A Shared-Frailty Spatial Scan Statistic Model for Time-to-Event Data.

Camille Frevent, Mohamed-Salem Ahmed, Sophie Dabo-Niang, Michaël Genin

Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

A Shared-Frailty Spatial Scan Statistic Model for Time-to-Event Data

Camille Frévent[1] | Mohamed-Salem Ahmed[1,2] | Sophie Dabo-Niang[3,4] | Michaël Genin[1]

[1]Université de Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, Université de Lille, Lille, France | [2]Alicante SARL, Lesquin, France | [3]CNRS, UMR 8524 - Laboratoire Paul Painlevé, Université de Lille, Lille, France | [4]MODAL team, INRIA Lille-Nord Europe, Villeneuve-d'Ascq, France

Correspondence: Camille Frévent (camille.frevent@univ-lille.fr)

Received: 20 July 2023 | **Revised:** 24 January 2024 | **Accepted:** 4 May 2024

Keywords: conditional autoregressive model | shared frailty model | spatial scan statistics | time-to-event data

ABSTRACT

Spatial scan statistics are well-known methods widely used to detect spatial clusters of events. Furthermore, several spatial scan statistics models have been applied to the spatial analysis of time-to-event data. However, these models do not take account of potential correlations between the observations of individuals within the same spatial unit or potential spatial dependence between spatial units. To overcome this problem, we have developed a scan statistic based on a Cox model with shared frailty and that takes account of the spatial dependence between spatial units. In simulation studies, we found that (i) conventional models of spatial scan statistics for time-to-event data fail to maintain the type I error in the presence of a correlation between the observations of individuals within the same spatial unit and (ii) our model performed well in the presence of such correlation and spatial dependence. We have applied our method to epidemiological data and the detection of spatial clusters of mortality in patients with end-stage renal disease in northern France.

1 | Introduction

In many applications, researchers look for unusual spatial aggregations (clusters) of data. In the field of public health, epidemiologists seek to identify the presence (within a geographical area) of spatial clusters in which the risk of disease is unusually high (or low); this makes it possible to (i) formulate hypotheses to guide etiological research and (ii) implement localized public health policies. By way of another example, researchers in the environmental sciences may be interested in determining the presence of environmental black spots defined by particularly unusual pollutant concentrations in a specific area—thus leading to local actions to prevent or solve the problem.

Spatial scan statistics are widely used to detect statistically significant spatial clusters with a scanning window and without preselection bias. These methods were introduced by Kulldorff and Nagarwalla (1995) and Kulldorff (1997) in the cases of Bernoulli and Poisson models, respectively. Since then, the scan statistics approach has been extended to many other spatial data distributions. In a univariate framework, for example, Gaussian (Kulldorff, Huang, and Konty 2009), ordinal (Jung, Kulldorff, and Klassen 2007), zero-inflated (Cançado, da Silva, and da Silva 2014; Cançado, Fernandes, and da Silva 2017; de Lima et al. 2015), and Poisson with overdispersion (de Lima et al. 2015; Zhang, Zhang, and Lin 2012) models have been developed. Similarly, in the context of multivariate or functional data, several spatial scan statistics have been developed (Cucala et al. 2017; Frévent et al. 2021; Frévent et al. 2023; Kulldorff et al. 2007; Neill and Cooper 2010; Smida et al. 2022). The reader is referred to Abolhassani and Prates (2021) for a comprehensive review of spatial scan statistics. These methods have been widely applied in many fields, such as epidemiology (Genin et al. 2020; Green et al. 2006; Khan et al. 2021; Marciano et al. 2018), environmental science (Shi, Liu,

and Zhong 2022; Wan et al. 2020), oncology (Leiser et al. 2020), criminology (Minamisava et al. 2009), and astronomy (De La Fuente Marcos and De La Fuente Marcos 2008).

In the field of spatial epidemiology, the study of the spatial distribution of time-to-event data can identify areas in which the survival time of patients is different from the rest of the geographical area (i.e., the survival time is longer or shorter). From an epidemiological point of view, the identification of these areas of unusual survival time is particularly useful for identifying local risk factors that condition survival. Moreover, this information can help public health decision-makers to develop and implement targeted, specific local policies. In the context of spatial cluster detection in time-to-event data, Huang, Kulldorff, and Gregorio (2007) and Bhatt and Tiwari (2014) developed spatial scan statistics based on an exponential model and a Weibull model, respectively. More recently, Usman and Rosychuk (2018) developed a parametric model considering a log-Weibull distribution. Although these methods are widely used in practice to detect spatial clusters of time-to-event data (Gregorio et al. 2007; Henry, Niu, and Boscoe 2009; Wan et al. 2012), they are totally parametric. The first semiparametric method (using a Cox model) was developed by Cook, Gold, and Li (2007).

Unlike other spatial scan statistics models, the above-mentioned exponential, Weibull, log-Weibull, and Cox models consider data measured at the individual level. However, in health data studies, the patient's exact geographic location is rarely known (e.g., for reasons of anonymity), and patients are located through an administrative spatial unit (e.g., municipalities). In this context, the above-mentioned methods are based on the strong assumption of independence between observations—a classical assumption in the field of spatial scan statistics. This assumption is associated with two major drawbacks. First, the methods do not take account of the potential correlation between the observations of individuals within the same spatial unit, namely the intraspatial unit correlation. The latter can be induced by characteristics of the spatial units that have not been measured in the study (e.g., healthcare supply) but that affect the patients' survival (Austin 2017). Second, the methods do not take account of potential spatial dependence between spatial units. However, one can logically expect geographically close units to be more strongly related than distant ones (Li 2009). Furthermore, it has been shown that ignoring spatial dependence when using spatial scan statistics leads to a significant increase in the type I error (Loh and Zhu 2007). Since then, several researchers have developed methods that take spatial dependence into account (Ahmed, Cucala, and Genin 2021; Lee, Sun, and Chang 2020; Lin 2014; Loh and Zhu 2007). However, none of these methods were designed for time-to-event data and the adjustment for intraspatial unit correlations.

In the analysis of time-to-event data, various models have been developed to take account of unobserved factors common to groups of individuals; for example, members of the same family share genetic factors and patients in the same hospital often receive much the same care. One way of taking this intragroup homogeneity into account involves introducing a random effect common to all individuals in a group, namely shared frailty (Clayton 1978; Hougaard 2000; Liang et al. 1995). The shared frailties are assumed to be independent between groups (Liang et al. 1995). However, when the groups correspond to spatial units, this assumption is unrealistic because close spatial units tend to be related (Arlinghaus 1995). To this end, Li and Ryan (2002) extended shared frailty models to the case of spatially correlated frailty, which take account of not only intraspatial unit correlation but also possible spatial dependence between spatial units. However, although this approach has been widely applied (Aswi et al. 2020; Banerjee, Wall, and Carlin 2003; Ojiambo and Kang 2013), it has never been investigated in the field of spatial scan statistics.

Here, we present a new spatial scan statistic for time-to-event data based on a semiparametric Cox model with spatially correlated shared frailties. Section 2 describes the methodological aspects of the scan statistic model. Section 3 presents both the design and the results of simulation studies evaluating (i) the performance of conventional methods on datasets with intraspatial unit correlation and (ii) the performance of our approach on datasets with both intraspatial unit correlation and spatial dependence between spatial units. Section 4 describes the application of our method to epidemiological data and the detection of spatial clusters of mortality in patients with end-stage renal disease in northern France. Lastly, we discuss results in Section 5.

2 | Methodology

2.1 | General Principle

Let us consider K nonoverlapping spatial locations $s_1, \dots, s_k, \dots, s_K$ of an observation domain $S \subset \mathbb{R}^2$ and let $i_1^{(k)}, \dots, i_n^{(k)}, \dots, i_{N_k}^{(k)}$ be N_k individuals at spatial location s_k. The total number of individuals in S is defined as $N = \sum_{k=1}^{K} N_k$. Here, we are interested in the time-to-event data measured on individuals: $T_{i_n^{(k)}}$ and $\delta_{i_n^{(k)}}$ are, respectively, the observation time of the i_nth individual in spatial location s_k and the censoring indicator, which is equal to 0 if the individual $i_n^{(k)}$ is censored and 1 otherwise. In the following, we only considered the cases of right censoring (i.e., the event of interest could not have occurred before the beginning of the study). Censoring was assumed to be uninformative, and the event times were assumed to be independent of the censoring times.

We sought to test for the presence of spatial clusters in which individuals have shorter (or longer) survival times than other individuals in the rest of S. In this context, spatial scan statistics are designed to detect spatial clusters and to test their statistical significance by testing a null hypothesis $\mathcal{H}_0$ (the absence of a cluster) against a composite alternative hypothesis $\mathcal{H}_1$ (the presence of at least one cluster $w \subset S$ presenting abnormal time-to-event values). According to Cressie (1977), a spatial scan statistic is the maximum of a concentration index over a set of potential clusters $\mathcal{W}$. In the following and without loss of generality, we focused on variable-size circular clusters. Hence, in line with Kulldorff (1997), the set of potential circular clusters $\mathcal{W}$ can be defined as $\mathcal{W} = \{w_{k,l} / 1 \leq |w_{k,l}| \leq \frac{N}{2}, 1 \leq k, l \leq K\}$, where $w_{k,l}$ is the disk centered on s_k that passes through s_l and $|w_{k,l}|$ is the number of individuals in $w_{k,l}$: a cluster comprises at most 50% of the study population (i.e., $N/2$) (Kulldorff and Nagarwalla 1995). It should be noted that other cluster shapes have been described in the literature, such as elliptical clusters (Kulldorff et al. 2006),

rectangular clusters (Chen and Glaz 2009), or arbitrarily shaped clusters (Tango and Takahashi 2005; Yin and Mu 2018; Zhou, Shu, and Su 2015).

2.2 | The Model

We assume that the instantaneous hazard rate at time t for the individual $i_n^{(k)}$ is

$$\lambda_{i_n^{(k)}}(t|\boldsymbol{Z}_{i_n^{(k)}}, \varphi_k) = \lambda_0(t)\exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_{i_n^{(k)}} + \varphi_k\right],$$

where $\boldsymbol{Z}_{i_n^{(k)}} = (Z_{i_n^{(k)},1}, \ldots, Z_{i_n^{(k)},p})^\top$ is a vector of p covariates associated with the individual $i_n^{(k)}$, and φ_k is the shared frailty associated with the spatial location s_k. The presence of a spatial cluster in the data results in an effect on the survival times in the spatial units involved. Hence, the effect of this cluster has been incorporated within the shared frailty: for each potential cluster w, φ_k can be decomposed into α_w (a cluster fixed effect) and X_k (effect specific to the spatial location s_k). Thus, the shared frailties φ_k associated with the potential cluster w can be rewritten as $\varphi_k^{(w)} = \alpha_w \mathbb{1}_{s_k \in w} + X_k$, where $\mathbb{E}[X_k] = 0$. In this context, the test hypotheses can be rewritten as $\mathcal{H}_0 : \forall w \in \mathcal{W}, \alpha_w = 0$ (the absence of a cluster), and the alternative hypothesis associated with the potential cluster w is $\mathcal{H}_1^{(w)} : \alpha_w \neq 0$ (the presence of a cluster w, in which the individuals present atypical survival times).

