazard of survival among the latter units (Cai et al., 2012; Beger et al., 2018; Box-Steffensmeier and Zorn, 1999; Box-Steffensmeier and Jones, 2004). The second assumption of conventional survival models is that each observation is conditionally independent after controlling for covariates. This assumption is violated if spatially clustered units share common unobserved features that influence their baseline risk of experiencing an event (Darmofal, 2009; Taylor and Rowlingson, 2017). If spatial autocorrelation matters for process survival and/or the probability of being immune to an event, ignoring it can lead to biased parameter estimates.

Although there are numerous methods for dealing with spatially autocorrelated time-to-event data (e.g., Li and Ryan, 2002; Banerjee et al., 2003; Henderson et al., 2002; Zhou et al., 2020), these spatial survival models do not differentiate between at-risk and immune observations. Conversely, existing SP survival models either ignore spatial autocorrelation altogether or only account for it in the survival stage (Banerjee and Carlin, 2004). We present the **BayesSPsurv** (Bolte et al., 2021b) R-package (available at https://CRAN.R-project.org/package=BayesSPsurv), which has the power to overcome these limitations as well as the flexibility for researchers to model spatial clustering in their survival data however they choose to define it. In particular, **BayesSPsurv** allows the user to estimate parametric Bayesian Spatial split-population survival (cure) models with spatial frailties in both the model's split and survival stages. These models account for spatial clustering in the immune and at-risk fractions of the data. The package also includes functions and code for pre-estimation spatial autocorrelation diagnostics, visualizing results, simulating multiple Markov chains, and implementing Markov Chain Monte Carlo (MCMC) estimation routines for SP survival models with independent (exchangeable) or no random effects. The user can also include time-varying covariates in either stage. In the next section, we outline previous work on SP and spatial survival models, including existing packages and their limitations. We then formally develop the Pooled, Exchangeable, and Spatial Bayesian SP survival models before describing the various functions available in the **BayesSPsurv** package. We demonstrate these functions using replication data from a published study on the survival of post-civil war peace.

2 Background and other R-packages

Scholars have identified at least two sources of conditional heterogeneity in survival data that hasve been separately addressed. The first occurs when a subset of the units in the data are immune from experiencing a "failure" event, which violates the assumption that even right-censored observations experience the event of interest (Box-Steffensmeier and Zorn, 1999; Box-Steffensmeier and Jones, 2004; Beger et al., 2017). Cure rate or split-population (SP) survival models account for this immune fraction by first estimating the probability that right-censored units are immune from an event in a "split-stage" and then the time until *at-risk* censored and non-censored units experience the event. Recent work has extended these models to include i.i.d. frailties (Peng and Taylor, 2011), time-varying covariates (Beger et al., 2017), and account for random right-censoring (Patilea and Van Keilegom, 2020). Split-population survival models have been used to study phenomena, ranging from the survival of breast cancer patients (Wang et al., 2005), criminal recidivism (Schmidt and Witte, 1989), the risk of coups (Beger et al., 2017), and parasite-induced mortality in river salmon (Ray et al., 2014).

Separately, conventional survival models have been extended to account for spatial autocorrelation among nearby units (Banerjee et al., 2003; Taylor and Rowlingson, 2017; Zhou et al., 2020). These models relax the assumption of spatial independence by incorporating spatially weighted frailties into the survival model's baseline hazard function. This allows for the possibility that adjacent units share unmodeled risk factors that influence their underlying propensities for experiencing a failure event. Apart from recent advances in modeling spatial frailties in different survival frameworks (e.g. Diva et al., 2008; Zhou et al., 2020), spatial survival models have been applied to analyze, for example, geographically referenced data on leukemia survival (Henderson et al., 2002), position announcements by U.S. House members (Darmofal, 2009), prostate cancer (Zhou et al., 2020), and fire service response times (Taylor and Rowlingson, 2017). Despite these considerable methodological developments, far less attention has been dedicated to accounting for spatial autocorrelation in SP survival settings. This is surprising because spatial autocorrelation between units may influence their probability of being immune and the survival rate among the units at risk of experiencing the failure event simultaneously. Banerjee and Carlin (2004) develop Bayesian spatial cure models but focus on modeling spatial autocorrelation in the survival stage using a conditionally autoregressive (CAR) prior. In fact, these authors emphasize that future research must "include covariates and spatial random effects as regressors in the cure rate portion of the model, instead of just the log-relative risk portion" (274).

In line with these trends in the literature, some R packages fit standard and split-population survival models but do not allow for the incorporation of spatial information. The **survival** package (Therneau, 2020) fits parametric and semiparametric Cox survival models via MLE, whereas **dynsurv** (Wang et al., 2020) fits the Cox Proportional Hazards (PH) model with dynamic coefficients using MCMC methods. Conventional semi-parametric cure models can be estimated via MLE with the **smcure** (Cai et al., 2012) or **nltm** (Garibotti et al., 2019) package. The **flexsurvcure** (Amdahl, 2020) package fits parametric mixture and non-mixture cure models for time-to-event data, and the **spduration** package fits parametric SP survival models with time-varying covariates (Beger et al., 2018).

A small handful of packages allow the user to incorporate spatial information into their survival models, but *never* in split-population settings. The **BayesX** package (Umlauf et al., 2019) and its associated interface to R **R2BayesX** (Belitz et al., 2017) fit spatial survival models and structured additive regression models with spatial frailties. The **spBayesSurv** package (Zhou and Hanson, 2020) fits several Bayesian survival models with spatial frailties that can be formulated on a PH, proportional odds, or AFT scale, all of which can include time-varying covariates. The **spatsurv** package also fits Bayesian spatial survival models, including a PH model that permits users to incorporate Gaussian process frailties (Taylor and Rowlingson, 2020). Beyond R, WinBUGS and GeoBUGs code has been developed to fit, for instance, survival and non-mixture cure models with spatial frailties (Banerjee et al., 2003, 2004; Thomas et al., 2004).

To our knowledge, our **BayesSPsurv** package is the first to allow users to fit parametric split-population survival models with not just time-varying covariates but also spatial frailties in both stages. The frailties in the parametric Bayesian spatial SP model are estimated using the CAR prior approach (Besag et al., 1991; Bernardinelli et al., 1995; Banerjee et al., 2003; Banerjee and Carlin, 2004). **BayesSPsurv** also includes functions and routines coded in C++ to fit nonspatial parametric SP survival models with exchangeable frailties in the model's split and survival-stage equation and without any frailties. Statistical inference of the models in **BayesSPsurv** is conducted via combined MCMC techniques that require little input from users. Our package and supplemental code also provide functions to implement spatial autocorrelation tests, produce country-level adjacency matrices, generate and compare multiple Markov chains, assess convergence, and conduct model comparison. Before outlining the functionality of the package in greater detail, we turn to briefly develop each of the three included Bayesian split-population survival models.

3 The Bayesian (*Spatial*) split-population survival model

Model development

Define $i = \{1, 2, ...N\}$ for the units that may fail or experience an event of interest in a continuous-time survival dataset. Let $f(t)$ and $F(t)$ represent the probability density function and cumulative distribution function. The survival distribution is $S(t) = 1 - F(t)$, and the hazard rate is $h(t) = \frac{f(t)}{S(t)}$. Some units will fail during the time period under observation ($\widetilde{C}_i = 1$), while others do not and are "censored" ($\widetilde{C}_i = 0$). The general likelihood of the conventional survival model in which all units eventually experience the event of interest is

$$\prod_{i=1}^{N} \left[f\left(t_i\right) \right]^{\widetilde{C}_i} \left[S\left(t\right) \right]^{1-\widetilde{C}_i}. \tag{1}$$

Suppose that the survival data includes two subpopulations: an "at-risk" fraction that can fail and an "immune" fraction that will *not* experience the (failure) event of interest (Maller and Zhou, 1996; Yin and Ibrahim, 2005; Beger et al., 2017). When presented with this data generation process, researchers typically employ split-population survival (cure) models with or without unit-specific frailties to simultaneously estimate the probability of observations being in the immune fraction and the effect of covariates on the hazard of survival among the at-risk fraction (Maller and Zhou, 1996; Lu, 2010; Peng and Taylor, 2014).

