AIMS Mathematics, 10(6): 14088–14101.
DOI: 10.3934/math.2025634
Received: 31 March 2025
Revised: 10 June 2025
Accepted: 16 June 2025
Published: 19 June 2025

https://www.aimspress.com/journal/Math

Review

A review of spatial scan statistics for survival data

Camille Frévent*

Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France

* **Correspondence:** Email: camille.frevent@univ-lille.fr.

Abstract: We propose a review of spatial scan statistics approaches in the context of survival data. After presenting the general principle of spatial scan statistics, we review the literature and find that few approaches exist. We distinguish between the first parametric approaches, based on a specific distribution model and therefore not very flexible, and a semi-parametric method. However, these approaches do not allow taking into account the spatial dependence frequently observed in the data. We then present a more recent approach allowing us to take them into account. Finally, we describe the adjustment of cluster detection on covariates before illustrating the methods on the detection of abnormal survival time clusters following the diagnosis of leukemia.

Keywords: cluster detection; scan statistic; spatial data; survival; time-to-event data
Mathematics Subject Classification: 62H30, 62M30, 62N03

1. Introduction

In epidemiology, the detection of spatial clusters enables the identification of geographical areas presenting an abnormally high (or low) risk. Spatial scan statistics are a well-known method for objectively detecting spatial clusters, and providing an indication of their statistical significance.

In the context of health data, spatial scan statistics based on Poisson and Bernoulli models have been proposed by Kulldorff and Nagarwalla [1] and Kulldorff [2]. From observed cases and at-risk populations, these can be used, for example, to detect geographical areas where the risk of having or dying from a disease is higher than elsewhere.

A spatial scan statistic based on an ordinal model has also been proposed by Jung et al. [3] and enables the detection of spatial clusters in which the stages of patients suffering from a disease are more severe than elsewhere.

Researchers may also be interested in detecting areas where the time to an event of interest (typically death or recovery from a disease) is higher or lower than elsewhere. In this case, the data observed is

not a number of cases and a number of people at risk, but for each individual, a time to the event of interest. However, some individuals may never experience the event during the study, in which case they are said to be censored.

This article provides a review of the literature on spatial scan statistics methods for survival data. Section 2 introduces the general principle of spatial scan statistics, while Section 3 presents the different methods proposed in the literature for survival data. Section 4 explains how to adjust cluster detection on covariates and illustrates the approaches on the `LeukSurv` dataset. Finally, Section 5 concludes the article.

2. Spatial scan statistics

Let $s_1, \ldots, s_K$ be K nonoverlapping spatial locations of an observation domain $\mathcal{S} \subset \mathbb{R}^2$. Spatial scan statistics aim at detecting spatial clusters in which the distribution of the observed data is different than elsewhere, as well as determining their statistical significance. More precisely, they consider the following test hypotheses:

$\mathcal{H}_0$: There is no cluster in the data,

vs.

$\mathcal{H}_1$: There is at least one spatial cluster in which the data present abnormal values compared with the rest of the domain.

These test hypotheses can be expressed more explicitly, depending on the type of data and model considered. They will be clarified for each survival data model in the following.

The spatial scan statistic approach is a two-stage process. The first stage is the scanning step. It consists in determining the most likely cluster (MLC) from a set of potential clusters $\mathcal{W}$. Here, we will consider the set of circular clusters containing between 1 and 50% of observations, but it should be noted that other approaches exist, especially elliptical clusters [4], graph-based [5] or arbitrarily-shaped clusters [6, 7]. Next, a concentration index is computed for each potential cluster $w \in \mathcal{W}$. It compares the distribution within the potential cluster with that outside, so that the greater the difference, the higher the concentration index. Finally, the spatial scan statistic Λ is defined as the maximum of the concentration index over $\mathcal{W}$, and the MLC is the potential cluster for which this maximum is reached.

The second step is to assess the statistical significance of the MLC. Since the distribution of the spatial scan statistic is generally impossible to determine under $\mathcal{H}_0$, two Monte Carlo approaches are commonly used: (i) M permutations of the data are generated, and the scan statistic $\Lambda^{(m)}$ is calculated on each permuted dataset $m \in \{1, \ldots, M\}$ [8–10]; (ii) if the distribution of the data is known under $\mathcal{H}_0$, M datasets are generated under $\mathcal{H}_0$, and the scan statistic is calculated on each generated dataset [1–3]. In both approaches, the p-value is then estimated by

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

3. Spatial scan statistics for survival data

In the context of spatial survival data, spatial scan statistics allow researchers to identify risk or protective factors related to a study event. The test hypotheses are the following:

$\mathcal{H}_0$: There is no cluster of abnormal survival times,

vs.

$\mathcal{H}_1$: There is at least one cluster $w \in \mathcal{W}$ of abnormal survival times.