Moreover, the spatial nature of the data requires one to take account of a possible spatial dependence between the spatial locations s_k, and thus between the X_k. This makes it possible to distinguish the effect of the cluster from the spatial correlation of unobserved factors on the scale of the spatial unit. Thus, we considered the conditional autoregressive (CAR) model developed by Leroux, Lei, and Breslow (2000) for the distribution of the X_k:

$$X_k|X_{-k} \sim \mathcal{N}\left(\frac{\rho\sum_{l=1}^{K} \upsilon_{k,l}X_l}{\rho\sum_{l=1}^{K}\upsilon_{k,l} + 1 - \rho}, \frac{\sigma_X^2}{\rho\sum_{l=1}^{K}\upsilon_{k,l} + 1 - \rho}\right),$$

where $X_{-k} = \{X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_K\}$, $\upsilon_{k,l} = 1$ if s_k and s_l are adjacent (i.e., they share a common boundary) and 0 if not, and $\rho \in [0,1]$ is the spatial dependence parameter. It should be noted that in the absence of a spatial dependence, the X_k are independent and identically distributed (i.i.d.) according to a normal distribution $\mathcal{N}(0, \sigma_X^2)$. Conversely, if $\rho = 1$ (i.e., complete spatial dependence between the spatial units), the X_k is distributed according to an intrinsic CAR (ICAR) model (Besag, York, and Mollié 1991).

The method comprises two steps. The first step (Section 2.2.1) consists of estimating the shared frailties φ_k and their spatial dependence parameter ρ. In a second step (Section 2.2.2), a scan procedure is developed and applied to the estimated shared frailties in order to identify clusters of spatial units in which the φ_k are significantly higher (corresponding to a higher risk) or significantly lower (corresponding to a lower risk) than elsewhere. Lastly, the procedure for determining the statistical significance of the identified spatial clusters is described in Section 2.2.3.

2.2.1 | Estimation of the φ_k and ρ

This first step consists of estimating the φ_k and ρ in a Bayesian framework by using the integrated nested Laplace approximation (INLA) (see Rue, Martino, and Chopin 2009 for details).

The φ_k are considered under both the $\mathcal{H}_0$ and $\mathcal{H}_1$ hypotheses. However, it should be noted that (i) neither X_k nor ρ depends on the clustering assumptions since they depend only on the spatial structure of the data and (ii) only a single vector of φ_k needs to be estimated in order to best fit the observed data. Therefore, the φ_k must be estimated under the true hypothesis among $\mathcal{H}_0$ and the set of alternative hypotheses $\mathcal{H}_1^{(w)}$, that is, the hypothesis under which the observations have been generated. In line with the approach developed by Ahmed, Cucala, and Genin (2021), we need to determine the "best model" among the candidate hypotheses ($\mathcal{H}_0$ and $\mathcal{H}_1^{(w)}$). To this end, for each potential cluster $w \in \mathcal{W}$, we considered the Bayes factor $\mathrm{BF}^{(w)}$, defined as the marginal likelihood ratio between the model under $\mathcal{H}_1^{(w)}$ ($\mathcal{M}_1^{(w)}$) and the model under $\mathcal{H}_0$ ($\mathcal{M}_0$):

$$\mathrm{BF}^{(w)} = \frac{\mathbb{P}\left[\left\{T_{i_n^{(k)}}, \delta_{i_n^{(k)}}, \boldsymbol{Z}_{i_n^{(k)}}, \mathbb{1}_{i_n^{(k)}\in w}\right\}\middle|\mathcal{M}_1^{(w)}\right]}{\mathbb{P}\left[\left\{T_{i_n^{(k)}}, \delta_{i_n^{(k)}}, \boldsymbol{Z}_{i_n^{(k)}}\right\}\middle|\mathcal{M}_0\right]}.$$

It should be noted that if adjustment of cluster detection on covariates is required, adjustment is performed at this stage: We consider the approach proposed by Jung (2009) and Ahmed and Genin (2020), which consists of estimating the coefficients associated with the covariates in $\mathcal{M}_0$ and then setting in each model $\mathcal{M}_1^{(w)}$ the coefficients associated with the covariates to the values estimated in $\mathcal{M}_0$.

Next, considering all the models under $\mathcal{H}_1^{(w)}$ we used the above criterion to select the "best model" $\mathcal{M}_1^{(w^*)}$, that is, the one associated with the potential cluster w maximizing $\mathrm{BF}^{(w)}$. Lastly, to decide whether the estimates should be kept under $\mathcal{H}_0$ or under $\mathcal{H}_1^{(w^*)}$, we followed the rule of thumb developed by Jeffreys (1961): if $\mathrm{BF}^{(w^*)} \geq 30$, then we keep the estimates (using the posterior mean) under $\mathcal{H}_1^{(w^*)}$; otherwise, we keep the estimates (using the posterior mean) under $\mathcal{H}_0$. This threshold of 30 corresponds to "very strong" evidence for $\mathcal{H}_1^{(w^*)}$. Note that if the selected model is $\mathcal{M}_1^{(w^*)}$, the chosen estimate of φ_k is $\varphi_k^* = \hat{\alpha}_{w^*}\mathbb{1}_{s_k\in w^*} + \hat{X}_k$ and if the selected model is $\mathcal{M}_0$, $\varphi_k^* = \hat{X}_k$.

2.2.2 | Scan Procedure

Here, we present a scan procedure on the φ_k^* that identifies spatial clusters of spatial units in which the φ_k^* are significantly higher (corresponding to a higher risk) or significantly lower (corresponding to a lower risk) than elsewhere. Thus, the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1^{(w)}$ are redefined in terms of the distribution of the φ_k^*, as follows:

$$\mathcal{H}_0 : \boldsymbol{\varphi}^* \sim \mathcal{N}(\alpha\mathbb{1}, \sigma^{2(0)}A^{-1}) \text{ and}$$

$$\mathcal{H}_1^{(w)} : \boldsymbol{\varphi}^* \sim \mathcal{N}(\alpha_w\mathbb{1}_w + \alpha_{w^c}\mathbb{1}_{w^c}, \sigma^{2(w)}A^{-1}),\ \alpha_w \neq \alpha_{w^c}$$

where $\boldsymbol{\varphi}^* = (\varphi_1^*, \ldots, \varphi_K^*)^\top$, $\mathbb{1}$ is the column vector composed only of 1, $\mathbb{1}_w$, and $\mathbb{1}_{w^c}$ are the column indicator vectors of w and w^c, respectively, and $A = \rho^* R + (1-\rho^*)\mathrm{I}_K$ in which R is the square matrix composed of the elements

$$R_{k,l} = \begin{cases} \sum_{j=1}^{K} v_{k,j} & \text{if } k = l \\ -v_{k,l} & \text{otherwise} \end{cases}.$$

Note that these assumptions are equivalent to considering the same variance–covariance structure under $\mathcal{H}_0$ and $\mathcal{H}_1^{(w)}$, as with the CAR model considered above (see the proof in Supporting Information A). Since $w \cap w^c = \emptyset$, one assumes under $\mathcal{H}_1^{(w)}$ that the frailty means in w and w^c are different (α_w and α_{w^c}, respectively).

The unknown parameters α, $\sigma^{2(0)}$, α_w, α_{w^c}, and $\sigma^{2(w)}$ are estimated by their maximum likelihood estimators (for proofs, see Supporting Information B):

$$\hat{\alpha} = \frac{\mathbb{1}^\top A \boldsymbol{\varphi}^*}{\mathbb{1}^\top A \mathbb{1}},$$

$$\widehat{\sigma^{2(0)}} = \frac{1}{K}[\boldsymbol{\varphi}^{*\top} A \boldsymbol{\varphi}^* - 2\hat{\alpha}\mathbb{1}^\top A \boldsymbol{\varphi}^* + \hat{\alpha}^2 \mathbb{1}^\top A \mathbb{1}],$$

$$\hat{\alpha}_{w^c} = \left[\mathbb{1}_{w^c}^\top A \mathbb{1}_{w^c} - \frac{\mathbb{1}_w^\top A \mathbb{1}_{w^c} \mathbb{1}_w^\top A \mathbb{1}_{w^c}}{\mathbb{1}_w^\top A \mathbb{1}_w}\right]^{-1} \left[\mathbb{1}_{w^c}^\top A \boldsymbol{\varphi}^* - \frac{\mathbb{1}_w^\top A \boldsymbol{\varphi}^* \mathbb{1}_w^\top A \mathbb{1}_{w^c}}{\mathbb{1}_w^\top A \mathbb{1}_w}\right],$$

$$\hat{\alpha}_w = \frac{\mathbb{1}_w^\top A \boldsymbol{\varphi}^* - \hat{\alpha}_{w^c} \mathbb{1}_w^\top A \mathbb{1}_{w^c}}{\mathbb{1}_w^\top A \mathbb{1}_w} \text{ and}$$

$$\widehat{\sigma^{2(w)}} = \frac{1}{K}[\boldsymbol{\varphi}^* - \hat{\alpha}_w \mathbb{1}_w - \hat{\alpha}_{w^c} \mathbb{1}_{w^c}]^\top A [\boldsymbol{\varphi}^* - \hat{\alpha}_w \mathbb{1}_w - \hat{\alpha}_{w^c} \mathbb{1}_{w^c}].$$

The log-likelihood function under $\mathcal{H}_0$ is then expressed as follows:

$$\ell_{\mathcal{H}_0}(\hat{\alpha}, \widehat{\sigma^{2(0)}}) = -\frac{K}{2}\ln[2\pi] - \frac{1}{2}\ln|A^{-1}| - \frac{K}{2}\ln[\widehat{\sigma^{2(0)}}] - \frac{K}{2},$$

while the log-likelihood function associated with $\mathcal{H}_1^{(w)}$ can be expressed as

$$\ell_{\mathcal{H}_1}(\hat{\alpha}_w, \hat{\alpha}_{w^c}, \widehat{\sigma^{2(w)}}) = -\frac{K}{2}\ln[2\pi] - \frac{1}{2}\ln|A^{-1}| - \frac{K}{2}\ln[\widehat{\sigma^{2(w)}}] - \frac{K}{2}.$$

Thus, the log-likelihood ratio associated with the potential cluster w is

$$\begin{aligned} LLR^{(w)} &= \ell_{\mathcal{H}_1}(\hat{\alpha}_w, \hat{\alpha}_{w^c}, \widehat{\sigma^{2(w)}}) - \ell_{\mathcal{H}_0}(\hat{\alpha}, \widehat{\sigma^{2(0)}}) \\ &= \frac{K}{2}\left[\ln \frac{\widehat{\sigma^{2(0)}}}{\widehat{\sigma^{2(w)}}}\right]. \end{aligned}$$

Lastly, the spatial scan statistic can be defined as

$$\Lambda = \max_{w \in \mathcal{W}} LLR^{(w)}.$$

The most likely cluster (MLC) is then defined as

$$\mathrm{MLC} = \arg\max_{w \in \mathcal{W}} LLR^{(w)}.$$

2.2.3 | Statistical Significance

Once the MLC has been detected, its statistical significance must be evaluated. However, the distribution of Λ does not have a closed form under $\mathcal{H}_0$. In the literature, this distribution is usually approximated with a Monte-Carlo procedure (Dwass 1957). Two main methods can be distinguished, depending on the presence (or not) of a distributional hypothesis for the data. The first method consists of generating datasets under $\mathcal{H}_0$, which thus requires a distributional hypothesis (Kulldorff 1997). The second method (random labeling) consists of randomly permuting the observations among the spatial locations (Kulldorff, Huang, and Konty 2009). In the present case, random labeling is not applicable because the permutation of the observations would change the spatial dependence. Therefore, we used the first method to approximate the distribution of Λ under $\mathcal{H}_0$: since we had assumed a distribution for φ_k, we can generate M datasets under $\mathcal{H}_0$ via $\hat{\alpha}$ and $\widehat{\sigma^{2(0)}}$, which correspond, respectively, to the estimators of the mean and variance of the φ_k under $\mathcal{H}_0$. For each generated dataset m ($1 \leq m \leq M$), one computes the associated spatial scan statistic $\Lambda^{(m)}$, giving an approximation of the distribution of Λ under $\mathcal{H}_0$. Lastly, the p-value associated with the MLC is estimated as

$$\hat{p} = \frac{1 + \sum_{m=1}^{M} \mathbb{1}_{\Lambda^{(m)} \geq \Lambda}}{M+1}.$$

3 | Simulation Studies

Huang, Kulldorff, and Gregorio (2007) and Cook, Gold, and Li (2007) developed spatial scan statistics for time-to-event data indexed in space. However, they supposed that the observations are independent, which is a strong and not very realistic hypothesis because the observations of individuals located in the same spatial unit can be correlated. Thus, in a first simulation study (Section 3.1), we investigated the impact of the presence of intraspatial unit correlation on the type I error of the methods developed by Huang, Kulldorff, and Gregorio (2007) and Cook, Gold, and Li (2007).