To understand these models in more detail, consider the split-population survival model for the duration t that splits the sample into an at-risk and an immune fraction. Let $\alpha_i = Pr(Y_i = 1)$ be the probability with which units enter the immune fraction. α_i can be estimated via a binary response function and is defined for the logit case as:

$$\alpha_i = \frac{\exp\left(\mathbf{Z}_i \gamma + V_i\right)}{1 + \exp\left(\mathbf{Z}_i \gamma + V_i\right)}, \tag{2}$$

where $\mathbf{Z}_i$ are p2-dimensional covariates, γ the parameter vector in $\mathbb{R}^{p_2}$, and $V_i \sim N(0, \sigma^2)$ are the nonspatial i.i.d unit-specific frailties (random effects). Equation 2 is the split-population model's split-stage equation, where the unit-specific frailties V_i, which are each independent of other individual random effects, account for unobserved heterogeneity that influences probability α_i. Let $W_i \sim N(0, \sigma^2)$ denote the nonspatial i.i.d unit-specific frailties that capture the possibility that some units are at different risks of experiencing the event of interest due to unobserved factors. The proportional hazards function of the SP survival model with nonspatial, unit-specific frailties is

$$h\left(t_i | \mathbf{X}_i \boldsymbol{\beta}, W_i\right) = h_0\left(t_i\right) \omega_i \exp\left(\mathbf{X}_i \boldsymbol{\beta}\right) = h_0\left(t_i\right) \exp\left(\mathbf{X}_i \boldsymbol{\beta} + W_i\right), \tag{3}$$

where $h_0(t_i)$ is the baseline hazard (e.g., Weibull, log-logistic), $\log \omega_i = W_i$, $\mathbf{X}_i$ is the p1-dimensional covariates, and $\boldsymbol{\beta}$ the parameter vector in $\mathbb{R}^{p_1}$. We focus on incorporating unit-specific frailty terms generally because they are most commonly used in the social sciences. However, our approach could plausibly be extended to a shared frailty framework if the researcher believes that the frailties occur in clusters such that within-cluster frailties are correlated while frailties between clusters are independent.

Suppose, however, that the survival data with the two aforementioned subpopulations should be fit with time-varying covariates. We can re-define this data with unique "entry time" duration as $t0$ and "exit time" as duration t for each period at which an observation is observed. Let $t0_{ij}$ denote unit i's elapsed time since inception until the beginning of time period j, t_{ij} the elapsed time since that unit's inception until the end of period j, and $\tilde{C}_{ij} = 1$ if that unit fails or is censored ($\tilde{C}_{ij} = 0$) at t_{ij}. The probability of survival up until period j is now $S(t0) = 1 - F(t0)$ where $F(t0) = \int_0^{t0} f(t0)$. In this case, both subpopulations contribute to the log-likelihood of the split-population survival model with nonspatial i.i.d frailties as:

$$\ln L = \sum_{i=1}^{N} \left\{ \widetilde{C}_{ij} \ln \left[\left(1 - \alpha_{ij}\right) \frac{f\left(t_{ij} | \mathbf{X}_{ij} \boldsymbol{\beta}, W_i\right)}{S\left(t0_{ij} | \mathbf{X}_{ij} \boldsymbol{\beta}, W_i\right)} \right] + \left(1 - \widetilde{C}_{ij}\right) \ln \left[\alpha_i + \left(1 - \alpha_i\right) \frac{S\left(t_{ij} | \mathbf{X}_{ij} \boldsymbol{\beta}, W_i\right)}{S\left(t0_{ij} | \mathbf{X}_{ij} \boldsymbol{\beta}, W_i\right)} \right] \right\}, \tag{4}$$

where the "split-stage" equation is $\alpha_{ij} = \frac{\exp\left(\mathbf{Z}_{ij} \gamma + V_i\right)}{1 + \exp\left(\mathbf{Z}_{ij} \gamma + V_i\right)}$. V_i and W_i are the nonspatial frailties. The model's survival stage estimates the probability of survival prior to the event of interest conditional upon being at-risk for that event given covariates $\mathbf{X}_{ij}$ and the baseline hazard function. If $V_i = W_i = 0$, then (4) defines the log-likelihood of the "Pooled" SP survival model (*without* unit-specific frailties)

with time-varying covariates (Ibrahim et al., 2001; Lu, 2010). However, if unobserved unit-specific heterogeneity influences the probability of immunity or survival time, it can be accounted for with the split and survival-stage frailty terms (V_i and W_i). In a Bayesian split-population survival framework, these frailties are incorporated into each stage of the model using the exchangeable normal prior,

$$W_i \sim N\left(0, 1/\tau\right) \text{ and } V_i \sim N\left(0, 1/\tau\right), \tag{5}$$

with τ as the precision parameter (Banerjee et al., 2003; Banerjee and Carlin, 2004). The prior in (5) is induced by treating each specified unit as exchangeable rather than assigning weights corresponding to each unit's spatial relationship to one another (Bernardinelli and Montomoli, 1992; Darmofal, 2009). This Exchangeable split-population survival model is appropriate if each unit's frailty is presumed to be independent of other individual random effects. Geographically, for instance, this means that the influence of each unit-specific frailty on that unit's probability of being immune or its risk propensity is completely unrelated to the neighboring units' frailties unobserved effects.

Suppose, however, that independence among the frailties *cannot* be assumed—that is, that the frailties exhibit spatial autocorrelation or clustering that influences each units' propensity for being immune to an event of interest and their survival time if they are not immune. In a Bayesian split-population survival model, the assumption of spatial independence is relaxed by assigning spatial weights to the unit-specific frailties in the model's split and survival stage and then statistically incorporating these spatially weighted frailties via the conditionally autoregressive (CAR) prior approach (Besag et al., 1991; Banerjee et al., 2003). The CAR prior accounts for spatially autocorrelated frailties by allowing the frailties to be spatially autocorrelated across geographically adjacent units.

To understand how the CAR prior is applied, first note that spatial data often take the form of a lattice in which a continuous spatial surface is divided into a grid of units such as counties, districts, or countries. The spatially weighted frailties are constructed by defining the relevant spatial relationship among adjacent units (this could, for example, be geographic distance or contiguity) in an adjacency matrix $\mathbf{A}$ with elements $a_{ii'}$. Each element $a_{ii'}$ in $\mathbf{A}$ is given a weight of 1 if units i and i' are "neighbors," and 0 if they are not. Once these spatial weights are defined via the matrix $\mathbf{A}$, this information is then incorporated into the CAR prior, which permits us to model spatially dependent frailties between these contiguous units. To employ the CAR prior approach in a Bayesian SP survival framework, the frailties V_i are collected into the vector $\mathbf{V} = \{V_1, ..., V_N\}$, and W_i into $\mathbf{W} = \{W_1, ..., W_N\}$. Separate CAR priors are then employed for $\mathbf{V}$ and $\mathbf{W}$, which implies the following model structure:

$$\mathbf{V}|\lambda \sim \text{CAR}\left(\lambda\right) \text{ and } \mathbf{W}|\lambda \sim \text{CAR}\left(\lambda\right). \tag{6}$$

λ is the precision parameter (Besag et al., 1991; Banerjee and Carlin, 2004). The CAR(λ) prior for $\mathbf{V}$ and $\mathbf{W}$ has a joint distribution in each case that has been formally characterized by scholars (Banerjee et al., 2003, 126).

The resulting conditional distributions of the spatial frailties for $\mathbf{V}$ and $\mathbf{W}$ are

$$V_i|V_{i' \neq i} \sim N\left(\overline{V_i}, 1/\left(\lambda m_i\right)\right), \quad W_i|W_{i' \neq i} \sim N\left(\overline{W_i}, 1/\left(\lambda m_i\right)\right). \tag{7}$$

$\overline{W_i} = m_i^{-1} \sum_{i' \; adj \; i} W_{i'}$ and $\overline{V_i} = m_i^{-1} \sum_{i' \; adj \; i} V_{i'}$. $\overline{W_i}$ and $\overline{V_i}$ are the averages of the neighboring $W_{i' \neq i}$ and $V_{i' \neq i}$, respectively, where $i' \; adj \; i$ denotes that i' is adjacent to i given $\mathbf{A}$, and m_i is the number of these adjacencies (Bernardinelli and Montomoli, 1992, 989; Thomas et al., 2004; Banerjee et al., 2003). Incorporating the spatial information in $\mathbf{A}$ in this way accounts for the possibility that spatially proximate units share common unmodeled factors that influence their probability of being immune or their survival time before experiencing the event of interest. Using this CAR prior approach to address spatial autocorrelation, the Spatial split-population survival model's log-likelihood is defined by substituting $\mathbf{V} = \{V_i\}$ and $\mathbf{W} = \{W_i\}$ in equation 4 (where $\alpha_{ij} = \frac{\exp(\mathbf{Z}_{ij}\gamma, \mathbf{V})}{1 + \exp(\mathbf{Z}_{ij}\gamma, \mathbf{V})}$ is the split-stage equation).

The log-likelihood of the Pooled (non-frailty), Exchangeable (nonspatial frailty), and Spatial split-population (SP) survival models are compatible with any survival distribution. The **BayesSPsurv** package, however, supports MCMC estimation of these models for the Weibull and log-logistic distributions. Our empirical application below focuses on the Weibull survival distribution. The density, survival, and hazard rate in the Weibull case are

$$f\left(t_{ij}|\rho, \theta\right) = \rho\theta\left(\theta t_{ij}\right)^{\rho-1} \exp\left(-\left(\theta t_{ij}\right)^{\rho}\right)$$

$$S\left(t_{ij}|\rho, \theta\right) = \exp\left(-\left(\theta t_{ij}\right)^{\rho}\right) \text{ and } h\left(t_{ij}|\rho, \theta\right) = \rho\theta\left(\theta t_{ij}\right)^{\rho-1}, \tag{8}$$

where $\theta = \exp\left(\mathbf{X}_{ij}\boldsymbol{\beta}, \mathbf{W}\right)$ for the Spatial, $\theta = \exp\left(\mathbf{X}_{ij}\boldsymbol{\beta} + W_i\right)$ for the Exchangeable, and $\theta = \exp\left(\mathbf{X}_{ij}\boldsymbol{\beta}\right)$ for the Pooled split-population Weibull model. The density, survival function and the hazard rate for the log-logistic case is defined in Bolte et al. (2021a).