We can also define the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w as

$\mathcal{H}_1^{(w)}$: $w \in \mathcal{W}$ is a cluster of abnormal survival times.

Then, $\mathcal{H}_1 = \bigcup_{w \in \mathcal{W}} \mathcal{H}_1^{(w)}$.

Let $i_1^{(1)}, \ldots, i_{N_1}^{(1)}, \ldots, i_1^{(K)}, \ldots, i_{N_K}^{(K)}$ be the observed individuals in $s_1, \ldots, s_K$, where $i_n^{(k)}$ corresponds to the n^{th} individual in spatial unit s_k. For each individual $i_n^{(k)}$, we observe survival data consisting of an observed delay $T_{i_n^{(k)}}$ and a censoring indicator $\delta_{i_n^{(k)}}$ (equal to 1 if $T_{i_n^{(k)}}$ corresponds to the true delay until the event, 0 otherwise, which corresponds to censoring). In the following, we only consider right-censoring (assuming that the event of interest could not have occurred before the beginning of the study). Censoring is assumed to be uninformative, and the event times are assumed to be independent of the censoring times.

This section presents the different scan statistics for survival data proposed in the literature, as well as their limitations.

3.1. Parametric models

Several parametric approaches have been proposed in the literature. Huang et al. [11] first proposed a scan statistic assuming that the true (but not necessarily observed) survival times $Y_{i_n^{(k)}}$ follow an exponential model. The test hypotheses can be rewritten as

$$\mathcal{H}_0\text{: For all } k \in \{1, \ldots, K\}, n \in \{1, \ldots, N_k\}, Y_{i_n^{(k)}} \sim \mathcal{E}\left(\tfrac{1}{\theta}\right),$$

vs.

$$\mathcal{H}_1\text{: There exists } w \in \mathcal{W} \text{ such that for all } i_n^{(k)} \text{ so that } s_k \in w,\ Y_{i_n^{(k)}} \sim \mathcal{E}\left(\tfrac{1}{\theta_w}\right), \text{ and for all } i_n^{(k)} \text{ so that } s_k \in w^c,\ Y_{i_n^{(k)}} \sim \mathcal{E}\left(\tfrac{1}{\theta_{w^c}}\right),\ \theta_w \neq \theta_{w^c}.$$

The alternative hypothesis associated with a potential cluster w is then

$$\mathcal{H}_1^{(w)}\text{: For all } i_n^{(k)} \text{ so that } s_k \in w,\ Y_{i_n^{(k)}} \sim \mathcal{E}\left(\tfrac{1}{\theta_w}\right), \text{ for all } i_n^{(k)} \text{ so that } s_k \in w^c,\ Y_{i_n^{(k)}} \sim \mathcal{E}\left(\tfrac{1}{\theta_{w^c}}\right),\ \theta_w \neq \theta_{w^c}.$$

Next, the log-likelihood under $\mathcal{H}_0$ is

$$\ell_{\mathcal{H}_0}(\theta) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \left[-\delta_{i_n^{(k)}} \ln(\theta) - \frac{T_{i_n^{(k)}}}{\theta} \right],$$

which is maximized when $\hat{\theta} = \dfrac{\sum_{k=1}^{K} \sum_{n=1}^{N_k} T_{i_n^{(k)}}}{\sum_{k=1}^{K} \sum_{n=1}^{N_k} \delta_{i_n^{(k)}}}$, and the log-likelihood under $\mathcal{H}_1^{(w)}$ is defined as

$$\ell_{\mathcal{H}_1^{(w)}}(\theta_w, \theta_{w^c}) = \sum_{i_n^{(k)}, s_k \in w} \left[-\delta_{i_n^{(k)}} \ln(\theta_w) - \frac{T_{i_n^{(k)}}}{\theta_w} \right] + \sum_{i_n^{(k)}, s_k \in w^c} \left[-\delta_{i_n^{(k)}} \ln(\theta_{w^c}) - \frac{T_{i_n^{(k)}}}{\theta_{w^c}} \right],$$

which is maximized when $\hat{\theta}_w = \dfrac{\sum\limits_{i_n^{(k)}, s_k \in w} T_{i_n^{(k)}}}{\sum\limits_{i_n^{(k)}, s_k \in w} \delta_{i_n^{(k)}}}$ and $\hat{\theta}_{w^c} = \dfrac{\sum\limits_{i_n^{(k)}, s_k \in w^c} T_{i_n^{(k)}}}{\sum\limits_{i_n^{(k)}, s_k \in w^c} \delta_{i_n^{(k)}}}$.