In Section 3.2, we conducted two simulation studies. The first one (Section 3.2.2) evaluated our method's ability to correctly estimate both the spatial dependence parameter and the cluster effect. The second one (Section 3.2.3) evaluated the performance of our approach in the context of cluster detection, with spatial dependence and also compared it with the particular i.i.d. ($\rho = 0$) and ICAR ($\rho = 1$) versions.

Section 3.3 evaluates the performance of the approach in the presence of covariates. Lastly, Section 3.4 investigates the performance of our approach in the presence of different levels of censoring of time-to-event data.

3.1 | The Impact of Intraspatial Unit Correlation on the Type I Error in Standard Methods

In this simulation study, we evaluated the type I errors of conventional spatial scan statistics for cluster detection in survival data (namely the exponential model developed by Huang, Kulldorff,

and Gregorio 2007; the method based on a log-rank test developed by Cook, Gold, and Li 2007) in the presence of intraspatial unit correlation.

3.1.1 | Design of the Simulation Study

We considered 1690 individuals distributed in 169 spatial units; the latter corresponded to administrative subdivisions in northern France and were located by their centroid. We defined a spatial cluster w (characterized by α) composed of 135 individuals located in 14 contiguous spatial units (the green area in Figure C.1 in Supporting Information C.1).

We considered the following simulation model for the individual $i_n^{(k)}$ in the spatial unit s_k:

$$\lambda_{i_n^{(k)}}(t|\varphi_k) = \lambda_0(t)\exp[\varphi_k],$$

with $\lambda_0(t) = \frac{1}{2}$ which results in an exponential model. The event times were simulated by inverse transform sampling: for each individual $i_n^{(k)}$, we generated a uniformly distributed random number $u_{i_n^{(k)}}$ on [0,1], which then allowed us to generate a survival time $T_{i_n^{(k)}}$ by $T_{i_n^{(k)}} = \inf_{t>0} 1 - S_{i_n^{(k)}}(t) > u_{i_n^{(k)}}$. Note that this results in $T_{i_n^{(k)}} = -2\ln[1-u_{i_n^{(k)}}]\exp[-\varphi_k]$.

The φ_k were defined as the vector $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_K)^\top$, such that

$$\boldsymbol{\varphi} \sim \mathcal{N}(\alpha \mathbb{1}_w, \sigma^2[\rho R + (1-\rho)I_K]^{-1}),$$

where $\mathbb{1}_w$ is the column indicator vector of w.

Here, we focused our analysis on the type I errors ($\alpha = 0$) in the exponential model (Huang, Kulldorff, and Gregorio 2007) and the log-rank test method (Cook, Gold, and Li 2007) in the presence of a nonspatially correlated ($\rho = 0$) shared frailty for the frailty variance σ^2, which ranged from 0.001 to 0.101 in increments of 0.010.

For each value of σ^2, 100 datasets were simulated. The statistical significance of the MLC was evaluated in the same way as in the original publications, that is, by using 999 permutations of the data. The type I error was set to 0.05.

3.1.2 | Results

Figure 1 shows the type I error as a function of σ^2. One can note that the type I error increases with σ^2, showing that the nominal level is not maintained. This can be explained by the fact that under hypothesis $\mathcal{H}_0$ (the absence of a cluster), the increase in σ^2 leads directly to an increase in the variance of X_k. Since the two standard models do not incorporate a shared frailty, the identification of false-positive spatial clusters is essentially due to the intraspatial unit correlation (i.e., the variance of X_k).

3.2 | Evaluation of the Method's Performance

Here, two simulation studies were conducted. The first (Section 3.2.2) assessed the ability of our method to accurately estimate both the spatial dependence parameter and the cluster effect. The second (Section 3.2.3) evaluated the performance of our method in the context of cluster detection and compared it with two particular versions of the model in the presence of spatial dependence: one assuming no spatial dependence (the i.i.d. frailty model) and one assuming complete spatial dependence (the ICAR frailty model).

3.2.1 | Design of the Simulation Studies

The designs of these simulation studies are very similar to those presented in Section 3.1. However, given that we wanted to investigate the impact of spatial dependence on cluster detection, we set σ^2 to 1 and considered several values for the parameters controlling the spatial dependence $\rho \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and the cluster effect $\alpha \in \{0, 0.5, 1, 1.5, 2\}$. Note that $\alpha = 0$ was considered in order to evaluate the maintenance of the type I error.

For each value of the spatial dependence parameter ρ and each value of α, 100 datasets were simulated. The statistical significance of the MLC was evaluated through 999 generations of the data under $\mathcal{H}_0$ (see Section 2.2.3 for more details), and the type I error was set to 0.05.

The performances were measured through four criteria: the power, the true positive rate, the false positive rate, and the positive predictive value. The power was estimated as the proportion of simulations leading to the rejection of $\mathcal{H}_0$, depending on the type I error. Using the simulated datasets leading to the rejection of $\mathcal{H}_0$, the true positive rate was defined as the mean proportion of individuals correctly detected among the individuals in w, the false positive rate was defined as the mean proportion of individuals in w^c that were included in the detected cluster, and the positive predictive value corresponded to the mean proportion of individuals in w within the detected cluster.

Since the estimations of the φ_k and ρ were performed in a Bayesian framework, we considered the following Leroux CAR prior for X_k: $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2[\rho R + (1-\rho)I_K]^{-1})$, with a $\boldsymbol{\beta}(1,1)$ prior for the spatial dependence parameter ρ and a $\Gamma(10^{-3}, 10^{-3})$ prior for the precision $1/\sigma^2$. For α_w, we chose a noninformative prior $\mathcal{N}(0, 10^3)$.

Lastly, for the baseline hazard λ_0, the observation times were divided into n_T time intervals. Here, n_T was set to the number of unique times divided by 20. Next, λ_0 was assumed to be constant in each time interval, and for each interval I we assumed that $\lambda_0 = \exp(c_I)$. We chose a Gaussian prior on the c_I increments with a precision τ such that $\tau \sim \Gamma(10^{-3}, 10^{-3})$: $\Delta c_I = c_I - c_{I-1} \sim \mathcal{N}(0, \tau^{-1})$.

3.2.2 | Evaluation of the Estimates of ρ and α_w

Section 2.2.1 presents the estimation of the φ_k. Briefly, it consisted in choosing either the estimates under the best hypothesis $\mathcal{H}_1^{(w)}$: $\mathcal{H}_1^{(w^*)}$ (in this case, the estimates are $\varphi_k^* = \hat{\alpha}_{w^*}\mathbb{1}_{s_k \in w^*} + \hat{X}_k$) or the estimates under $\mathcal{H}_0$ ($\varphi_k^* = \hat{X}_k$). The present section focuses on the bias of the estimates obtained for the spatial dependence parameter (ρ^*) and for the cluster effect ($\hat{\alpha}_{w^*}$). Note that for

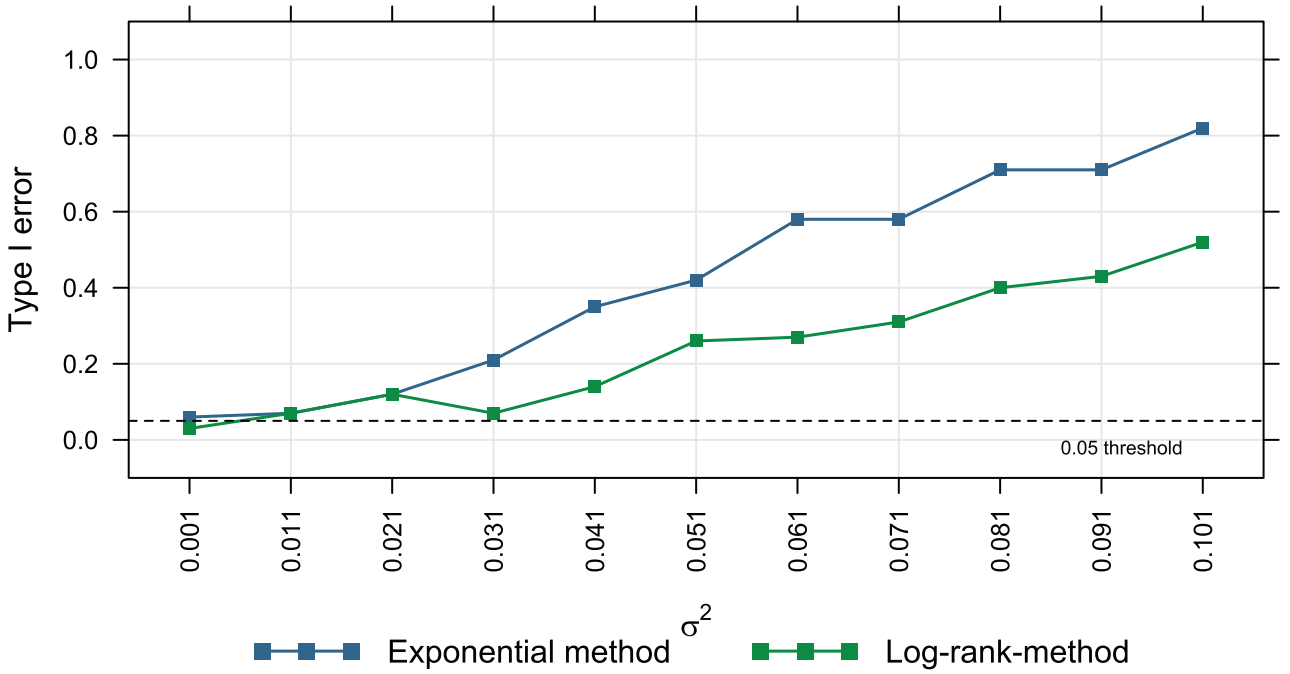

FIGURE 1 | Type I error in the exponential method (Huang, Kulldorff, and Gregorio 2007) and the log-rank test method (Cook, Gold, and Li 2007) as a function of the degree of intraspatial unit correlation (characterized by the simulated values of the shared frailty variance σ^2).

the cluster effect, we only considered simulations that did not retain $\mathcal{H}_0$ (otherwise, an estimate $\hat{\alpha}_{w^*}$ was unavailable). Thus, the estimates obtained were compared with the true values of the spatial dependence parameter and the cluster effect.

Figure 2 shows the selected ρ^* as a function of the parameters ρ and α, and the estimations $\hat{\alpha}_{w^*}$ with the INLA method when one selects $\mathcal{H}_1$ according to the Bayes factor criterion.

Our approach estimates the cluster effect well when the simulated values of α are 1, 1.5 and 2. Although the cluster effect might appear to be poorly estimated for α values of 0 and 0.5, this was because our approach rarely selected $\mathcal{H}_1$ for these values and so few estimates were made.

The parameter ρ was well estimated generally but was slightly overestimated for $\rho = 0$ and slightly underestimated for $\rho = 0.8$.