Bayesian inference

Following standard practice for Bayesian inference (Carlin and Louis, 2000), we assign the multivariate normal (MVN) prior to $\boldsymbol{\beta} = \{\beta_1, ..., \beta_{p_1}\}$ and $\boldsymbol{\gamma} = \{\gamma_1, ..., \gamma_{p_2}\}$, and the Gamma prior for ρ with shape and scale parameters a_ρ and b_ρ for *each* of the three Bayesian split-population survival models in the **BayesSPsurv** package.

$$\rho \sim \text{Gamma}\left(a_\rho, b_\rho\right), \quad \boldsymbol{\beta} \sim \text{MVN}_{p_1}\left(\mu_\beta, \Sigma_\beta\right), \quad \boldsymbol{\gamma} \sim \text{MVN}_{p_2}\left(\mu_\gamma, \Sigma_\gamma\right)$$
$$\Sigma_\beta \sim \text{IW}\left(S_\beta, \nu_\beta\right); \ \Sigma_\gamma \sim \text{IW}\left(S_\gamma, \nu_\gamma\right),$$

(9)

where a_ρ, b_ρ, S_β, ν_β, S_γ, ν_γ are the hyperparameters. We use Bayesian hierarchical modeling to estimate Σ_β and Σ_γ employing the Inverse-Wishart (IW) distribution when using the MVN (a weakly informative) prior. For Bayesian MCMC estimation of the Spatial SP survival (Weibull) model, we assign the hyperprior $p(\lambda)$ to λ given the CAR prior approach. Specifically, we assign the Gamma hyperprior $\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$ for λ (Banerjee and Carlin, 2004; Darmofal, 2009).[1] To estimate the Exchangeable SP Weibull model, we assign the (multivariate) normal prior for the model's split and survival-stage frailties (V_i, W_i), and the priors defined in (9) for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and ρ. To identify the Exchangeable and Spatial SP models' intercepts, we impose the constraint that the frailties sum to zero ($\sum_i V_i = 0$ and $\sum_i W_i = 0$).

The joint posterior distribution of the Bayesian Spatial SP survival model with time-varying covariates—our main model of interest—is

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \mathbf{W}, \mathbf{V}, \lambda, \Sigma_\beta, \Sigma_\gamma | \mathbf{C}, \mathbf{X}, \mathbf{Z}, \mathbf{t}, \mathbf{t0}, \boldsymbol{\gamma}\right) \quad \propto \quad L\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \mathbf{W}, \mathbf{V} | \mathbf{C}, \mathbf{X}, \mathbf{Z}, \mathbf{t}, \mathbf{t0}\right)$$
$$\pi\left(\mathbf{W} | \lambda\right) \pi\left(\mathbf{V} | \lambda\right) \pi\left(\beta | \mu_\beta, \Sigma_\beta\right) \pi\left(\gamma | \mu_\gamma, \Sigma_\gamma\right) \pi\left(\rho\right) \pi\left(\lambda\right) \pi\left(\Sigma_\beta\right) \pi\left(\Sigma_\gamma\right),$$

(10)

where $L\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \mathbf{W}, \mathbf{V} | \mathbf{C}, \mathbf{X}, \mathbf{Z}, \mathbf{t}, \mathbf{t0}\right)$ is defined in (4) with $\mathbf{V} = \{V_1, ..., V_N\}$ and $\mathbf{W} = \{W_1, ..., W_N\}$. The density, survival function, and hazard rate for this likelihood is given by the Weibull (or log-logistic) distribution in our R-package. $\pi\left(\mathbf{W} | \lambda\right)$ and $\pi\left(\mathbf{V} | \lambda\right)$ are defined via their respective conditional distributions in (7). $\pi\left(\beta | \mu_\beta, \Sigma_\beta\right)$, $\pi\left(\gamma | \mu_\gamma, \Sigma_\gamma\right)$, and $\pi\left(\rho\right)$ are from (9), $\pi\left(\Sigma_\beta\right)$ and $\pi\left(\Sigma_\gamma\right)$ are from (10). $\pi\left(\lambda\right)$ is the Gamma hyperprior for the Spatial SP survival model. From (10), we can define the joint posterior distribution of the time-varying (i) Exchangeable SP survival model by incorporating the frailties V_i and W_i (instead of $\mathbf{W}$, $\mathbf{V}$, and their CAR priors) given by (5) and (ii) Pooled SP survival model by excluding frailty terms.

The three split-population survival models in **BayesSPsurv** are each estimated with an MCMC algorithm for Bayesian inference. Because closed-forms for the posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and ρ are not available, these parameters in each model are updated in the MCMC algorithm via slice-sampling (with stepout and shrinkage) from their respective full conditional distribution. The closed form of the full conditional distributions (e.g., $\pi\left(\Sigma_\beta | \beta\right)$ and $\pi\left(\Sigma_\gamma | \gamma\right)$) and details about slice-sampling is provided in Bolte et al. (2021a).[2] Further, because the closed-forms for the posterior distributions of λ, $\mathbf{W}$, and $\mathbf{V}$ are not available for the Spatial split-population survival model, our MCMC algorithm incorporates Gibbs Sampling for estimating λ. Our MCMC update scheme then employs the Metropolis-Hastings algorithm for estimating $\mathbf{V}$ given λ and then uses the Metropolis-Hastings algorithm to estimate $\mathbf{W}$ given λ.[3] In the Exchangeable SP survival model, the nonspatial i.i.d frailties V_i and W_i are updated via Metropolis-Hastings, while the Pooled SP model excludes frailty terms. The MCMC update scheme to fit each model in **BayesSPsurv** are described in detail in Bolte et al. (2021a).

---

[1] We specify the vague prior $(a_\lambda, b_\lambda) = (0.001, 1/0.001) = (0.001, 1000)$ as done for ρ.

[2] For more information on slice-sampling, see Neal (2003).

[3] The closed-form of the full conditional distribution of $\pi(\lambda | \mathbf{W}, \mathbf{V})$ used for the MCMC update scheme is formally characterized in Bolte et al. (2021a).

4   Using the BayesSPsurv R package

Each of the three Bayesian SP survival models in **BayesSPsurv** incorporates a cure rate fraction and assumes that the time-to-event baseline hazard follows a Weibull or log-logistic distribution (which the user specifies). Users can also incorporate time-varying covariates in either stage. **BayesSPsurv** contains compiled C++ code using the package **Rcpp** (Eddelbuettel et al., 2020) to maximize computational efficiency when estimating the included Bayesian SP survival models. In addition to the pre-estimation spatial autocorrelation (Join Count and Global Moran's I) tests described below, the **BayesSPsurv** package also permits users to calculate the deviance information criterion (DIC) and log-likelihood statistics from the Spatial, Exchangeable, and Pooled SP survival models' MCMC output. The DIC is a measure of model fit that also penalizes the effective number of parameters. Like the Akaike Information Criterion, models with smaller values are preferable to those with larger values. Users can also conduct MCMC diagnostics with various extant packages designed to handle mcmc objects such as **coda** (Plummer et al., 2020).

Function	Description
spatialSPsurv()	Fits Bayesian Spatial SP survival model
exchangeSPsurv()	Fits Bayesian Exchangeable SP survival model
pooledSPsurv()	Fits Bayesian Pooled SP survival model
plot_JoinCount()	Implement and plot Join Count statistics
plot_Moran.I()	Implement and plot global Moran's I statistics
spatial_SA()	Generate spatial weights (adjacency) matrix
SPstats()	Calculate DIC and log-likelihood from fitted models

Table 1: Functions in the **BayesSPsurv** package.

Loading the package, dataset, and assessing spatial autocorrelation

To demonstrate the utility and various functions in **BayesSPsurv**, we use replication data from Walter's (2015) global study on post-civil war peace duration (denoted in the package as Walter_2015_JCR). Walter's (2015) panel data consist of 1,237 observations from 46 countries observed between 1962 and 2009. As discussed in Bolte et al. (2021a), her data are well-suited for SP survival analysis because they include two underlying populations: an "at-risk" fraction of countries in which civil wars can potentially recur, and an "immune" fraction of countries in which civil conflict recurrence is structurally improbable. The Walter_2015_JCR data are a subset of Walter's (2015) most important variables, including lgdpl for log per capita income, unpko for the presence of UN peacekeeping operations, the binary variable victory, coded as 1 when one side in the civil war wins the conflict militarily, and the dummy variable comprehensive, coded as 1 when the combatants sign a comprehensive peace agreement. The dataset also includes the binary indicator renewed_war, coded as 1 for the year in which a civil war recurs and 0 otherwise. This variable will serve as our failure event.