Then the spatial scan statistic is defined as

$$\Lambda^{\exp} = \max_{w \in \mathcal{W}} \ell_{\mathcal{H}_1^{(w)}}(\hat{\theta}_w, \hat{\theta}_{w^c}) - \ell_{\mathcal{H}_0}(\hat{\theta}).$$

However, the exponential distribution is somewhat too simplistic in reality, since it assumes a constant hazard rate over time. An alternative parametric model is the Weibull model, which allows for increasing or decreasing hazard rate over time. Bhatt and Tiwari [12] proposed a scan statistic in this context, where the test hypotheses can be rewritten as

$$\mathcal{H}_0\text{: For all } k \in \{1, \ldots, K\}, n \in \{1, \ldots, N_k\}, Y_{i_n^{(k)}} \sim \text{Wei}\left(\tfrac{1}{\theta}, \alpha\right),$$
$$\text{vs.}$$
$$\mathcal{H}_1\text{: There exists } w \in \mathcal{W} \text{ such that for all } i_n^{(k)} \text{ so that } s_k \in w,\ Y_{i_n^{(k)}} \sim \text{Wei}\left(\tfrac{1}{\theta_w}, \alpha_w\right), \text{ and for all } i_n^{(k)} \text{ so that } s_k \in w^c,\ Y_{i_n^{(k)}} \sim \text{Wei}\left(\tfrac{1}{\theta_{w^c}}, \alpha_{w^c}\right),\ \theta_w \neq \theta_{w^c}.$$

The log-likelihood under $\mathcal{H}_0$ is

$$\ell_{\mathcal{H}_0}(\theta, \alpha) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \left\{ \delta_{i_n^{(k)}} \left[\ln(\alpha) + (\alpha - 1) \ln(T_{i_n^{(k)}}) - \ln(\theta) \right] - \frac{T_{i_n^{(k)}}^{\alpha}}{\theta} \right\},$$

which is maximized when $\hat{\theta} = \dfrac{\sum\limits_{k=1}^{K} \sum\limits_{n=1}^{N_k} T_{i_n^{(k)}}^{\hat{\alpha}}}{\sum\limits_{k=1}^{K} \sum\limits_{n=1}^{N_k} \delta_{i_n^{(k)}}}$, and the log-likelihood under $\mathcal{H}_1^{(w)}$ is defined as

$$\begin{aligned} \ell_{\mathcal{H}_1^{(w)}}(\theta_w, \theta_{w^c}, \alpha_w, \alpha_{w^c}) = & \sum_{i_n^{(k)}, s_k \in w} \left[\delta_{i_n^{(k)}} \left(\ln(\alpha_w) + (\alpha_w - 1) \ln(T_{i_n^{(k)}}) - \ln(\theta_w) \right) - \frac{T_{i_n^{(k)}}^{\alpha_w}}{\theta_w} \right] \\ & + \sum_{i_n^{(k)}, s_k \in w^c} \left[\delta_{i_n^{(k)}} \left(\ln(\alpha_{w^c}) + (\alpha_{w^c} - 1) \ln(T_{i_n^{(k)}}) - \ln(\theta_{w^c}) \right) - \frac{T_{i_n^{(k)}}^{\alpha_{w^c}}}{\theta_{w^c}} \right], \end{aligned}$$

which is maximized when $\hat{\theta}_w = \dfrac{\sum\limits_{i_n^{(k)}, s_k \in w} T_{i_n^{(k)}}^{\hat{\alpha}_w}}{\sum\limits_{i_n^{(k)}, s_k \in w} \delta_{i_n^{(k)}}}$ and $\hat{\theta}_{w^c} = \dfrac{\sum\limits_{i_n^{(k)}, s_k \in w^c} T_{i_n^{(k)}}^{\hat{\alpha}_{w^c}}}{\sum\limits_{i_n^{(k)}, s_k \in w^c} \delta_{i_n^{(k)}}}$.

For α, α_w and α_{w^c}, the expressions of the maximum likelihood estimators are more complicated, and, since $\hat{\alpha}, \hat{\alpha}_w$ and $\hat{\alpha}_{w^c}$ appear in the formulas of $\hat{\theta}, \hat{\theta}_w$ and $\hat{\theta}_{w^c}$, we can in practice use an optimization algorithm to estimate all the parameters.

Finally, the spatial scan statistic is

$$\Lambda^{\text{Wei}} = \max_{w\in\mathcal{W}} \ell_{\mathcal{H}_1^{(w)}}(\hat{\theta}_w, \hat{\theta}_{w^c}, \hat{\alpha}_w, \hat{\alpha}_{w^c}) - \ell_{\mathcal{H}_0}(\hat{\theta}, \hat{\alpha}).$$

These models have been generalized by Bhatt and Tiwari [13] to any density function for the $Y_{i_n^{(k)}}$ of the form

$$f(t;\gamma,a,b,c) = \frac{ct^{ac-1}\exp\left(-\frac{t^c}{\gamma^b}\right)}{\gamma^{ab}\Gamma(a)}, t>0,$$

where $a, b, c > 0$ are known and γ is to be estimated from the data.