3.2.3 | The Impact of Spatial Dependence on Cluster Detection

We next evaluated the performance of our new method in the context of cluster detection. Two particular versions of the method were also considered, in order to investigate the impact of taking account of potential intraspatial unit correlation but not spatial dependence between spatial units (the i.i.d. model with $\rho = 0$) or taking account of spatial dependence between spatial units without adjusting its intensity (i.e., by considering it to be complete: the ICAR model, $\rho = 1$). Note that the ICAR model with $\rho = 1$ leads to a noninvertible matrix A, and so it was not possible to generate data under $\mathcal{H}_0$ to estimate the p-value associated with the most likely cluster (see Section 2.2.3 for more details). To overcome this problem, the value of the spatial dependence was set to 0.999 (instead of 1) in the scan procedure (Section 2.2.2) for the ICAR model.

Figure 3 shows the type I error, the power curves, true positive rates, false positive rates, and positive predictive values obtained with our method and with its two special cases (i.i.d. and ICAR).

For the Leroux CAR model, the performances were relatively stable as a function of ρ, although the type I error was slightly above the 5% threshold. This was not the case for $\rho = 0$ because then ρ^* slightly overestimates ρ (Figure 2), which makes the method quite conservative). It should be noted that the true positive rates, the false positive rates, and the positive predictive values appear to be less stable when $\alpha = 0.5$. This was because these indicators are only computed for simulations that lead to the rejection of $\mathcal{H}_0$, which were not numerous when $\alpha = 0.5$.

The i.i.d. model failed to maintain a reasonable type I error as ρ increased. Moreover, the power as a function of ρ was less stable than that of the CAR model.

The ICAR model tended to absorb the cluster effect into the spatial dependence parameter ρ. This was particularly the case when the true value of ρ was low. Thus, the type I errors remain reasonable but the power tended to decrease as ρ decreased.

The false positive rates were very low for the three approaches. However, the true positive rates and the positive predictive values were lower for the i.i.d. and the ICAR models than for the CAR model.

We also investigated the performance of our method with other thresholds for the Bayes factor (i.e., 3, 10, and 100, which correspond, respectively, to "substantial," "strong," and

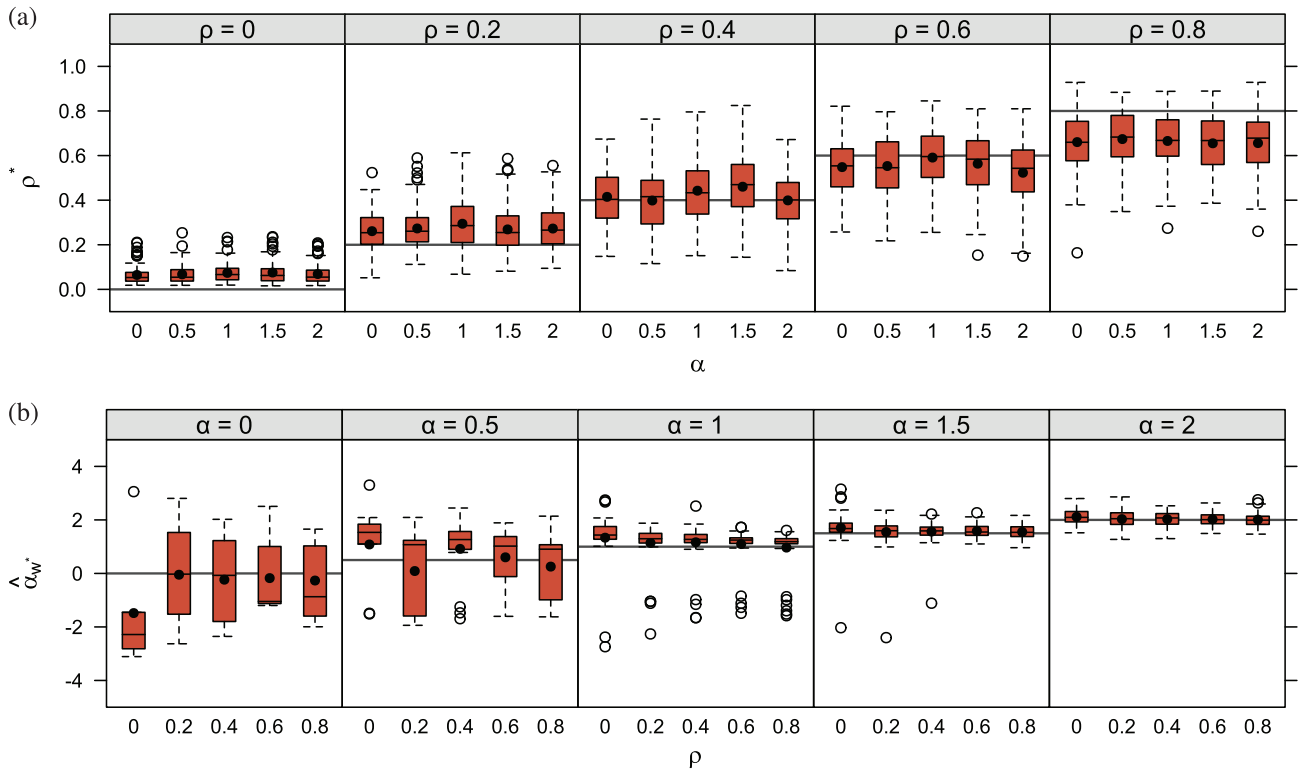

FIGURE 2 | Simulation study: the selected values of ρ^* according to the parameters ρ and α (panel a) and $\hat{\alpha}_{w^*}$ obtained with the INLA method when we selected $\mathcal{H}_1$ according to the Bayes factor criterion (panel b). The horizontal lines in panels a and b correspond, respectively, to the true values of the parameters ρ and α, and the black points represent the mean estimates obtained.

"decisive" levels of evidence for $\mathcal{H}_1^{(w^*)}$; Jeffreys 1961). The results are presented in Figure C.3 in Supporting Information C.2.

3.3 | Adjustment on Covariates

The purpose of this simulation study is to ensure that our approach correctly adjusts cluster detection on covariates. To this end, we considered a design similar to that of Section 3.2 but added a covariate $Z_{i_n^{(k)}}$ such that

$$\lambda_{i_n^{(k)}}(t|\varphi_k) = \lambda_0(t)\exp\left[\varphi_k + \beta Z_{i_n^{(k)}}\right].$$

Two different scenarios were then considered. In the first case, we considered a distribution on the $Z_{i_n^{(k)}}$ that does not depend on the simulated cluster. The second one consists of simulating a "true cluster" as previously (the green area in Figure C.1 in Supporting Information C.1), whose intensity does not depend on the covariate, and similarly to Ahmed and Genin (2020), a "fake cluster" whose intensity depends on the effect of the covariate. In this case, the covariate is a confounding factor for the association between the survival of individuals and the "true cluster."

To be more precise, for Scenario 1, $Z_{i_n^{(k)}} \underset{i.i.d.}{\sim} \mathcal{N}(2,1)$, whereas for Scenario 2, $Z_{i_n^{(k)}} \underset{i.i.d.}{\sim} \mathcal{N}(2,1)$ if $i_n^{(k)} \in f$ and $Z_{i_n^{(k)}} \underset{i.i.d.}{\sim} \mathcal{N}(1,1)$ otherwise, where f, namely the "fake cluster" is the red area in Figure C.1 in Supporting Information C.1 composed of 149 individuals in 14 spatial units. f is thus only characterized by the effect of the covariate. Two values of β were considered: $\beta = 0.5, 1.5$.

For each scenario, each value of β, each value of the spatial dependence parameter ρ, and each value of α, 100 datasets were simulated. The statistical significance of the MLC was evaluated through 999 generations of the data under $\mathcal{H}_0$ (see Section 2.2.3 for more details) and the type I error was set to 0.05.

The performances were measured through the same four criteria as in Section 3.2: the power, the true positive rate, the false positive rate, and the positive predictive value. We also computed the true positive rates and the positive predictive values for f for Scenario 2, defined as the mean proportion of individuals detected among the individuals in f and the mean proportion of individuals in f within the detected cluster, respectively, to ensure that f is not detected as a cluster.

The prior distributions were the same as in Section 3.2 and for β, we chose the noninformative prior $\mathcal{N}(0, 10^3)$.

3.3.1 | Evaluation of the Estimate of β

As explained in Section 2.2.1, β is estimated under the null hypothesis $\mathcal{H}_0$. Then its estimation is used under all hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1^{(w^*)}$. This section presents the bias of the estimates obtained for β: $\hat{\beta}$. Figure 4 shows the estimations $\hat{\beta}$ obtained with the INLA method. It indicates that whatever the scenario considered, the value of the spatial dependence parameter ρ or the cluster effect α, the parameter β is well estimated.

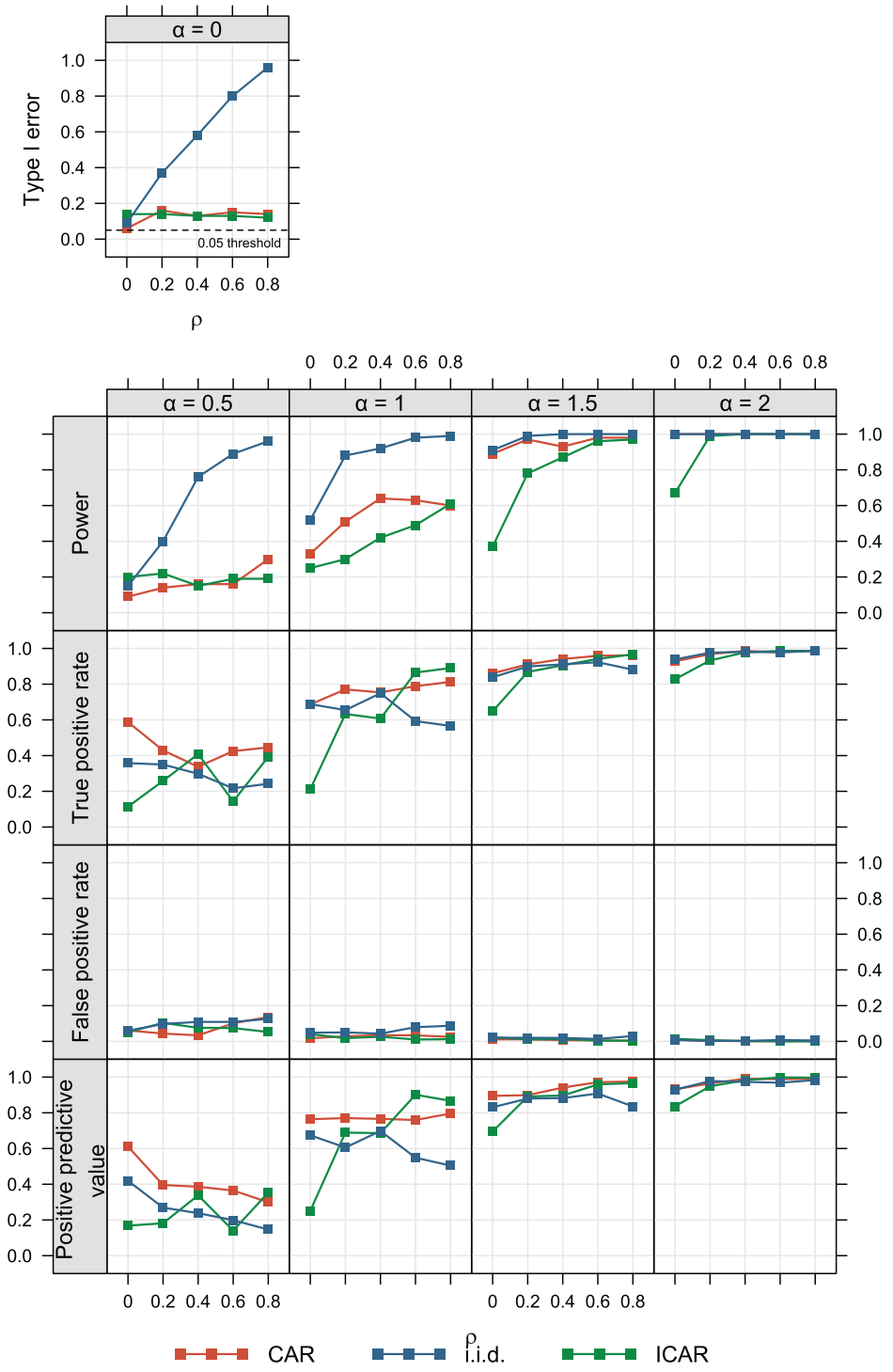

FIGURE 3 | Simulation study: Comparison of the type I error, power curves, true positive rates, false positive rates, and positive predictive values for the CAR, ICAR, and i.i.d. models. α is the parameter that controls the cluster intensity and ρ controls the spatial dependence.