First, however, we load the **BayesSPsurv** package along with the dataset and then use the add_duration() function from the **spduration** package to add several variables that allow us to capture the survival characteristics of the data (Beger et al., 2017). unitID indicates the unique unit identifier (in this case, a unique id for each civil conflict), and timeID specifies the temporal variable.

```
library(BayesSPsurv)
data(Walter_2015_JCR)
walter <- spduration::add_duration(Walter_2015_JCR,"renewed_war",
                                   unitID = "id", tID = "year",
                                   freq = "year", ongoing = FALSE)
str(walter)
'data.frame':        1237 obs. of  21 variables:
 $ year        : num  2002 2003 2004 2005 2003 ...
 $ lastyear    : num  0 0 0 1 0 0 0 0 0 0 ...
 $ renewed_war : num  0 0 0 1 0 0 0 0 0 0 ...
 $ fhcompor1   : num  -1.17 -1.17 -1.08 -1 -1.08 ...
 $ lgdpl       : num  5.75 6.29 6.36 6.4 8 ...
 ...
 $ failure     : num  0 0 0 1 0 0 0 0 0 0 ...
 $ ongoing     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ end.spell   : num  0 0 0 1 0 0 0 0 0 0 ...
```

```
$ cured       : num  0 0 0 0 1 1 1 1 1 1 ...
$ atrisk      : num  1 1 1 1 0 0 0 0 0 0 ...
$ censor      : num  0 0 0 0 0 0 0 0 0 0 ...
$ duration    : num  1 2 3 4 1 2 3 4 5 6 ...
$ t.0         : num  0 1 2 3 0 1 2 3 4 5 ...
```

These new variables include `duration`, a cumulative count of the years of post-war peace survived, and the dummy variable `atrisk`, coded as 1 for all observations that eventually experience war recurrence in the sample period. Altogether the data include 77 post-civil war peace spells and 24 instances of civil war recurrence.

If the preferred frailty unit is at the country-level, users can take advantage of the `spatial_SA()` function in the **BayesSPsurv** package to generate their binary spatial weights matrix. In most cases, analysts define spatial clustering in terms of geographic proximity. This often requires the researcher to define some maximum distance threshold below which two units are considered "neighbors." `spatial_SA()` allows users to specify their own distance threshold. For this example, we use the `spatial_SA()` function to generate an adjacency matrix of countries in the data whereby "proximity" ($a_{ii'} = 1$) is defined as having capitals that are within 800 km of each other (and $a_{ii'} = 0$ otherwise). We simply specify the unique identifier for each country (ccode) and our distance threshold of 800 km. The `spatial_SA()` function will produce a $1 \times N$ vector with identifying information (e.g., country ID) for each observation that matches the rows and columns of the matrix. The result is a list object called `walter` that includes both the original data frame and the associated adjacency matrix.

```
walter <- BayesSPsurv::spatial_SA(data = walter, var_ccode = "ccode",
walter[[2]][1:6,1:6]
```

```
        42  90  92  93  135  155
    42   0   0   0   0    0    0
    90   0   0   1   1    0    0
    92   0   1   0   1    0    0
    93   0   1   1   0    0    0
   135   0   0   0   0    0    0
   155   0   0   0   0    0    0
```

Note that spatial adjacency between units does not need to be defined geographically; users can conceptualize "space" as any form of the dyadic relationship between units, but typically spatial clustering is substantively captured with some measure of geographic distance. Users can also create their own adjacency matrix from scratch to incorporate into the Bayesian estimation routine if their units of analysis are something other than countries.

Having generated a matrix that records the spatial relationship of all pairs of units, we may now be interested in assessing the presence and degree of spatial autocorrelation in the data with respect to our outcomes of interest. The **BayesSPsurv** package provides two functions to conduct these preliminary tests on country-level data. The first is the `plot_JoinCount()` function, which generates an adjacency matrix with a user-defined distance threshold (as above), implements the join count test for each cross-section in the data, and then automatically plots the test statistics with user-specified confidence intervals across each observed year. The join count test is a widely used correlational statistic for evaluating whether the expected count of categorically distinct adjacent units is greater than what we would expect by chance alone (Cliff and Ord, 1981). More formally, if we assume two categories, Black and White, then the join count test statistic is

$$Z\left(BW\right) = \frac{BW - E\left(BW\right)}{\sqrt{\sigma_{BW}^2}}, \tag{11}$$

where BW and E(BW) are the observed and expected counts of adjacent Black and White units, respectively, and $\sigma_{BW}^2 = E\left(BW^2\right) - E\left(BW\right)^2$. We use this test as a preliminary exercise to determine whether countries that *never* experience conflict recurrence in the sample exhibit spatial autocorrelation. Note that `plot_JoinCount` also has an argument specifying the minimum number of units that must be in a cross-section for the test statistic to be calculated. In our case, there were fewer than 12 countries in every year prior to 1975 in the data, and none were geographically proximate. By specifying n=12, we are simply dropping years prior to 1975. The following code produces the plot in Figure 1a.

```
plot_JoinCount(data = walter[[1]], var_cured = "cured", var_id = "ccode",
               var_time = "year", n = 12)
```

Negative values indicate clustering (positive spatial autocorrelation), and positive values indicate spatial dispersion. Figure 1a clearly depicts spatially correlated patterns of potential "peace consolidation,"
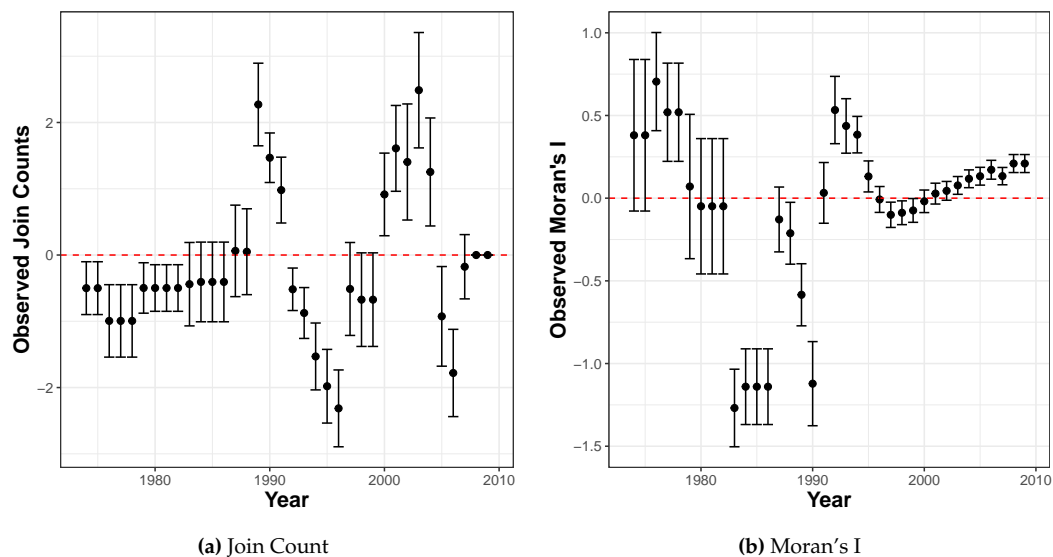
(a) Join Count

(b) Moran's I

Figure 1: Pre-estimation Spatial Autocorrelation Tests

though the direction of the autocorrelation varies over time.

The second function, `plot_Moran.I()`, implements the Global Moran's I test, which assesses the direction and degree of spatial autocorrelation in continuous or ordinal data (Moran, 1950; Paradis et al., 2020). We use this test to examine whether the duration of post-civil war peace among geographically proximate countries exhibits spatial autocorrelation (or dispersion). Like the previous function, `plot_Moran.I()` plots the Moran's I values based on user-defined adjacencies and, in this case, 90% confidence intervals for each year.

```
plot_Moran.I(data = walter[[1]], var_duration = "duration", var_id = "ccode",
             var_time = "year", n = 12)
```

Figure 1b reveals that the spatial relationship of post-war peace survival oscillates between positive autocorrelation (in the 1970s, early 1990s, and late 2000s) and spatial dispersion (in much of the 1980s and late 1990s). Taken together, the Join Count and Moran's I tests suggest that unobserved heterogeneous risk factors that transcend the borders of a single state may lead to spatial autocorrelation in both the consolidation and duration of post-war peace in Walter's (2015) data. This suggests that a Spatial SP survival model is an appropriate method of analysis, particularly if proximate countries share common unobservable risk factors that affect their odds of being immune to war recurrence or the time it takes for renewed war to occur.

Applying the Bayesian Spatial SP survival model

The `spatialSPsurv()` function fits the Bayesian Spatial SP survival model, which includes spatially autocorrelated frailties in the model's split and survival stage and can incorporate time-varying covariates. We illustrate the features of the `spatialSPsurv()` function by fitting the Bayesian Spatial *Weibull* cure model on Walter's (2015) data on post-war peace.[4] The log-logistic results are reported in Bolte et al. (2021a). We include the `lgdpl` variable in the estimated model's split-stage equation, which estimates the probability of units being in the "consolidated" post-war peace group (the "immune" fraction), though it is important to note that `atrisk` is used here as the dependent variable. We estimate the probability of post-war survival among the at-risk fraction as a function of UN peacekeeping (`unpko`), outright military victory in the previous war (`victory`), the resolution of the previous war via peace agreement (`comprehensive`), and, again, GDP/capita (`lgdpl`). The spatial relationship of each country is incorporated into the model via the spatial weights matrix generated from the `spatial_SA()` function described earlier.