In this context, the test hypotheses are

$\mathcal{H}_0$: For all $i_n^{(k)}$, the density function of $Y_{i_n^{(k)}}$ is $f(\,.\,;\gamma,a,b,c)$,

vs.

$\mathcal{H}_1$: There exists $w \in \mathcal{W}$ such that for all $i_n^{(k)}$ so that $s_k \in w$, the density function of $Y_{i_n^{(k)}}$ is $f(\,.\,;\gamma_w,a,b,c)$ and for all $i_n^{(k)}$ so that $s_k \in w^c$, the density function of $Y_{i_n^{(k)}}$ is $f(\,.\,;\gamma_{w^c},a,b,c)$, $\gamma_w \neq \gamma_{w^c}$.

γ, γ_w and γ_{w^c} can be estimated as in previous models using the maximum likelihood estimators and then the scan statistic is

$$\Lambda^{\text{gen}} = \max_{w\in\mathcal{W}} \ell_{\mathcal{H}_1^{(w)}}(\hat{\gamma}_w, \hat{\gamma}_{w^c}) - \ell_{\mathcal{H}_0}(\hat{\gamma}).$$

It should be noted that if $a = b = c = 1$, this approach is equivalent to the exponential model; if $a = 1$ and $b = c$, it results in a Weibull model; and if $a = b = 1$ and $c = 2$, it is equivalent to a Rayleigh model.

More recently, Usman and Rosychuk [14] proposed an approach based on a log-Weibull distribution and considering the following test hypotheses:

$\mathcal{H}_0$: For all $i_n^{(k)}$, the density function of $Y_{i_n^{(k)}}$ is of the form $f(t\,;a,b) = \frac{1}{b}\exp\left(\frac{t-a}{b}\right)\exp\left[-\exp\left(\frac{t-a}{b}\right)\right]$,

vs.

$\mathcal{H}_1$: There exists $w \in \mathcal{W}$ such that for all $i_n^{(k)}$ so that $s_k \in w$, the density function of $Y_{i_n^{(k)}}$ is of the form $f(t\,;a_w,b_w) = \frac{1}{b_w}\exp\left(\frac{t-a_w}{b_w}\right)\exp\left[-\exp\left(\frac{t-a_w}{b_w}\right)\right]$ and for all $i_n^{(k)}$ so that $s_k \in w^c$, the density function of $Y_{i_n^{(k)}}$ is of the form $f(t\,;a_{w^c},b_{w^c}) = \frac{1}{b_{w^c}}\exp\left(\frac{t-a_{w^c}}{b_{w^c}}\right)\exp\left[-\exp\left(\frac{t-a_{w^c}}{b_{w^c}}\right)\right]$, $b_w \neq b_{w^c}$.

The log-likelihoods under $\mathcal{H}_0$ and under $\mathcal{H}_1^{(w)}$ are, respectively,

$$\ell_{\mathcal{H}_0}(a,b) = \sum_{k=1}^{K}\sum_{n=1}^{N_k}\left[\delta_{i_n^{(k)}}\left(-\ln(b) + \frac{T_{i_n^{(k)}}-a}{b}\right) - \exp\left(\frac{T_{i_n^{(k)}}-a}{b}\right)\right]$$

and

$$\ell_{\mathcal{H}_1^{(w)}}(a_w, a_{w^c}, b_w, b_{w^c}) = \sum_{i_n^{(k)}, s_k \in w} \left[\delta_{i_n^{(k)}} \left(-\ln(b_w) + \frac{T_{i_n^{(k)}} - a_w}{b_w} \right) - \exp\left(\frac{T_{i_n^{(k)}} - a_w}{b_w} \right) \right] + \sum_{i_n^{(k)}, s_k \in w^c} \left[\delta_{i_n^{(k)}} \left(-\ln(b_{w^c}) + \frac{T_{i_n^{(k)}} - a_{w^c}}{b_{w^c}} \right) - \exp\left(\frac{T_{i_n^{(k)}} - a_{w^c}}{b_{w^c}} \right) \right].$$

And the spatial scan statistic is

$$\Lambda^{\text{log-Wei}} = \max_{w \in \mathcal{W}} \ell_{\mathcal{H}_1^{(w)}}(\hat{a}_w, \hat{a}_{w^c}, \hat{b}_w, \hat{b}_{w^c}) - \ell_{\mathcal{H}_0}(\hat{a}, \hat{b}).$$

Once the spatial scan statistic $\Lambda \in \{\Lambda^{\text{exp}}, \Lambda^{\text{Wei}}, \Lambda^{\text{gen}}, \Lambda^{\text{log-Wei}}\}$ is computed, the MLC is defined as the potential cluster of $\mathcal{W}$ corresponding to this maximum. The statistical significance of the MLC is then determined using a Monte Carlo procedure with permutations of the individuals (that is, the $(T_{i_n^{(k)}}, \delta_{i_n^{(k)}})$).