3.3.2 | The Impact of Covariates on Cluster Detection

We evaluated the impact of the presence of a covariate on the performance of our approach in the context of cluster detection. Figure 5 shows the power curves, true positive rates, false positive rates, and positive predictive values obtained with our method in the presence of a covariate and compares them with the performances obtained in the absence of a covariate (the red curves). The type I errors are also presented in Figure C.5 in Supporting Information C.3. No impact of the covariate on the performances is visible in these figures.

In particular, in the case of a confounding factor (Scenario 2), the power and the type I error remain the same as in its absence. This indicates that the presence of a "fake cluster" does not lead to the detection of a statistically significant cluster corresponding to it.

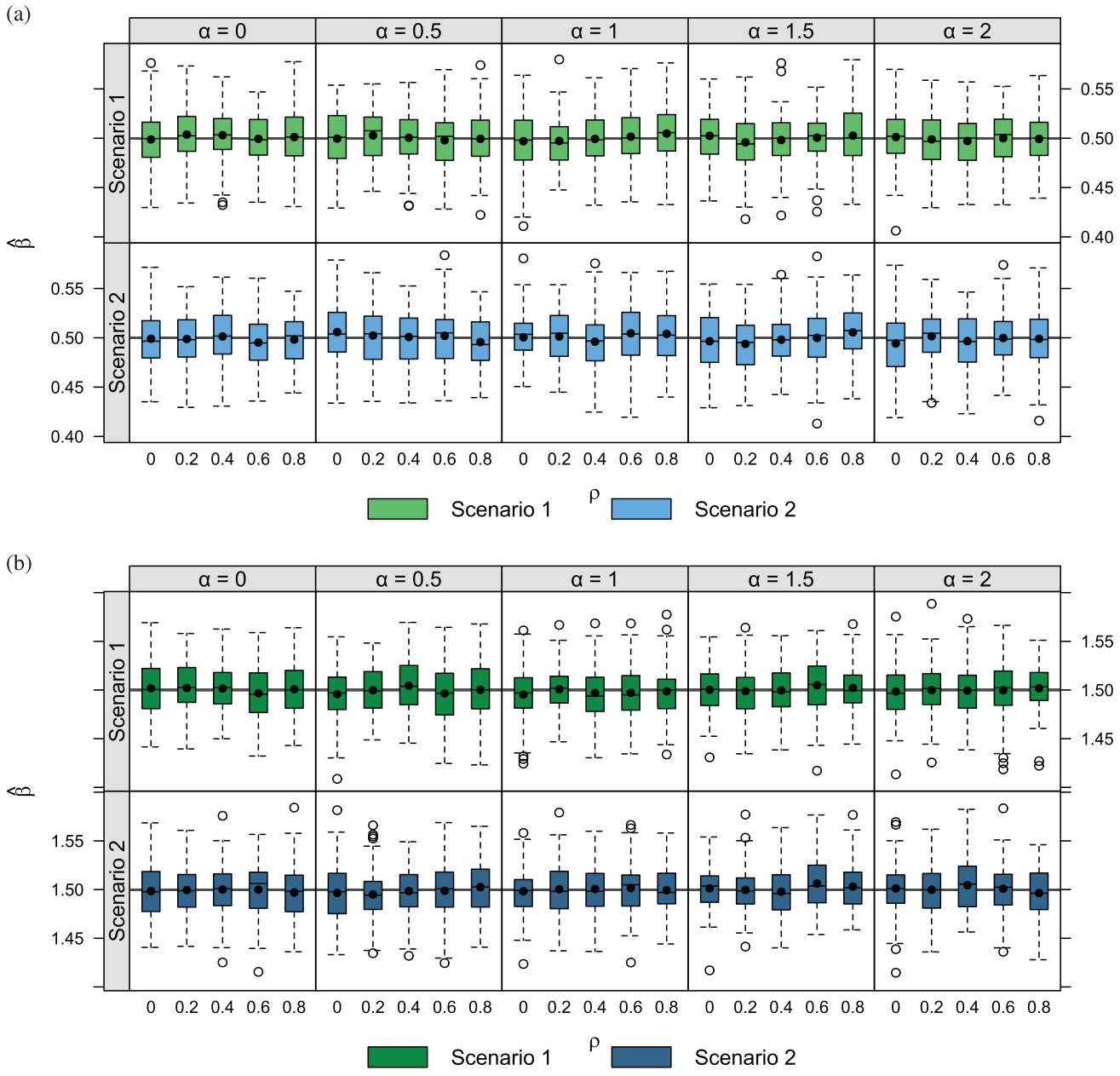

FIGURE 4 | Simulation study: the estimations of β ($\hat{\beta}$) according to the parameters ρ and α obtained with the INLA method. Panel (a) corresponds to $\beta = 0.5$, and Panel (b) corresponds to $\beta = 1.5$. The horizontal lines correspond to the true value of β, and the black points represent the mean estimates obtained.

Figure 5 also shows that the true positive rates and the positive predictive values for the "fake cluster," as well as the false positive rates for the "true cluster," are extremely low, while the true positive rates and the positive predictive values for the "true cluster" are elevated. These results show that the cluster detected corresponds to the "true cluster" and that the adjustment has been correctly performed.

3.4 | The Influence of Censoring

Here, we describe the simulation study that was designed to evaluate the performance of our approach in the presence of different levels of data censoring.

3.4.1 | Design of the Simulation Study

Due to computational time constraints, the simulation's design differed slightly from those of the previous studies: we considered 940 individuals distributed in the 94 French *départements* (counties) located by their centroid. The simulated cluster contains 73 individuals in the eight *départements* of the Île-de-France region (the green area in Figure C.2 in Supporting Information C.1).

The data were generated in the same way as in Section 3.2, except that different proportions of the observations were censored (10%, 20%, 30%, and 40%). Administrative censoring was considered according to Montez-Rath et al. (2017) method. Briefly, the end

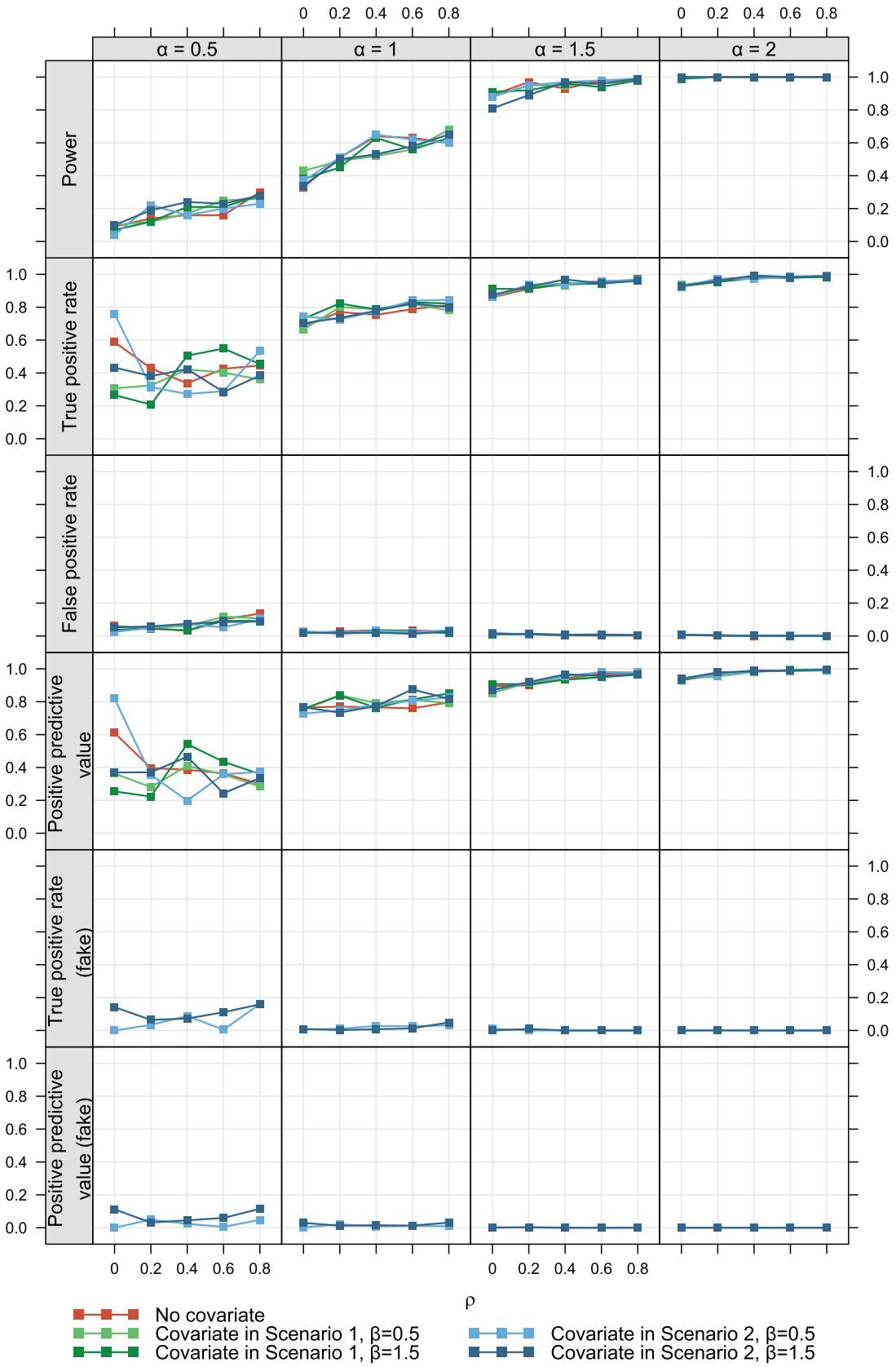

FIGURE 5 | Simulation study: Comparison of the power curves, true positive rates, false positive rates, and positive predictive values in the presence and the absence of a covariate. α is the parameter that controls the cluster intensity, ρ controls the spatial dependence, and β controls the covariate effect on the survival. The true positive rates and the positive predictive values for a *fake cluster* only characterized by a confounder (Scenario 2) are also presented.

of the study was determined so that the intended proportion of censoring was achieved.

For each value of the spatial dependence parameter ρ, each value of α, and each censoring percentage, 100 datasets were simulated. The statistical significance of the MLC was evaluated through 999 generations of the data under $\mathcal{H}_0$ (see Section 2.2.3 for more details), and the type I error was set to 0.05.

The performances were measured through the same four criteria as in Section 3.2: the power, the true positive rate, the false positive rate, and the positive predictive value.

Note that in this simulation study, the prior distributions were the same as in Section 3.2.

3.4.2 | Results

The results of the simulation study are shown in Figure 6. First, we found that the power of our method increases as the proportion of censoring increases. This was also the case for the type I error (see Figure C.6 in Supporting Information C.4). Although the type I error remained stable and close to the nominal value (whatever the values of ρ) when 10% of the observations were censored, it tended to move away from the nominal value as the censoring percentage increased.