The Spatial SP survival model is fit via Bayesian methods for inference. The MVN prior is used for the model's split-stage (γ) and survival-stage (β) parameters (with the following hyperparameters $s_\beta = I_{p1}$, $v_\beta = p1$, $s_\gamma = I_{p2}$, $v_\gamma = p2$). The Gamma prior is used for the hazard's shape parameter ρ (with hyperparameters $a\rho = 1$, $b\rho = 1$), and separate CAR priors—with the Gamma hyperprior assigned to λ for CAR (λ)[5]—incorporate the spatially autocorrelated frailties in the model's split-stage

---

[4]The results for all models were calculated in **R** version 4.0.3. on a PC with a 6-core i7 processor.
[5]With vague prior $(a_\lambda, b_\lambda) = (0.001, 1/0.001) = (0.001, 1000)$ as done for ρ

(V) and survival stage (W). Estimation proceeds via the MCMC algorithm described in the previous section. To this end, we specify the MCMC sampler as follows: the argument `N = 15,000` sets the number of MCMC iterations to 15,000, `burn = 5,000` discards the first 5,000 states of the Markov chain, `thin = 15` specifies the number of steps that are employed to prevent autocorrelation, and `m = 10` limits the steps in the slice sampling to 10. The argument `w = c(1,1,1)` specifies the default values of the size of the slice sampling for γ, β, and ρ.

For simplicity, we retain the default 0 as the initial value for all parameters with the `ini.beta`, `ini.gamma`, `ini.W`, and `ini.V` arguments. We incorporate separate proposal variances, namely `prop.varV = 1e-05` and `prop.varW = 1e-05`, for the split and survival stage frailties (**V**, **W**) and then assign separate Metropolis-Hastings proposal steps for estimating these parameters to optimize the acceptance rates. To specify the covariates in the model, `duration ~` precedes the survival-stage covariates, while `immune ~` precedes the split-stage variables. `Y0` indicates the time elapsed since inception until time t - 1, while `LY` is a dummy that captures the last observation year due to censoring or failure. `A = walter[[2]]` calls the spatial weights matrix from our list object. `S= 'sp_id'` (generated by `spatial_SA()`) gives the unit identifier in the spatial weights matrix that matches the units in the data frame. The argument `form` is used to specify the parametric survival distribution (which can be either "Weibull" or "loglog").

Importantly, the model functions in the **BayesSPsurv** package generate single chains for each parameter, but we provide a routine for estimating multiple chains in parallel on the GitHub repository for the package (https://github.com/Nicolas-Schmidt/BayesSPsurv). We first discuss the results and diagnostics for the single-chain results using the above MCMC specification and then illustrate the multiple-chain results and diagnostics for the Spatial SP survival model.

```
set.seed(123456)
model <-  spatialSPsurv(duration = duration ~ victory + comprehensive + lgdpl + unpko ,
                        immune = atrisk ~ lgdpl,
                        Y0 = 't.0',
                        LY = 'lastyear',
                        S = 'sp_id' ,
                        data = walter[[1]],
                        N = 15000,
                        burn = 5000,
                        thin = 15,
                        w = c(1,1,1),
                        m = 10,
                        ini.beta =  0,
                        ini.gamma = 0,
                        ini.W = 0,
                        ini.V= 0,
                        form = "Weibull",
                        prop.varV = 1e-05,
                        prop.varW = 1e-05,
                        A = walter[[2]])
```

The function automatically provides a progress bar for users to track the percent of the computation that has been completed. The elapsed run time for the Spatial SP Weibull specification being examined here was just under 23 minutes (calculated with `system.time()`). Once completed, the generic `print()` function displays the results[6]:

```
print(model)

Call:
spatialSPsurv(duration = duration ~ victory + comprehensive +
    lgdpl + unpko, immune = atrisk ~ lgdpl, Y0 = "t.0",
    LY = "lastyear", S = "sp_id", A = walter[[2]],
    data = walter[[1]], N = 15000, burn = 5000, thin = 15, w = c(1,
        1, 1), m = 10, ini.beta = 0, ini.gamma = 0, ini.W = 0,
    ini.V = 0, form = "Weibull", prop.varV = 1e-05, prop.varW = 1e-05)


Iterations = 1:666
Thinning interval = 1
Number of chains = 1
```

---

[6]95% Bayesian Credible Intervals are reported in Bolte et al. (2021a).

```
Sample size per chain = 666

Empirical mean and standard deviation for each variable,
plus standard error of the mean:


Duration equation:
                    Mean        SD   Naive SE Time-series SE
(Intercept)    0.3434823 1.0681878 0.041391437    0.078812901
victory        0.1028612 0.5321456 0.020620224    0.020620224
comprehensive  0.2036460 0.6178274 0.023940326    0.023940326
lgdpl          0.4831914 0.1424928 0.005521485    0.009230762
unpko          0.3057078 0.7695278 0.029818596    0.029818596

Immune equation:
                   Mean       SD Naive SE Time-series SE
(Intercept) -0.4306327 4.545655 0.1761405      0.3648687
lgdpl       -2.4722920 3.113144 0.1206319      0.1646276
```

The posterior mean estimates suggest that peace agreements, military victory, and peacekeeping units may increase the survival of post-civil war peace, though none of these parameters are highly reliable given their 95% BCIs, as reported in (Bolte et al., 2021a). However, lgdpl appears to reliably *increase* the survival of peace among at-risk cases while also reliably *decreasing* the probability of peace "consolidation" in the split-stage. Calling the SPstats() function calculates the Deviance Information Criterion (DIC) and log-likelihood (LL) statistics from the estimated model, where $DIC = -2 \times (L - P)$, L is the log-likelihood of the data given the posterior means of the covariate parameters, and P is the effective number of parameters in the model.

```
SPstats(model)

$DIC
[1] -1726.128

$Loglik
[1] 5301.714
```

The spatialSPsurv() function produces an object of class "mcmc", which is compatible with all standard summary and diagnostic methods for the single Markov chains in the **coda** package (Plummer et al., 2020). For instance, we can test for convergence with both the Geweke (1992) test using the geweke.diag() function and the Heidelberger and Welch (1983) stationarity test via the heidel.diag() function in **coda**. We report the Heidelberger and Welch (1983) stationarity test results in Bolte et al. (2021a) to save space. In this case, the Geweke tests indicate little evidence against convergence for each split-stage and survival-stage covariate, because the test statistics are reasonably close to zero.

```
geweke.diag(model$gammas)

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

(Intercept)       lgdpl
    -1.3755      0.4429

geweke.diag(model$betas)

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  (Intercept)      victory comprehensive        lgdpl         unpko
      -0.8248      -0.4271        1.8237       0.3601       -0.9471

geweke.diag(model$rho)

Fraction in 1st window = 0.1
```

```
Fraction in 2nd window = 0.5

  var1
0.3544
```

coda's plot() function can then be used to display trace and posterior density plots for the single Markov chains of the model's split-stage γ, survival-stage β, and ρ parameters (which we do not report here to save space). Users may also wish to validate MCMC convergence using multiple chains with different initial values for each parameter rather than only examining single-chain diagnostics. A routine for generating multiple chains in parallel using the **doParallel** (Wallig et al., 2020) and **doRNG** (Gaujoux, 2020) packages is available on the **BayesSPsurv** GitHub repository, though single chains with different starting values can also be run sequentially. The user can then collect them into a single mcmc.list object and easily plot the trace and density plots of the multiple chains. Using the code on GitHub, the Spatial SP survival model was re-run four times (to generate four chains) with 0, 1, 10, and 50 as the initial starting values for each of the estimated parameters (β, γ, **W**, and **V**). The trace and density plots for the multiple chains are reported below (we excluded the β intercept for aesthetic reasons), and the total run time was approximately 28 minutes for all four chains to be generated simultaneously.

```
par(mfrow=c(2,2))
plot(gammas, density = FALSE, auto.layout = FALSE, col =c (28,26,27,29))
plot(gammas, trace = FALSE, auto.layout = FALSE)

par(mfrow=c(2,4))
plot(betas[,2:5], density= FALSE, auto.layout = FALSE, col = c(28,26,27,29))
plot(betas[,2:5], trace= FALSE, auto.layout = FALSE)
```
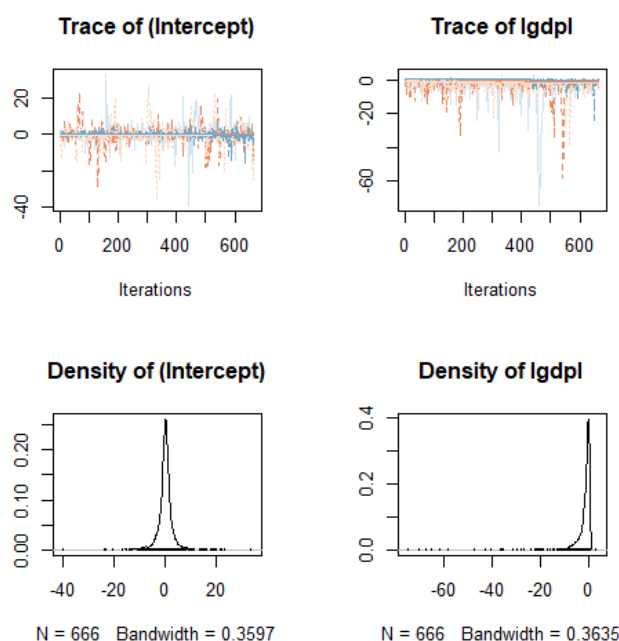