Although these approaches are based on conventional models for survival data, they remain parametric and are therefore less flexible than nonparametric or semiparametric approaches. Thus, a method based on a Cox model has been developed by Cook et al. [15].

3.2. Cox model

Cook et al. [15] proposed a spatial scan statistic based on a Cox model, which presents the advantage of not assuming a distribution for the data.

They considered the following Cox model on the hazard function λ for a potential cluster w:

$$\lambda_{i_n^{(k)}}^{(w)}(t) = \lambda_0^{(w)}(t) \exp(\alpha_w \mathbb{1}_{s_k \in w}).$$

In the context of cluster detection, the test hypotheses are

$$\mathcal{H}_0\text{: For all } w \in \mathcal{W}, \alpha_w = 0, \text{ that is for all } i_n^{(k)}, \lambda_{i_n^{(k)}}(t) = \lambda_0(t),$$
vs.
$$\mathcal{H}_1\text{: There exists } w \in \mathcal{W} \text{ such that } \alpha_w \neq 0, \text{ that is there exists } w \in \mathcal{W}, \text{ such that for all } i_n^{(k)} \text{ so that } s_k \in w, \lambda_{i_n^{(k)}}(t) = \lambda_0^{(w)}(t) \exp(\alpha_w), \text{ for all } i_n^{(k)} \text{ so that } s_k \in w^c, \lambda_{i_n^{(k)}}(t) = \lambda_0^{(w)}(t).$$

In this context, the partial log-likelihood under $\mathcal{H}_1^{(w)}$ is

$$\ell_{\mathcal{H}_1^{(w)}}(\alpha_w) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \delta_{i_n^{(k)}} \left[\alpha_w \mathbb{1}_{s_k \in w} - \ln \left(\sum_{l=1}^{K} \sum_{\substack{m=1 \\ T_{i_m^{(l)}} \geq T_{i_n^{(k)}}}}^{N_l} \exp(\alpha_w \mathbb{1}_{s_l \in w}) \right) \right].$$

In order to test $\mathcal{H}_0 : \alpha_w = 0$ vs. $\mathcal{H}_1^{(w)} : \alpha_w \neq 0$, Cook et al. [15] proposed to use the score statistic defined as

$$LR^{(w)} = \frac{U(0)}{\sqrt{I(0)}},$$

where $U(\alpha_w) = \dfrac{\partial \ell_{\mathcal{H}_1^{(w)}}(\alpha_w)}{\partial \alpha_w}$ and $I(\alpha_w) = -\mathbb{E}\left(\dfrac{\partial U(\alpha_w)}{\partial \alpha_w}\right)$. We obtain

$$U(0) = \sum_{k=1}^{K}\sum_{n=1}^{N_k} \delta_{i_n^{(k)}} \left[\mathbb{1}_{s_k \in w} - \frac{\text{Card}\left(\left\{i_m^{(l)}, s_l \in w, T_{i_m^{(l)}} \geq T_{i_n^{(k)}}\right\}\right)}{\text{Card}\left(\left\{i_m^{(l)}, T_{i_m^{(l)}} \geq T_{i_n^{(k)}}\right\}\right)} \right],$$

$$I(0) = \sum_{k=1}^{K}\sum_{n=1}^{N_k} \delta_{i_n^{(k)}} \left\{ \frac{\text{Card}\left(\left\{i_m^{(l)}, s_l \in w, T_{i_m^{(l)}} \geq T_{i_n^{(k)}}\right\}\right)}{\text{Card}\left(\left\{i_m^{(l)}, T_{i_m^{(l)}} \geq T_{i_n^{(k)}}\right\}\right)} - \left[\frac{\text{Card}\left(\left\{i_m^{(l)}, s_l \in w, T_{i_m^{(l)}} \geq T_{i_n^{(k)}}\right\}\right)}{\text{Card}\left(\left\{i_m^{(l)}, T_{i_m^{(l)}} \geq T_{i_n^{(k)}}\right\}\right)} \right]^2 \right\},$$

and the spatial scan statistic and the MLC are defined as $\Lambda^{\text{Cox}} = \max\limits_{w \in \mathcal{W}} |LR^{(w)}|$ and $\text{MLC}^{\text{Cox}} = \arg\max\limits_{w \in \mathcal{W}} |LR^{(w)}|$, respectively.

Finally, the statistical significance of the MLC is determined as previously, by permuting the individuals (that is, the $(T_{i_n^{(k)}}, \delta_{i_n^{(k)}})$).