The false positive rates remained very low regardless of the censoring rate. However, the true positive rates and positive predictive values decreased as the censoring rate increased. Lastly, the impact of censoring on the performance indicators decreased when the cluster's intensity α increased.

4 | Application to Epidemiological Data

4.1 | End-Stage Renal Disease Mortality and Related Confounding Factors

We considered data provided by the French renal epidemiology and information network (REIN) registry on end-stage renal disease (ESRD) in northern France between 2004 and 2020. The methodology of the REIN registry has been described elsewhere (Couchoud et al. 2005). Here, we focused on the analysis of mortality, measured by the survival time after the initiation of dialysis in ESRD patients aged 70 and over. This patient population is characterized by (i) a high mortality rate, thus leading to a high number of observed deaths, and (ii) a low frequency of kidney transplantation, thus minimizing the effect of this known competing risk of death among ESRD patients (Ayav et al. 2016; Hallan et al. 2012). The data covered 6071 individuals but the exact time to survival after the initiation of dialysis was not known in 17% of cases. These censored observations are either patients still alive at the end of the study (15.7%), patients lost to follow-up (0.7%), or patients having received a kidney transplant (in which case, the censoring time corresponds to the date of transplantation; 0.6%). The geographical region studied (the *Nord-Pas-de-Calais* region of northern France) is divided into 80 *cantons* (a French administrative subdivision), and each individual's stated place of residence was linked to the corresponding *canton*.

We also considered 18 variables measured at the individual level, and that are known to be confounders of survival in patients with ESRD (Couchoud et al. 2015; Fu et al. 2021). Thus, spatial cluster detection was adjusted by introducing the following confounders into each model as covariates: age (in years), sex, body mass index (in kg/m^2), the type of nephropathy (polycystic, primitive glomerulonephritis, hypertension, or vascular, diabetic, pyelonephritis, other), the number of cardiovascular comorbidities (none, one, two, or more), mobility (independent walking, need for help from a third party, total disability), the blood hemoglobin level (in g/dL), the serum albumin level (in g/dL), the dialysis method (hemodialysis or peritoneal dialysis), the glomerular filtration rate (below 7, between 7 and 10 or over 10 mL/min/1.73 m^2), the period of treatment initiation (2004–2009, 2010–2015, or 2016–2020), whether or not the treatment was initiated urgently, and the presence or absence of diabetes, chronic respiratory disease, respiratory assistance, cirrhosis, severe behavioral disorder, or active malignant cancer. Details of these confounding factors are available in Supporting Information D.

4.2 | Spatial Cluster Detection

In order to detect spatial clusters of atypical (shorter or longer) survival times among patients with ESRD, five models were considered: the exponential model (Model 1) developed by Huang, Kulldorff, and Gregorio (2007), the log-rank method developed by Cook, Gold, and Li (2007) (Model 2), and versions of the Cox-model-based method presented here for considering three types of shared frailty: i.i.d. ($\rho = 0$) (Model 3), CAR ($\rho \in]0, 1[$) (Model 4), and ICAR ($\rho = 1$) (Model 5).

Each model was used to detect spatial clusters of atypical survival times among the patients with ESRD when compared with patients in the rest of the region studied. To adjust survival times for the confounders in Model 1, we used an exponential regression method as proposed by Huang, Kulldorff, and Gregorio (2007). Regarding Model 2, we adopted the approach developed by Jung (2009) and Ahmed and Genin (2020), which consists of estimating the coefficients associated with the confounders in the model under $\mathcal{H}_0$ and then setting their value to this estimate in the scan statistic developed by Cook, Gold, and Li (2007). Regarding Models 3–5 (the shared frailty models), we also adopted this approach by setting (under each alternative hypothesis $\mathcal{H}_1^{(w)}$) the coefficients associated with the confounding factors to the values estimated in the model under $\mathcal{H}_0$ in the φ_k estimation step (Section 2.2.1).

In order to provide an indicator of the cluster-associated risk that is independent of the model considered, we estimated the hazard ratio (HR) associated with each cluster in a conventional Cox model adjusted for the confounding factors.

The MLC was considered, as were secondary clusters that had a high Λ value and did not cover the MLC (Kulldorff 1997). The statistical significance of the detected spatial clusters was

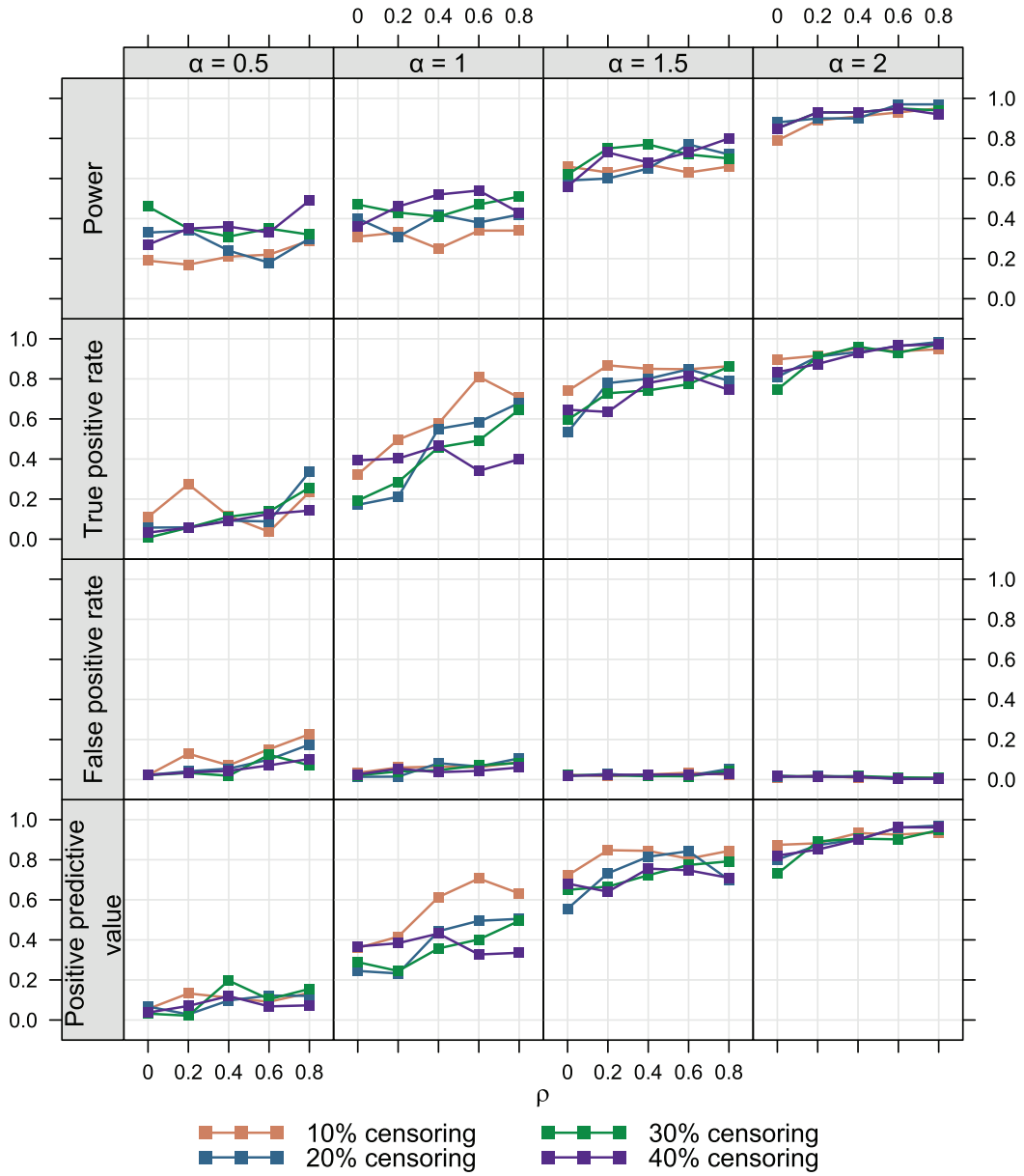

FIGURE 6 | Simulation study: Comparison of the power curves, true positive rates, false positive rates, and positive predictive values according to the percentage of censored observations. α is the parameter that controls the cluster intensity, and ρ controls the spatial dependence.

evaluated by performing 999 Monte Carlo simulations, with a type I error of 0.05.

4.3 | Results

The spatial clusters detected by each of the five models (exponential, log-rank, i.i.d., CAR, and ICAR frailty) are presented in Figure 7. Detailed information on the spatial clusters is presented in Table 1.

Both the exponential model (Model 1, panel a in Figure 7) and the method based on the log-rank test (Model 2, panel b in Figure 7) identified the same two statistically significant spatial clusters. The MLC (located in the northeast of the region, shown in green) had similar levels of statistical significance in the two models ($\hat{p} = 0.004$ and $\hat{p} = 0.005$, respectively) and had longer survival times than in the rest of the geographical area studied (HR = 0.84 for both models). The first secondary cluster (located in the western part of the region, shown in red) also had similar levels of statistical significance in the two models ($\hat{p} = 0.025$ and $\hat{p} = 0.043$, respectively) and was characterized by shorter survival times (HR = 1.13 for both methods).

The i.i.d. frailty model (Model 3, panel c) identified the same statistically significant MLC as the exponential model and the method based on the log-rank test ($\hat{p} = 0.006$). The CAR model (Model 4, panel d) and ICAR model (Model 5, panel e) both detected the same MLC, which contained three more spatial units than the MLC detected by the other models. This MLC was characterized by longer survival times (HR = 0.86). The MLC was statistically significant for the CAR model but not for the ICAR model ($\hat{p} = 0.011$ and $\hat{p} = 0.178$, respectively). The first secondary cluster detected by the three