Figure 2: Posterior Densities of Bayesian Spatial Split-stage Parameters (γ)

The trace plots for the Spatial SP Weibull model's β and γ reveal decent mixing between the first three chains, though the fourth chain appears to take longer to converge. This is unsurprising given its extreme initial value. We can assess the convergence of the multiple chains using the Gelman-Rubin diagnostic, which compares the within-chain variance to the between-chain variance (Gelman and Rubin, 1992). If the difference in these variances is large, then the multiple chains likely have not converged to a proper stationary distribution. We can easily calculate the potential scale reduction factor (PSRF) with the gelman.diag() function in **coda**, which should be approximately 1 if all chains have converged to a common distribution.
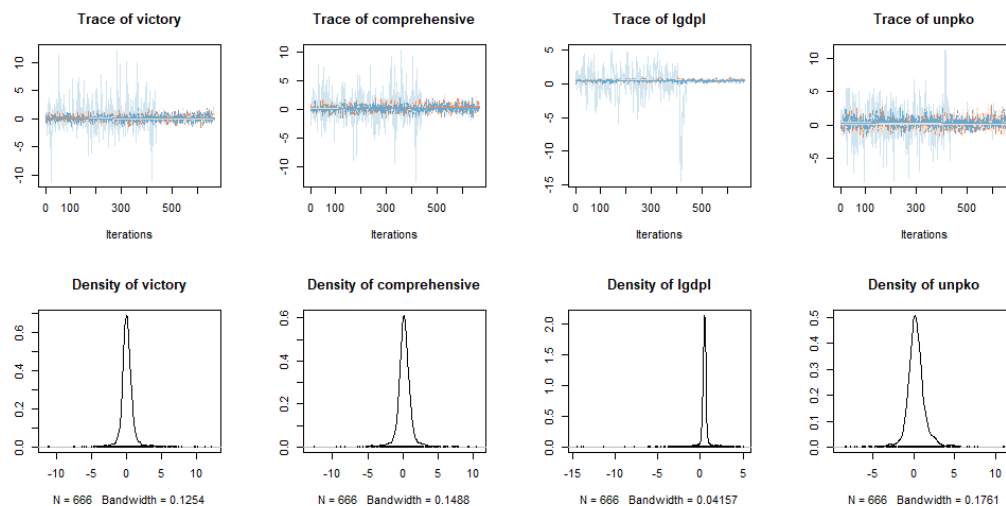
Figure 3: Posterior Densities of Bayesian Spatial Survival-stage Parameters (β)

```
Potential scale reduction factors:

              Point est. Upper C.I.
(Intercept)         1.19       1.38
victory             1.17       1.20
comprehensive       1.17       1.19
lgdpl               1.32       2.40
unpko               1.11       1.14

Multivariate psrf

1.05

Potential scale reduction factors:

            Point est. Upper C.I.
(Intercept)       1.05       1.07
lgdpl             1.14       1.33

Multivariate psrf

1.05
```

The multivariate PSRF is less than 1.1 in both stages, but the point estimates for all parameters are still insufficiently far from 1, suggesting that extending the chains would likely improve the accuracy of the estimates.

One way to substantively interpret the spatial frailties is to display a map and determine whether adjacent units share similar frailty values (e.g., Darmofal, 2009). The following code from the **countrycode** (Arel-Bundock, 2020) and **rworldmap** (South, 2016) packages permits users to create choropleth maps of the spatial frailty posterior means. Figures 4a-4b display the single-chain split-stage (V) and survival-stage (W) frailties from our Spatial SP Weibull model (the code generates Figure 4a; Figure 4b simply uses W in place of V.

```
library(rworldmap)
library(countrycode)
library(classInt)

spv  <- matrix(apply(model$V, 2, mean), ncol = 1, nrow = ncol(model$V))
ISO3 <- countrycode(colnames(model$V) ,'gwn','iso3c')
spv  <- data.frame(ccode = colnames(model$V), ISO3 = ISO3, spv = spv)
map  <- joinCountryData2Map(spv, joinCode = "ISO3", nameJoinColumn = "ISO3")
```

```
classInt <- classIntervals(map[["spv"]], n = 6, style = "quantile")
mapParams <- mapCountryData(map,
                           nameColumnToPlot = "spv",
                           addLegend = FALSE,
                           catMethod = classInt[["brks"]],
                           colourPalette = palette(RColorBrewer::brewer.pal(6,"RdBu")),
                           mapTitle = "")

do.call(addMapLegend, c(mapParams, legendLabels = "all", legendWidth = 0.5,
                                  legendIntervals = "data", legendMar = 2))
```
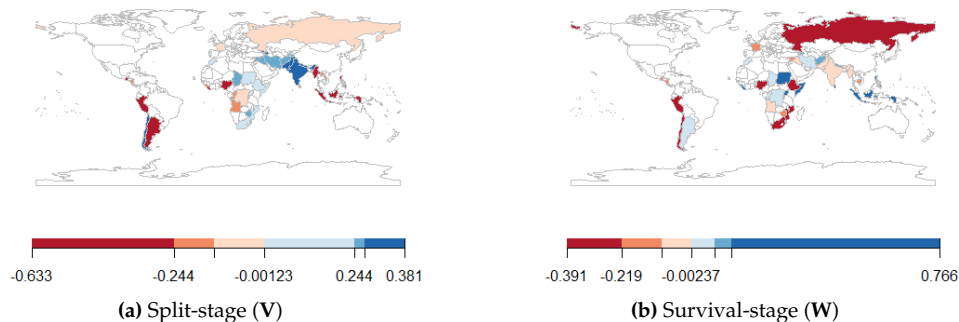

(a) Split-stage (**V**)   **(b)** Survival-stage (**W**)

Figure 4: Spatial Frailty Posterior Means

Note that in certain regions, there are distinct spatial bands in the (i) split-stage frailties, which range from -0.633 to 0.381 with a corresponding standard deviation of 0.34 and (ii) survival-stage frailties that range from -0.391 to 0.766 with a corresponding standard deviation of 0.32. These spatial bands reveal geographic clustering in both consolidation and duration of post-war peace since similar frailty values often occur near one another.

Applying the Bayesian Exchangeable and Pooled SP survival models

The **BayesSPsurv** package can also be used to estimate Bayesian Exchangeable and Pooled SP survival models. The Exchangeable model incorporates frailties that are assumed to be statistically independent in both stages, while the Pooled model excludes frailties altogether. We again use Walter's (2015) data on post-civil war peace to illustrate these models. Beginning with the Exchangeable Weibull SP survival model, we again assign the MVN prior to the model's γ and β parameters, the Gamma prior to ρ, and the same default settings for the hyperparameters ($s_\beta = I_{p1}$, $v_\beta = p1$, $s_\gamma = I_{p2}$, $v_\gamma = p2$, $a_\rho = 1$, $b_\rho = 1$). We also incorporate separate proposal variance variables for the nonspatial i.i.d frailties (V_i, W_i) and assign separate Metropolis-Hastings steps to estimate these parameters. Further, we use almost the same MCMC sampler specification as before. However, rather than incorporating spatial information via an adjacency matrix, we let id_WV=country to define the country-level nonspatial frailties.

```
set.seed(123456)
country  <- countrycode(unique(walter[[1]]$ccode),'gwn','iso3c')
model1 <- exchangeSPsurv(duration = duration ~ victory + comprehensive + lgdpl + unpko,
                         immune = atrisk ~ lgdpl,
                         Y0 = 't.0',
                         LY = 'lastyear',
                         S = 'sp_id' ,
                         data = walter[[1]],
                         N = 15000,
                         burn = 5000,
                         thin = 15,
                         w = c(1,1,1),
                         m = 10,
                         ini.beta =  0,
                         ini.gamma = 0,
```

```
                          ini.W = 0,
                          ini.V= 0,
                          form = "Weibull",
                          prop.varV = 1e-05,
                          prop.varW = 1e-05,
                          id_WV=country)
```