The spatial scan statistics described until now make the conventional assumption of independence of the observations. However, this assumption is very strong and rather unrealistic in practice, since the spatial nature of the observations leads to potential spatial autocorrelation, as specified by Tobler's first law of geography [16]. Moreover, for confidentiality reasons, survival data are often only available on an aggregated spatial level. Thus, we can distinguish two phenomena: (i) the survival times of individuals located in the same spatial unit may be correlated (intra-spatial unit correlation), for example, due to similar healthcare supply, and (ii) there may be the presence of spatial dependence at the level of spatial units. Thus, Frévent et al. [17] proposed a spatial scan statistic based on a Cox model with spatially correlated shared frailties. This takes into account both of the above-mentioned phenomena.

3.3. Cox model with spatially correlated shared frailties

Frévent et al. [17] considered the following Cox model with shared frailties:

$$\text{for all } i_n^{(k)} \text{ within spatial unit } s_k,\ \lambda_{i_n^{(k)}}(t|\varphi_k) = \lambda_0(t)\exp(\varphi_k),$$

where $\varphi_1, \ldots, \varphi_K$ are the shared frailties associated with the spatial locations $s_1, \ldots, s_K$, respectively, and include the cluster effect.

Thus, Frévent et al. [17] decomposed the frailties into two terms:

$$\text{for a potential cluster } w,\ \varphi_k^{(w)} = \alpha_w \mathbb{1}_{s_k \in w} + X_k \text{ where } \mathbb{E}(X_k) = 0.$$

The test hypotheses can be written as

$$\begin{gathered} \mathcal{H}_0 : \forall w \in \mathcal{W}, \alpha_w = 0, \\ \text{vs.} \\ \mathcal{H}_1 : \exists w \in \mathcal{W}, \alpha_w \neq 0. \end{gathered}$$

The shared frailties allow us to take into account the potential intra-spatial unit correlation. To take into account the potential spatial dependence between the spatial units, a spatial model, namely the

conditional autoregressive (CAR) model, is assumed on the X_k:

$$X_k|\{X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_K\} \sim \mathcal{N}\left(\frac{\rho\sum_{l=1}^{K} w_{k,l}X_l}{\rho\sum_{l=1}^{K} w_{k,l}X_l + 1 - \rho}, \frac{\sigma_X^2}{\rho\sum_{l=1}^{K} w_{k,l}X_l + 1 - \rho}\right),$$

where $\rho \in [0,1]$ is the spatial dependence parameter, and $w_{k,l} = 1$ if s_k and s_l share a common boundary and 0 if not. It should be noted that if ρ is assumed to be 0, then the model takes into account the intra-spatial unit correlation but not spatial dependence.

The proposed method is decomposed into two stages. The first one consists in estimating the φ_k and ρ. To this end, Frévent et al. [17] proposed to estimate the φ_k and ρ under $\mathcal{H}_0$ and under each alternative hypothesis $\mathcal{H}_1^{(w)}$, and then to extract the estimates $\{\varphi_1^*,\ldots,\varphi_K^*,\rho^*\}$ associated with the "best model" according to the Bayes factor criterion, i.e., under $\mathcal{H}_0$ if the Bayes factors comparing the models under each $\mathcal{H}_1^{(w)}$ to the model under $\mathcal{H}_0$ never exceed 30, and the model under $\mathcal{H}_1^{(w)}$ associated with the highest Bayes factor otherwise.

Next, the second stage consists in computing a scan statistic on the φ_k^*. At this stage, the test hypotheses can be rewritten on $\boldsymbol{\varphi}^* = (\varphi_1^*,\ldots,\varphi_K^*)^\top$ as

$$\mathcal{H}_0 : \boldsymbol{\varphi}^* \sim \mathcal{N}(\alpha\mathbb{1}, \sigma^{2(0)}A^{-1}),$$
$$\text{vs.}$$
$$\mathcal{H}_1 : \exists w \in \mathcal{W}, \boldsymbol{\varphi}^* \sim \mathcal{N}(\alpha_w\mathbb{1}_w + \alpha_{w^c}\mathbb{1}_{w^c}, \sigma^{2(w)}A^{-1}), \alpha_w \neq \alpha_{w^c},$$

where $\mathbb{1}$, $\mathbb{1}_w$ and $\mathbb{1}_{w^c}$ correspond, respectively, to the column vector composed only of 1, the column vector composed of 1 for the locations in w and 0 otherwise, and the column vector composed of 1 for the locations outside w and 0 otherwise. A is a squared matrix that results in the variance-covariance structure of the CAR model (see [17] for more details).