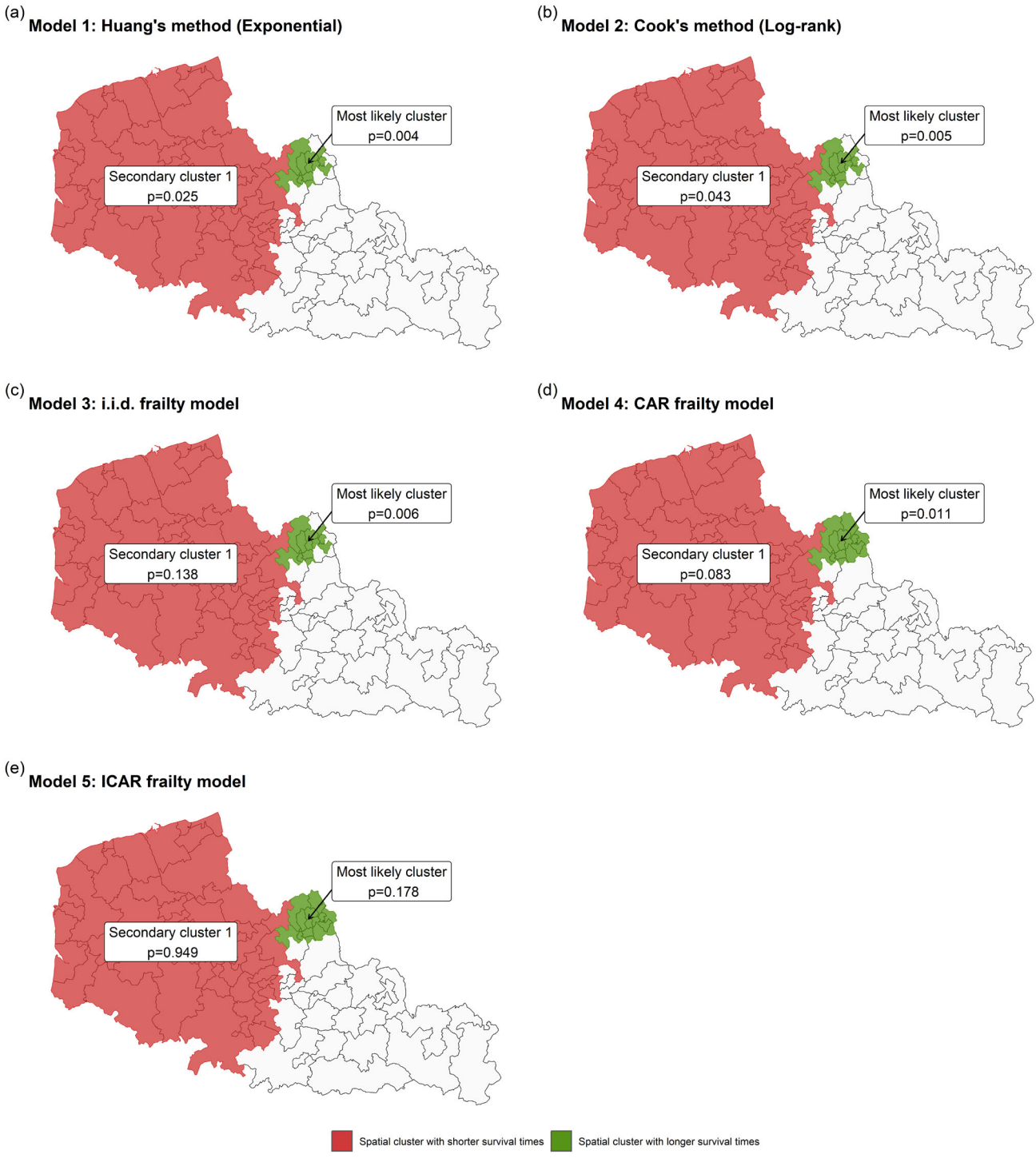

FIGURE 7 | Spatial clusters detected by the method developed by Huang, Kulldorff, and Gregorio (2007) (Model 1 (exponential), panel a), the method developed by Cook, Gold, and Li (2007) (Model 2 (log-rank), panel b), and those highlighted by the shared frailty models (Model 3 (i.i.d.), panel c; Model 4 (CAR), panel d; Model 5 (ICAR), panel e) after adjustment for confounding factors. Spatial clusters in green indicate longer survival times for patients with ESRD, compared with the rest of the region studied. Conversely, spatial clusters in red indicate shorter survival times for patients with ESRD.

frailty models is the same as that detected by the exponential model and the method based on the log-rank test. However, it was not statistically significant for any of the shared frailty models ($\hat{p} = 0.138$ for the i.i.d. frailty model, $\hat{p} = 0.083$ for the CAR frailty model, and $\hat{p} = 0.949$ for the ICAR frailty model).

The small differences between the conventional spatial scan statistics methods (Cook, Gold, and Li 2007; Huang, Kulldorff, and Gregorio 2007) and the three shared frailty models developed here can be explained by the low variance of the shared frailties (see Figure D.1 in Supporting Information D for the posterior distribution of σ^2 with each model).

TABLE 1 | Description of the statistically significant spatial clusters detected by the method developed by Huang, Kulldorff, and Gregorio (2007) (Model 1 (exponential)), the method of Cook, Gold, and Li (2007) (Model 2 (log-rank)) and those detected by the shared frailty models (Model 3 (i.i.d.), Model 4 (CAR) and Model 5 (ICAR)), after adjustment for confounding factors.

Model	Cluster	*p* value	Number of spatial units	Number of individuals	Number of events	Hazard ratio[a]
Model 1	MLC	0.004	10	1091	890	0.84
(Exponential)	Secondary cluster 1	0.025	43	2632	2163	1.13
Model 2	MLC	0.005	10	1091	890	0.84
(Log-rank)	Secondary cluster 1	0.043	43	2632	2163	1.13
Model 3	MLC	0.006	10	1091	890	0.84
(i.i.d. frailty)	Secondary cluster 1	0.138	43	2632	2163	1.13
Model 4	MLC	0.011	13	1346	1094	0.86
(CAR frailty)	Secondary cluster 1	0.083	43	2632	2163	1.13
Model 5	MLC	0.178	13	1346	1094	0.86
ICAR frailty	Secondary cluster 1	0.949	43	2632	2163	1.13

[a]The hazard ratio was computed using a Cox model with adjustment for confounders.

5 | Discussion

Here, we developed a new spatial scan statistic for survival data indexed in space. It allows one to (i) take account of both potential intraspatial unit correlation and spatial dependence between spatial units and (ii) adjust the cluster detection for confounding factors. This method is based on a Cox model that includes spatially structured shared frailty distributed according to a Leroux CAR model.

In a simulation study, we showed that in the presence of intraspatial unit correlation, the existing methods (Cook, Gold, and Li 2007; Huang, Kulldorff, and Gregorio 2007) are confronted by a huge increase in the type I error. Thereafter, the performance of the CAR model was evaluated in the context of cluster detection and compared with two particular versions of it: the i.i.d. frailty model and the ICAR frailty model. The CAR model presented the best performances in the presence of spatial dependence, which thus demonstrated good-quality adjustment. We have also investigated the goodness of fit of our method on covariates through the bias of the estimator obtained for the effect of the covariate, and the performances for the detection of the simulated cluster. This simulation study showed that the proposed approach allows to fit correctly on covariates (including confounders) according to these criteria. In the last simulation study, we showed that the performance of the CAR model is adequate as long as the percentage of censored observations does not exceed 20%.

These approaches were then applied to epidemiological data, that is, the detection of clusters of abnormally low or high survival times in elderly patients with ESRD in northern France during the period 2004–2020. The conventional approaches (Cook, Gold, and Li 2007; Huang, Kulldorff, and Gregorio 2007) detected two statistically significant clusters: one in the northeast of the region (corresponding to longer survival times, that is, a lower risk than elsewhere) and the other containing the whole western part of the region (corresponding to lower survival times, i.e. a higher risk). The i.i.d. shared frailty model only detected the cluster in the northeast of the region as being statistically significant. Assuming a complete spatial dependence, the ICAR model also identified an MLC in the northeast of the region but this was not statistically significant. When we considered the CAR frailty model that allowed flexibility of the spatial dependence, a statistically significant cluster was detected in the northeast of the region. The cluster's *p*-value was slightly higher than that provided by the i.i.d. shared frailty model; this can be explained by the fact that the CAR model takes account of spatial dependence. These results are consistent with those of the simulation study.

In both the simulation study and the application to epidemiological data, circular potential clusters were considered. However, other cluster shapes (e.g., elliptical clusters; Kulldorff et al. 2006) could be considered, since the shape of the scanning window has an impact on the power of cluster detection. It should be noted that other scanning window shapes can be easily implemented in our method because this only changes the set of potential clusters $\mathcal{W}$.

In the population of elderly patients with ESRD, only a low percentage had received a kidney transplant. However, this percentage is higher in the general population (Couchoud et al. 2015). It is well known that kidney transplantation is a competing risk for death in patients with ESRD, and failure to take account of this risk in the analysis might bias the estimate of survival (Hallan et al. 2012). In this context, the method developed here should be modified to account for competing risks by considering (for instance) the model developed by Fine and Gray (1999).

Here, the spatial dependence parameter ρ was assumed to be constant over the whole study area. This assumption may be too simplistic because this coefficient can vary spatially (Crawford 2009). However, the integration of a varying spatial dependence coefficient would be challenging because it is necessary to clearly distinguish between the effect of spatial dependence and the effects of spatial clusters in the data. Adapting the method developed here to this context could be the subject of future research.

In our model, we included covariates as fixed effects. However, it is possible to consider them as random effects that may have a spatial dependence. One way of taking these spatially structured effects into account involves the use of CAR models for the prior distributions of the coefficients associated with the covariates.

Moreover, it should be noted that the methodology proposed in this article considers a Bayesian approach in order to estimate the spatial frailties and then a frequentist approach for the realization of the scanning process. This method is rather singular since it does not allow to use the richness of the posterior distributions in the scanning process. However, the development of a fully Bayesian spatial scan statistics method (Cançado, Fernandes, and da Silva 2017; Neill, Moore, and Cooper 2005; Neill and Cooper 2010) in the context of survival data, allowing to take into account the uncertainties, both on the frailty estimates and on spatial dependence estimate, requires complex mathematical developments that could be the subject of future work.

Lastly, our spatial scan statistics can be extended to deal with recurrent events. For example, one might be interested in the time until an asthma attack in patients treated for asthma, and a patient might experience several asthma attacks during the study period. One possible approach is to consider shared frailty at the individual level, making it possible to take account of unobserved, subject-specific factors (Kleinbaum and Klein 2012). However, the time to an asthma attack might also exhibit an intraspatial unit correlation, due (for instance) to environmental factors. In this context, one approach would be to consider a nested frailty model (Rondeau 2010), that is, a model with both shared frailties at the level of spatial units with a potential spatial dependence, and shared frailties at the level of individuals, in order to take account of unobserved factors that are specific to spatial units (e.g., air quality) and those that are specific to individuals (e.g., tobacco consumption).

Acknowledgments

The authors would like to thank Sebastien Gomis, Aghiles Hamroun, François Glowacki, the REIN network, and the Nephronor registry for providing the renal disease data.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Research data are not shared due to privacy or ethical restrictions.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues and computational complexity.

References

Abolhassani, A., and M. O. Prates. 2021. "An Up-to-Date Review of Scan Statistics." *Statistics Surveys* 15: 111–153.

Ahmed, M.-S., L. Cucala, and M. Genin. 2021. "Spatial Autoregressive Models for Scan Statistic." *Journal of Spatial Econometrics* 2, no. 1: 1–20.

Ahmed, M.-S., and M. Genin. 2020. "A Functional-Model-Adjusted Spatial Scan Statistic." *Statistics in Medicine* 39, no. 8: 1025–1040.

Arlinghaus, S. 1995. *Practical Handbook of Spatial Statistics*. Boca RAton, FL: CRC Press.

Aswi, A., S. Cramb, E. Duncan, W. Hu, G. White, and K. Mengersen. 2020. "Bayesian Spatial Survival Models for Hospitalisation of Dengue: A Case Study of Wahidin Hospital in Makassar, Indonesia." *International Journal of Environmental Research and Public Health* 17, no. 3: 878.

Austin, P. C. 2017. "A Tutorial on Multilevel Survival Analysis: Methods, Models and Applications." *International Statistical Review* 85, no. 2: 185–203.

Ayav, C., J.-B. Beuscart, S. Briançon, A. Duhamel, L. Frimat, and M. Kessler. 2016. "Competing Risk of Death and End-Stage Renal Disease in Incident Chronic Kidney Disease (Stages 3 to 5): The EPIRAN Community-Based Study." *BMC Nephrology* 17, no. 1: 1–13.

Banerjee, S., M. M. Wall, and B. P. Carlin. 2003. "Frailty Modeling for Spatially Correlated Survival Data, With Application to Infant Mortality in Minnesota." *Biostatistics* 4, no. 1: 123–142.

Besag, J., J. York, and A. Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics." *Annals of the Institute of Statistical Mathematics* 43, no. 1: 1–20.

Bhatt, V., and N. Tiwari. 2014. "A Spatial Scan Statistic for Survival Data Based on Weibull Distribution." *Statistics in Medicine* 33, no. 11: 1867–1876.

Cançado, A. L., C. Q. da Silva, and M. F. da Silva. 2014. "A Spatial Scan Statistic for Zero-Inflated Poisson Process." *Environmental and Ecological Statistics* 21, no. 4: 627–650.

Cançado, A. L., L. B. Fernandes, and C. Q. da Silva. 2017. "A Bayesian Spatial Scan Statistic for Zero-Inflated Count Data." *Spatial Statistics* 20: 57–75.