Again, users can assess either single or multiple (parallel) Markov chains with different starting parameter values when fitting the Exchangeable SP survival model. We report the single-chain results here for simplicity. The run time for the above specification was approximately 16 minutes and 30 seconds. The single-chain results and model fit statistics are obtained with the `print()` and `SPstats()` functions.

```
print(model1)

Call:
exchangeSPsurv(duration = duration ~ victory + comprehensive +
    lgdpl + unpko, immune = atrisk ~ lgdpl, Y0 = "t.0",
    LY = "lastyear", S = "sp_id", data = walter[[1]],
    N = 15000, burn = 5000, thin = 15, w = c(1, 1, 1), m = 10,
    ini.beta = 0, ini.gamma = 0, ini.W = 0, ini.V = 0, form = "Weibull",
    prop.varV = 1e-05, prop.varW = 1e-05, id_WV = country)


Iterations = 1:666
Thinning interval = 1
Number of chains = 1
Sample size per chain = 666

Empirical mean and standard deviation for each variable,
plus standard error of the mean:


Duration equation:
                    Mean        SD    Naive SE Time-series SE
(Intercept)   0.31054460 1.0938158 0.042384501     0.08335381
victory       0.14520456 0.5328378 0.020647044     0.02064704
comprehensive 0.09721372 0.6374526 0.024700786     0.02470079
lgdpl         0.48473536 0.1455731 0.005640844     0.01061988
unpko         0.44731384 0.7421925 0.028759376     0.02875938

Immune equation:
                 Mean       SD Naive SE Time-series SE
(Intercept) -0.8993176 5.924550 0.2295716      0.4969172
lgdpl       -3.7286951 5.846367 0.2265421      0.5271203

SPstats(model1)
$DIC
[1] 3549.075

$Loglik
[1] 7506.608
```

Based on these results, `lgdpl` is still a highly reliable predictor of longer peace duration among conflicts at risk of recurrence, while the other survival stage covariates remain positive but statistically unreliable. In contrast to the spatial frailty model, however, the effect of `lgdpl` on the probability of peace consolidation is now statistically unreliable. Users interested in displaying and interpreting the exchangeable frailties can do so in a variety of ways, but if these results are being compared to those of a spatial frailty model, we recommend mapping the exchangeable frailty values and comparing the clustering (or lack thereof) of the i.i.d. frailty terms to those produced from the spatial model (Darmofal, 2009).

Finally, the `pooledSPsurv()` function fits the Pooled Bayesian SP survival model. For this application, we again assign the MVN prior to the model's γ and β parameters, the Gamma prior to ρ, and the same default settings for the hyperparameters described earlier.

```
set.seed(123456)
model2 <- pooledSPsurv(duration = duration ~ victory + comprehensive + lgdpl + unpko,
                       immune = atrisk ~ lgdpl,
                       Y0 = 't.0',
                       LY = 'lastyear',
                       data = walter[[1]],
                       N = 15000,
                       burn = 5000,
                       thin = 15,
                       w = c(1,1,1),
                       m = 10,
                       ini.beta =  0,
                       ini.gamma = 0,
                       form = "Weibull")
```

Note that the MCMC sampler is almost identical to that of the previous two models, but we now
exclude the separate Metropolis-Hastings proposal step for the frailties in the MCMC algorithm as
well as the initial values for the frailty estimates since this is a non-frailty model. The run time for the
Pooled model was approximately 10 minutes. The single-chain output is again easily displayed.

```
print(model2)

Call:
pooledSPsurv(duration = duration ~ victory + comprehensive +
    lgdpl + unpko, immune = atrisk ~ lgdpl, Y0 = "t.0",
    LY = "lastyear", data = walter[[1]], N = 15000, burn = 5000,
    thin = 15, w = c(1, 1, 1), m = 10, ini.beta = 0, ini.gamma = 0,
    form = "Weibull")


Iterations = 1:666
Thinning interval = 1
Number of chains = 1
Sample size per chain = 666

Empirical mean and standard deviation for each variable,
plus standard error of the mean:


Duration equation:
                 Mean        SD   Naive SE Time-series SE
(Intercept)   0.3014335 1.0572594 0.040967971    0.07197438
victory       0.1060453 0.5485730 0.021256770    0.02125677
comprehensive 0.1900127 0.6054421 0.023460406    0.02481878
lgdpl         0.4647260 0.1552465 0.006015679    0.01166616
unpko         0.2692275 0.7740804 0.029995006    0.02999501

Immune equation:
               Mean       SD  Naive SE Time-series SE
(Intercept) -1.213902 9.617551 0.3726725      1.3195278
lgdpl       -2.896038 4.939374 0.1913968      0.5267751


SPstats(model2)

$DIC
[1] 1804.527

$Loglik
[1] 6229.248
```

For both the Exchangeable and Pooled SP survival models, users can again employ the code
and routines available in our GitHub repository to view the trace plots from the single or multiple
Markov chains associated with the posterior densities for the γ, β, and ρ parameters and conduct
any relevant convergence diagnostics. The plots and convergence test statistics for the models

presented here are reported in Bolte et al. (2021a) to save space. Additional information is available at https://github.com/Nicolas-Schmidt/BayesSPsurv.

5 Conclusion

Survival data often include two populations with different underlying risk propensities: the immune fraction of right-censored units that will never experience the event of interest and the at-risk fraction of units that have or will. Numerous R-packages such as **smcure**, **nltm**, and **spduration** have been developed to fit parametric or semi-parametric cure models for nonspatial survival data, but none of these packages allow the user to account for spatial autocorrelation in both stages. A separate set of packages allows the user to include spatially weighted frailties in conventional survival models (Taylor and Rowlingson, 2017; Zhou et al., 2020), but spatial autocorrelation may also influence the probability of immunity from an event of interest. The **BayesSPsurv** package addresses this lacuna by offering a suite of functions to fit Bayesian SP survival models. Specifically, users can estimate Bayesian Pooled, Exchangeable, and Spatial frailty SP survival models with time-varying covariates, specify their own spatial weights matrices, and easily examine diagnostic statistics. The applied potential of the Spatial SP survival model, in particular, is considerable, as researchers studying anything from the survival of cancer patients to the survival of political regimes may have a methodological need to model spatial autocorrelation in the immune fraction.

Future work can build upon the spatial and nonspatial cure models presented here to develop estimation routines for survival data with multiple stages, competing risks, or recurrent events. Moreover, although **BayesSPsurv** is the first to allow spatially autocorrelated frailties in both stages, future work can extend the frailty models in our package by employing a similar approach as Yin (2005) to incorporate shared frailties with factor loadings to account for any correlation between the frailty terms (**W** and **V**) in either stage. Subsequent iterations of the package will include alternative baseline hazards based on penalized splines or nonparametric Gaussian processes as well as the option to use a semiparametric Cox model with exchangeable or spatial frailties. Implementing these future developments with an external MCMC sampling dependency built on C++ subroutines like STAN would be useful, given the Hamiltonian Monte Carlo algorithm's efficiency in sampling from high dimensional posterior distributions. We also plan to extend **BayesSPsurv** to accommodate left and interval censoring and delayed entry.

Bibliography

J. Amdahl. *flexsurvcure: Flexible Parametric Cure Models*, 2020. URL https://CRAN.R-project.org/package=flexsurvcure. R package version 1.2.0. [p596]

V. Arel-Bundock. *countrycode: Convert Country Names and Country Codes*, 2020. URL https://CRAN.R-project.org/package=countrycode. R package version 1.2.0. [p606]

S. Banerjee and B. P. Carlin. Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics*, 60(1):268–275, 2004. URL https://doi.org/10.1111/j.0006-341X.2004.00032.x. [p595, 596, 598, 599]

S. Banerjee, M. M. Wall, and B. P. Carlin. Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, 4(1):123–142, 2003. URL https://doi.org/10.1093/biostatistics/4.1.123. [p595, 596, 598]

S. Banerjee, A. Gelfand, J. R. Knight, and C. Sirmans. Spatial modeling of house prices using normalized distance-weighted sums of stationary processes. *Journal of Business & Economic Statistics*, 22(2):206–213, 2004. URL https://doi.org/10.1198/073500104000000091. [p596]

A. Beger, D. W. Hill, N. W. Metternich, S. Minhas, and M. D. Ward. Splitting it up: the spduration split-population duration regression package for time-varying covariates. *The R Journal*, 9(2):474–486, 2017. URL https://doi.org/10.32614/RJ-2017-056. [p595, 596, 597, 600]

A. Beger, D. Chiba, D. W. Hill, Jr., N. W. Metternich, S. Minhas, and M. D. Ward. *spduration: Split-Population Duration (Cure) Regression*, 2018. URL https://CRAN.R-project.org/package=spduration. R package version 0.17.1. [p595, 596]

C. Belitz, A. Brezger, T. Kneib, S. Lang, and N. Umlauf. *BayesX: Software for Bayesian Inference in Structured Additive Regression Models*, 2017. URL http://www.BayesX.org/. Version 1.1. [p596]

L. Bernardinelli and C. Montomoli. Empirical bayes versus fully bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 11(8):983–1007, 1992. URL http://dx.doi.org/10.1002/sim.4780110802. [p598]

L. Bernardinelli, D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini. Bayesian analysis of space—time variation in disease risk. *Statistics in medicine*, 14(21-22):2433–2443, 1995. URL http://dx.doi.org/10.1002/sim.4780142112. [p596]

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991. URL https://doi.org/10.1007/BF00116466. [p596, 598]

B. Bolte, N. Huynh, B. Mukherjee, S. Béjar, and N. Schmidt. Bayesian split-population survival models for the social sciences (version 2). Available at SSRN, 2021a. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3765112. [p599, 600, 602, 603, 604, 610]

B. Bolte, N. Schmidt, S. Bejar, B. Mukherjee, M. M. Joo, and N. K. Huynh. *BayesSPsurv: Bayesian Spatial Split Population Survival Model*, 2021b. URL https://CRAN.R-project.org/package=BayesSPsurv. R package version 0.1.3. [p595]