Then, the spatial scan statistic and the MLC are defined as

$$\Lambda^{\text{frail.Cox}} = \max_{w\in\mathcal{W}} \ell_{\mathcal{H}_1^{(w)}}(\hat{\alpha}_w, \hat{\alpha}_{w^c}, \widehat{\sigma^{2(w)}}) - \ell_{\mathcal{H}_0}(\hat{\alpha}, \widehat{\sigma^{2(0)}}) = \max_{w\in\mathcal{W}} \frac{K}{2}\ln\left(\frac{\widehat{\sigma^{2(0)}}}{\widehat{\sigma^{2(w)}}}\right),$$

and

$$\text{MLC}^{\text{frail.Cox}} = \arg\max_{w\in\mathcal{W}} \ell_{\mathcal{H}_1^{(w)}}(\hat{\alpha}_w, \hat{\alpha}_{w^c}, \widehat{\sigma^{2(w)}}) - \ell_{\mathcal{H}_0}(\hat{\alpha}, \widehat{\sigma^{2(0)}}) = \arg\max_{w\in\mathcal{W}} \frac{K}{2}\ln\left(\frac{\widehat{\sigma^{2(0)}}}{\widehat{\sigma^{2(w)}}}\right),$$

respectively, where

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

and

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

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

Since the distribution of the φ_k^* is known, Frévent et al. [17] generates M datasets of the φ_k^* under $\mathcal{H}_0$ to estimate the p-value associated with the MLC. It should be noted that using permutations of the φ_k^* is not possible here, as this would alter the spatial dependence of the data.

4. Including covariates in the models

In many applications it may be relevant to adjust cluster detection on covariates such as the age of individuals. Thus, this section presents the adjustment procedure proposed by the authors.

4.1. Including covariates in the parametric approaches

For the exponential model, Huang et al. [11] considered the following model to adjust for p covariates $Z^{(1)}, \ldots, Z^{(p)}$:

$$\ln\left(Y_{i_n^{(k)}}\right) = \beta_0 + \beta_1 Z^{(1)}_{i_n^{(k)}} + \cdots + \beta_p Z^{(p)}_{i_n^{(k)}} + \varepsilon_{i_n^{(k)}},$$

where $\varepsilon_{i_n^{(k)}}$ is an error term with density function $f_\varepsilon(e) = \exp(e)\exp[-\exp(e)]$. $\beta_0, \ldots, \beta_p$ can be estimated from the $(T_{i_n^{(k)}}, \delta_{i_n^{(k)}})$ using an exponential regression. Next, the observed times are adjusted as

$$T^{\text{adj}}_{i_n^{(k)}} = T_{i_n^{(k)}} \times \exp\left[-\hat{\beta}_1\left(Z^{(1)}_{i_n^{(k)}} - \mu_1\right) - \cdots - \hat{\beta}_p\left(Z^{(p)}_{i_n^{(k)}} - \mu_p\right)\right],$$

where $\mu_j = \min\limits_{i_n^{(k)}} Z^{(j)}_{i_n^{(k)}}$. Finally, the spatial scan statistic is applied on the $(T^{\text{adj}}_{i_n^{(k)}}, \delta_{i_n^{(k)}})$.

For the approaches based on the Weibull model, its generalization, or the log-Weibull model, the authors did not propose any adjustment on the covariates.

4.2. Including covariates in the Cox-based approaches

When the models can directly include covariates, the conventional approach for adjusting cluster detection on covariates in spatial scan statistics is to fit the model under $\mathcal{H}_0$, in order to make the effect of covariates independent of the potential cluster.

In the presence of covariates, the Cox model considered by Cook et al. [15] is as follows:

$$\lambda^{(w)}_{i_n^{(k)}}(t) = \lambda^{(w)}_0(t)\exp\left(\beta_1 Z^{(1)}_{i_n^{(k)}} + \cdots + \beta_p Z^{(p)}_{i_n^{(k)}} + \alpha_w \mathbb{1}_{s_k \in w}\right).$$

Thus, the score statistic is now expressed as

$$LR^{(w)\text{cov}}(\hat{\beta}_1, \ldots, \hat{\beta}_p) = \frac{U^{\text{cov}}(0; \hat{\beta}_1, \ldots, \hat{\beta}_p)}{\sqrt{I^{\text{cov}}(0; \hat{\beta}_1, \ldots, \hat{\beta}_p)}}$$

with

$$U(0; \beta_1, \ldots, \beta_p) = \sum_{k=1}^{K}\sum_{n=1}^{N_k} \delta_{i_n^{(k)}}\left[\mathbb{1}_{s_k \in w} - \frac{\displaystyle\sum_{s_l \in w}\sum_{\substack{m=1 \\ T_{i_m^{(l)}} \geq T_{i_n^{(k)}}}}^{N_l} \exp\left(\beta_1 Z^{(1)}_{i_m^{(l)}} + \cdots + \beta_p Z^{(p)}_{i_m^{(l)}}\right)}{\displaystyle\sum_{l=1}^{K}\sum_{\substack{m=1 \\ T_{i_m^{(l)}} \geq T_{i_n^{(k)}}}}^{N_l} \exp\left(\beta_1 Z^{(1)}_{i_m^{(l)}} + \cdots + \beta_p Z^{(p)}_{i_m^{(l)}}\right)}\right]$$