Chen, J., and J. Glaz. 2009. "Approximations for Two-Dimensional Variable Window Scan Statistics." In *Scan Statistics*, edited by J. Glaz, V. Pozdnyakov, and S. Wallenstein, 109–128. Boston: Birkhäuser.

Clayton, D. G. 1978. "A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence." *Biometrika* 65, no. 1: 141–151.

Cook, A. J., D. R. Gold, and Y. Li. 2007. "Spatial Cluster Detection for Censored Outcome Data." *Biometrics* 63, no. 2: 540–549.

Couchoud, C., B. Stengel, P. Landais, et al. 2005. "The Renal Epidemiology and Information Network (REIN): A New Registry for End-Stage Renal Disease in France." *Nephrology Dialysis Transplantation* 21, no. 2: 411–418.

Couchoud, C. G., J.-B. R. Beuscart, J.-C. Aldigier, P. J. Brunet, and O. P. Moranne. 2015. "Development of a Risk Stratification Algorithm to Improve Patient-Centered Care and Decision Making for Incident Elderly Patients with End-Stage Renal Disease." *Kidney International* 88, no. 5: 1178–1186.

Crawford, T. 2009. "Scale Analytical." In *International Encyclopedia of Human Geography*, edited by R. Kitchin and N. Thrift, 29–36. Oxford, UK: Elsevier.

Cressie, N. 1977. "On Some Properties of the Scan Statistic on the Circle and The Line." *Journal of Applied Probability* 14: 272–283.

Cucala, L., M. Genin, C. Lanier, and F. Occelli. 2017. "A Multivariate Gaussian Scan Statistic for Spatial Data." *Spatial Statistics* 21: 66–74.

De La Fuente Marcos, R., and C. De La Fuente Marcos. 2008. "From Star Complexes to the Field: Open Cluster Families." *The Astrophysical Journal* 672, no. 1: 342–351.

de Lima, M. S., L. H. Duczmal, J. C. Neto, and L. P. Pinto. 2015. "Spatial Scan Statistics for Models with Overdispersion and Inflated Zeros." *Statistica Sinica* 25: 225–241.

Dwass, M. 1957. "Modified Randomization Tests for Nonparametric Hypotheses." *Annals of Mathematical Statistics* 28, no. 1: 181–187.

Fine, J. P., and R. J. Gray. 1999. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association* 94, no. 446: 496–509.

Frévent, C., M.-S. Ahmed, M. Marbac, and M. Genin. 2021. "Detecting Spatial Clusters in Functional Data: New Scan Statistic Approaches." *Spatial Statistics* 46: 100550.

Frévent, C., M.-S. Ahmed, S. Dabo-Niang, and M. Genin. 2023. "Investigating Spatial Scan Statistics for Multivariate Functional Data." *Journal of the Royal Statistical Society Series C* 72: 450–475.

Fu, E. L., M. Evans, J.-J. Carrero, et al. 2021. "Timing of Dialysis Initiation to Reduce Mortality and Cardiovascular Events in Advanced Chronic Kidney Disease: Nationwide Cohort Study." *BMJ* 375: e066306.

Genin, M., M. Fumery, F. Occelli, et al. 2020. "Fine-Scale Geographical Distribution and Ecological Risk Factors for Crohn's Disease in France (2007-2014)." *Alimentary Pharmacology & Therapeutics* 51, no. 1: 139–148.

Green, C., L. Elliott, C. Beaudoin, and C. N. Bernstein. 2006. "A Population-Based Ecologic Study of Inflammatory Bowel Disease: Searching for Etiologic Clues." *American Journal of Epidemiology* 164, no. 7: 615–623.

Gregorio, D. I., L. Huang, L. M. DeChello, H. Samociuk, and M. Kulldorff. 2007. "Place of Residence Effect on Likelihood of Surviving Prostate Cancer." *Annals of Epidemiology* 17, no. 7: 520–524.

Hallan, S. I., K. Matsushita, Y. Sang, et al. 2012. "Age and Association of Kidney Measures With Mortality and End-Stage Renal Disease." *JAMA* 308, no. 22: 2349–2360.

Henry, K. A., X. Niu, and F. P. Boscoe. 2009. "Geographic Disparities in Colorectal Cancer Survival." *International Journal of Health Geographics* 8, no. 1: 1–13.

Hougaard, P. 2000. *Shared Frailty Models*, 215–262. New York, NY: Springer New York.

Huang, L., M. Kulldorff, and D. Gregorio. 2007. "A Spatial Scan Statistic for Survival Data." *Biometrics* 63, no. 1: 109–118.

Jeffreys, H. 1961. *Theory of Probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jung, I. 2009. "A Generalized Linear Models Approach to Spatial Scan Statistics for Covariate Adjustment." *Statistics in Medicine* 28, no. 7: 1131–1143.

Jung, I., M. Kulldorff, and A. C. Klassen. 2007. "A Spatial Scan Statistic for Ordinal Data." *Statistics in Medicine* 26, no. 7: 1594–1607.

Khan, M. M., S. Roberson, K. Reid, M. Jordan, and A. Odoi. 2021. "Geographic Disparities and Temporal Changes of Diabetes Prevalence and Diabetes Self-Management Education Program Participation in Florida." *PloS ONE* 16, no. 7: e0254579.

Kleinbaum, D. G., and M. Klein. 2012. "Recurrent Event Survival Analysis." In *Survival Analysis*, 363–423. Berlin: Springer.

Kulldorff, M. 1997. "A Spatial Scan Statistic." *Communications in Statistics—Theory and Methods* 26: 1481–1496.

Kulldorff, M., L. Huang, and K. Konty. 2009. "A Scan Statistic for Continuous Data Based on the Normal Probability Model." *International Journal of Health Geographics* 8: 58.

Kulldorff, M., L. Huang, L. Pickle, and L. Duczmal. 2006. "An Elliptic Spatial Scan Statistic." *Statistics in Medicine* 25: 3929–3943.

Kulldorff, M., F. Mostashari, L. Duczmal, W. Katherine Yih, K. Kleinman, and R. Platt. 2007. "Multivariate Scan Statistics for Disease Surveillance." *Statistics in Medicine* 26, no. 8: 1824–1833.

Kulldorff, M., and N. Nagarwalla. 1995. "Spatial Disease Clusters: Detection and Inference." *Statistics in Medicine* 14, no. 8: 799–810.

Lee, J., Y. Sun, and H. H. Chang. 2020. "Spatial Cluster Detection of Regression Coefficients in a Mixed-Effects Model." *Environmetrics* 31, no. 2: e2578.

Leiser, C. L., M. Taddie, R. Hemmert, et al. 2020. "Spatial Clusters of Cancer Incidence: Analyzing 1940 Census Data Linked to 1966–2017 Cancer Records." *Cancer Causes & Control* 31, no. 7: 609–615.

Leroux, B. G., X. Lei, and N. Breslow. 2000. "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence." In *Statistical Models in Epidemiology, The Environment, and Clinical Trials*, 179–191. Berlin: Springer.

Li, Y. 2009. "Modeling and Analysis of Spatially Correlated Data." In *New Developments in Biostatistics and Bioinformatics*, edited by J. Fan, X. Lin, and J. S. Liu, 72–98. Frontiers of Statistics, vol. 1: Singapore: WorldScientific.

Li, Y., and L. Ryan. 2002. "Modeling Spatial Survival Data Using Semiparametric Frailty Models." *Biometrics* 58, no. 2: 287–297.

Liang, K.-Y., S. G. Self, K. J. Bandeen-Roche, and S. L. Zeger. 1995. "Some Recent Developments for Regression Analysis of Multivariate Failure Time Data." *Lifetime Data Analysis* 1, no. 4: 403–415.

Lin, P.-S. 2014. "Generalized Scan Statistics for Disease Surveillance." *Scandinavian Journal of Statistics* 41, no. 3: 791–808.

Loh, J. M., and Z. Zhu. 2007. "Accounting for Spatial Correlation in the Scan Statistic." *The Annals of Applied Statistics* 1, no. 2: 560–584.

Marciano, L. H. S. C., A. de Faria Fernandes Belone, P. S. Rosa, et al. 2018. "Epidemiological and Geographical Characterization of Leprosy in a Brazilian Hyperendemic Municipality." *Cadernos de Saude Publica* 34, no. 8: e00197216.

Minamisava, R., S. S. Nouer, O. L. de Morais Neto, L. K. Melo, and A. L. S. Andrade. 2009. "Spatial Clusters of Violent Deaths in a Newly Urbanized Region of Brazil: Highlighting the Social Disparities." *International Journal of Health Geographics* 8, no. 1: 1–10.

Montez-Rath, M. E., K. Kapphahn, M. B. Mathur, A. A. Mitani, D. J. Hendry, and M. Desai. 2017. "Guidelines for Generating Right-Censored Outcomes from a Cox Model Extended to Accommodate Time-Varying Covariates." *Journal of Modern Applied Statistical Methods* 16, no. 1: 86–106.

Neill, D., A. Moore, and G. Cooper. 2005. "A Bayesian Spatial Scan Statistic." *Advances in Neural Information Processing Systems* 18: 1003–1010.

Neill, D. B., and G. F. Cooper. 2010. "A Multivariate Bayesian Scan Statistic for Early Event Detection and Characterization." *Machine Learning* 79, no. 3: 261–282.

Ojiambo, P., and E. Kang. 2013. "Modeling Spatial Frailties in Survival Analysis of Cucurbit Downy Mildew Epidemics." *Phytopathology* 103, no. 3: 216–227.

Rondeau, V. 2010. "Statistical Models for Recurrent Events and Death: Application to Cancer Events." *Mathematical and Computer Modelling* 52, no. 7–8: 949–955.

Rue, H., S. Martino, and N. Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71, no. 2: 319–392.

Shi, G., J. Liu, and X. Zhong. 2022. "Spatial and Temporal Variations of pm2.5 Concentrations in Chinese Cities During 2015-2019." *International Journal of Environmental Health Research* 32, no. 12: 2695–2707.

Smida, Z., L. Cucala, A. Gannoun, and G. Durif. 2022. "A Wilcoxon-Mann-Whitney Spatial Scan Statistic for Functional Data." *Computational Statistics & Data Analysis* 167: 107378.

Tango, T., and K. Takahashi. 2005. "A Flexibly Shaped Spatial Scan Statistic for Detecting Clusters." *International Journal of Health Geographics* 4, no. 1: 1–15.

Usman, I., and R. J. Rosychuk. 2018. "A Log-Weibull Spatial Scan Statistic for Time to Event Data." *International Journal of Health Geographics* 17, no. 1: 1–12.

Wan, L., Y. Sun, I. Lee, W. Zhao, and F. Xia. 2020. "Industrial Pollution Areas Detection and Location via Satellite-Based Iiot." *IEEE Transactions on Industrial Informatics* 17, no. 3: 1785–1794.

Wan, N., F. B. Zhan, Y. Lu, and J. P. Tiefenbacher. 2012. "Access to Healthcare and Disparities in Colorectal Cancer Survival in Texas." *Health & Place* 18, no. 2: 321–329.

Yin, P., and L. Mu. 2018. "A Hybrid Method for Fast Detection of Spatial Disease Clusters in Irregular Shapes." *GeoJournal* 83, no. 4: 693–705.

Zhang, T., Z. Zhang, and G. Lin. 2012. "Spatial Scan Statistics with Overdispersion." *Statistics in Medicine* 31, no. 8: 762–774.

Zhou, R., L. Shu, and Y. Su. 2015. "An Adaptive Minimum Spanning Tree Test for Detecting Irregularly-Shaped Spatial Clusters." *Computational Statistics & Data Analysis* 89: 134–146.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.