J. M. Box-Steffensmeier and B. S. Jones. *Event History Analysis*. Cambridge University Press, New York, 2004. [p595, 596]

J. M. Box-Steffensmeier and C. Zorn. Modeling heterogeneity in duration models. In *Summer Meeting of the Political Methodology Society, July15-17*, 1999. [p595, 596]

C. Cai, Y. Zou, Y. Peng, and J. Zhang. *smcure: Fit Semiparametric Mixture Cure Models*, 2012. URL https://CRAN.R-project.org/package=smcure. R package version 2.0. [p595, 596]

B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis, 2nd Edition*. New York: Chapman and Hall/CRC, 2000. [p599]

A. D. Cliff and J. K. Ord. *Spatial processes: models & applications*. Taylor & Francis, 1981. [p601]

D. Darmofal. Bayesian spatial survival models for political event processes. *American Journal of Political Science*, 53(1):241–257, 2009. URL https://doi.org/10.1111/j.1540-5907.2008.00368.x. [p595, 596, 598, 599, 606, 608]

U. Diva, D. K. Dey, and S. Banerjee. Parametric models for spatially correlated survival data for individuals with multiple cancers. *Statistics in medicine*, 27(12):2127–2144, 2008. URL https://doi.org/10.1002/sim.3141. [p596]

D. Eddelbuettel, R. Francois, J. Allaire, K. Ushey, Q. Kou, N. Russell, D. Bates, and J. Chambers. *Rcpp: Seamless R and C++ Integration*, 2020. URL https://CRAN.R-project.org/package=Rcpp. R package version 1.0.5. [p600]

G. Garibotti, A. Tsodikov, and M. Clements. *nltm: Non-Linear Transformation Models*, 2019. URL https://CRAN.R-project.org/package=nltm. R package version 1.4.2. [p596]

R. Gaujoux. *doRNG: Generic Reproducible Parallel Backend for 'foreach' Loops*, 2020. URL https://cran.r-project.org/web/packages/doRNG/index.html. R package version 1.8.2. [p605]

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511, 1992. [p605]

J. F. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. Bernardo, J. Berger, A. Dawid, and A. Smith, editors, *Bayesian Statistics*. Oxford: Clarendon Press, 1992. URL https://doi.org/10.21034/sr.148. [p604]

P. Heidelberger and P. D. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 3(6):1109–1144, 1983. URL https://doi.org/10.1287/opre.31.6.1109. [p604]

R. Henderson, S. Shimakura, and D. Gorst. Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association*, 97(460):965–972, 2002. URL https://doi.org/10.1198/016214502388618753. [p595, 596]

J. G. Ibrahim, M.-H. Chen, and D. Sinha. Bayesian semiparametric models for survival data with a cure fraction. *Biometrics*, 57(2):383–388, 2001. URL http://dx.doi.org/10.1111/j.0006-341X.2001.00383.x. [p598]

Y. Li and L. Ryan. Modeling spatial survival data using semiparametric frailty models. *Biometrics*, 58 (2):287–297, 2002. URL https://doi.org/10.1111/j.0006-341X.2002.00287.x. [p595]

W. Lu. Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*, 20(2):661–674, 2010. [p597, 598]

R. A. Maller and X. Zhou. *Survival analysis with long-term survivors*. John Wiley & Sons, New York, 1996. [p595, 597]

P. A. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950. URL http://dx.doi.org/10.1093/biomet/37.1-2.17. [p602]

R. M. Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003. URL http://dx.doi.org/10.1214/aos/1056562461. [p599]

E. Paradis, S. Blomberg, B. Bolker, J. Brown, S. Claramunt, J. Claude, H. S. Cuong, R. Desper, G. Didier, B. Durand, J. Dutheil, R. Ewing, O. Gascuel, T. Guillerme, C. Heibl, A. Ives, B. Jones, F. Krah, D. Lawson, V. Lefort, P. Legendre, J. Lemon, G. Louvel, E. Marcon, R. McCloskey, J. Nylander, R. Opgen-Rhein, A.-A. Popescu, M. Royer-Carenzi, K. Schliep, K. Strimmer, and D. de Vienne. *ape: Analyses of Phylogenetics and Evolution*, 2020. URL https://CRAN.R-project.org/package=ape. R package version 5.4-1. [p602]

V. Patilea and I. Van Keilegom. A general approach for cure models in survival analysis. *Annals of Statistics*, 48(4):2323–2346, 2020. URL https://doi.org/10.1214/19-AOS1889. [p596]

Y. Peng and J. M. Taylor. Mixture cure model with random effects for the analysis of a multi-center tonsil cancer study. *Statistics in medicine*, 30(3):211–223, 2011. URL https://doi.org/10.1002/sim.4098. [p596]

Y. Peng and J. M. Taylor. Cure models. *Handbook of survival analysis*, pages 113–134, 2014. URL https://doi.org/10.1201/b16248-8. [p595, 597]

M. Plummer, N. Best, K. Cowles, K. Vines, D. Sarkar, D. Bates, R. Almond, and A. Magnusson. *coda: Output Analysis and Diagnostics for MCMC*, 2020. URL https://CRAN.R-project.org/package=coda. R package version 0.19-4. [p600, 604]

R. A. Ray, R. W. Perry, N. A. Som, and J. L. Bartholomew. Using cure models for analyzing the influence of pathogens on salmon survival. *Transactions of the American Fisheries Society*, 143(2):387–398, 2014. URL https://doi.org/10.1080/00028487.2013.862183. [p596]

P. Schmidt and A. D. Witte. Predicting criminal recidivism using 'split population'survival time models. *Journal of Econometrics*, 40(1):141–159, 1989. URL https://doi.org/10.1016/0304-4076(89)90034-1. [p596]

A. South. *rworldmap: Mapping Global Data*, 2016. URL https://CRAN.R-project.org/package=rworldmap. R package version 1.3-6. [p606]

B. Taylor and B. Rowlingson. spatsurv: an r package for bayesian inference with spatial survival models. *Journal of Statistical Software*, 77(4):1–32, 2017. URL https://doi.org/10.18637/jss.v077.i04. [p595, 596, 610]

B. M. Taylor and B. S. Rowlingson. *spatsurv: Bayesian Spatial Survival Analysis with Parametric Proportional Hazards Models*, 2020. URL https://CRAN.R-project.org/package=spatsurv. R package version 1.5. [p596]

T. M. Therneau. *A Package for Survival Analysis in R*, 2020. URL https://CRAN.R-project.org/package=survival. R package version 3.1-12. [p596]

A. Thomas, N. Best, D. Lunn, D. Arnold, and D. Spiegelhalter. Geobugs version 1.2 user manual. *MRC Biostatistics Unit*, 2004. URL https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/geobugs12manual.pdf. [p596, 598]

N. Umlauf, T. Kneib, and N. Klein. *BayesX: R Utilities Accompanying the Software Package BayesX*, 2019. URL https://CRAN.R-project.org/package=BayesX. R package version 0.3-1. [p596]

M. Wallig, Microsoft Corporation, S. Weston, and D. Tenenbaum. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2020. URL https://cran.r-project.org/web/packages/doParallel/index.html. R package version 1.0.16. [p605]

B. F. Walter. Why bad governance leads to repeat civil war. *Journal of Conflict Resolution*, 59(7):1242–1272, 2015. URL https://doi.org/10.1177/0022002714528006. [p600, 602, 607]

W. Wang, M.-H. Chen, X. Wang, and J. Yan. *dynsurv: Dynamic Models for Survival Data*, 2020. URL https://CRAN.R-project.org/package=dynsurv. R package version 0.4-2. [p596]

Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005. URL https://doi.org/10.1016/S0140-6736(05)17947-1. [p596]

G. Yin. Bayesian cure rate frailty models with application to a root canal therapy study. *Biometrics*, 61 (2):552–558, 2005. [p610]

G. Yin and J. G. Ibrahim. Cure rate models: a unified approach. *Canadian Journal of Statistics*, 33(4): 559–570, 2005. URL http://dx.doi.org/10.1002/cjs.5550330407. [p597]

H. Zhou and T. Hanson. *spBayesSurv: Bayesian Modeling and Analysis of Spatially Correlated Survival Data*, 2020. URL https://CRAN.R-project.org/package=spBayesSurv. R package version 1.1.4. [p596]

H. Zhou, T. Hanson, and J. Zhang. spBayesSurv: Fitting Bayesian spatial survival models using R. *Journal of Statistical Software*, 92(9):1–33, 2020. URL http://dx.doi.org/10.18637/jss.v092.i09. [p595, 596, 610]

Brandon Bolte
Department of Political Science
Penn State University
USA
blb72@psu.edu

Nicolás Schmidt
Department of Political Science
Universidad de la República
Uruguay
nschmidt@cienciassociales.edu.uy

Sergio Béjar
Department of Political Science
San José State University
USA
sergio.bejar@sjsu.edu

Nguyen Huynh
Department of Political Science
Penn State University
USA
nkh8@psu.edu

Bumba Mukherjee
Department of Political Science
Penn State University
USA
sxm73@psu.edu