and

$$I(0;\beta_1,\ldots,\beta_p)=\sum_{k=1}^{K}\sum_{n=1}^{N_k}\delta_{i_n^{(k)}}\left[\frac{\displaystyle\sum_{s_l\in w}\sum_{\substack{m=1\\ T_{i_m^{(l)}}\geq T_{i_n^{(k)}}}}^{N_l}\exp\left(\beta_1 Z_{i_m^{(l)}}^{(1)}+\cdots+\beta_p Z_{i_m^{(l)}}^{(p)}\right)}{\displaystyle\sum_{l=1}^{K}\sum_{\substack{m=1\\ T_{i_m^{(l)}}\geq T_{i_n^{(k)}}}}^{N_l}\exp\left(\beta_1 Z_{i_m^{(l)}}^{(1)}+\cdots+\beta_p Z_{i_m^{(l)}}^{(p)}\right)}-\left(\frac{\displaystyle\sum_{s_l\in w}\sum_{\substack{m=1\\ T_{i_m^{(l)}}\geq T_{i_n^{(k)}}}}^{N_l}\exp\left(\beta_1 Z_{i_m^{(l)}}^{(1)}+\cdots+\beta_p Z_{i_m^{(l)}}^{(p)}\right)}{\displaystyle\sum_{l=1}^{K}\sum_{\substack{m=1\\ T_{i_m^{(l)}}\geq T_{i_n^{(k)}}}}^{N_l}\exp\left(\beta_1 Z_{i_m^{(l)}}^{(1)}+\cdots+\beta_p Z_{i_m^{(l)}}^{(p)}\right)}\right)^2\right].$$

The spatial scan statistic is still defined as $\Lambda^{\text{Cox}}=\max_{w\in\mathcal{W}}|LR^{(w)\text{cov}}(\hat{\beta}_1,\ldots,\hat{\beta}_p)|$.

The adjustment on covariates is performed similarly in the approach based on shared frailties. Frévent et al. [17] considered the following Cox model: for an individual $i_n^{(k)}$ within spatial unit s_k,

$$\lambda_{i_n^{(k)}}\left(t|Z_{i_n^{(k)}}^{(1)},\ldots,Z_{i_n^{(k)}}^{(p)},\varphi_k\right)=\lambda_0(t)\exp\left(\beta_1 Z_{i_n^{(k)}}^{(1)}+\cdots+\beta_p Z_{i_n^{(k)}}^{(p)}+\varphi_k\right).$$

$\beta_1,\ldots,\beta_p$ are estimated under $\mathcal{H}_0$ and fixed to these values in the models under each alternative hypothesis $\mathcal{H}_1^{(w)}$. Next, the estimates $\{\varphi_1^*,\ldots,\varphi_K^*,\rho^*\}$ of $\{\varphi_1,\ldots,\varphi_K,\rho\}$ retained are those obtained with the "best model" according to the Bayes factor criterion, and the scan step is performed on them, as described above.

4.3. Application to the `LeukSurv` *dataset*

In this section, we illustrate the covariate adjustment procedure on the `LeukSurv` dataset studied by Henderson et al. [18] and available in the R package spBayesSurv.

The dataset consists of 1,043 patients with acute myeloid leukemia within 24 districts in northwest England. For each patient, the survival time in days, status (dead or censored), age, sex, white blood cell count at diagnosis (wbc, truncated at 500), Townsend score (tpi, higher values indicate less affluent areas), and district of residence are available. The median survival times are presented in Figure 1.

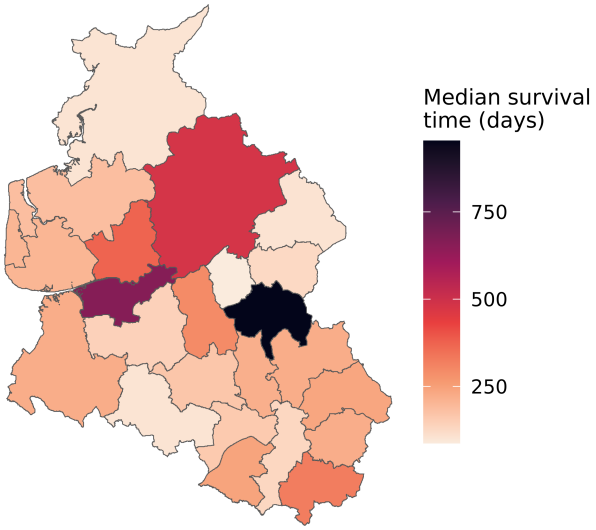

Figure 1. Median survival time for acute myeloid leukemia in each district in the `LeukSurv` dataset.

We applied the exponential scan statistic as well as the Cox-based approach with and without shared frailties, using the adjustment procedure described above. Cluster detection was first adjusted on age, sex and wbc at diagnosis, and then we also adjusted the clusters on the Townsend score. The estimated frailties are presented in Figure 2.

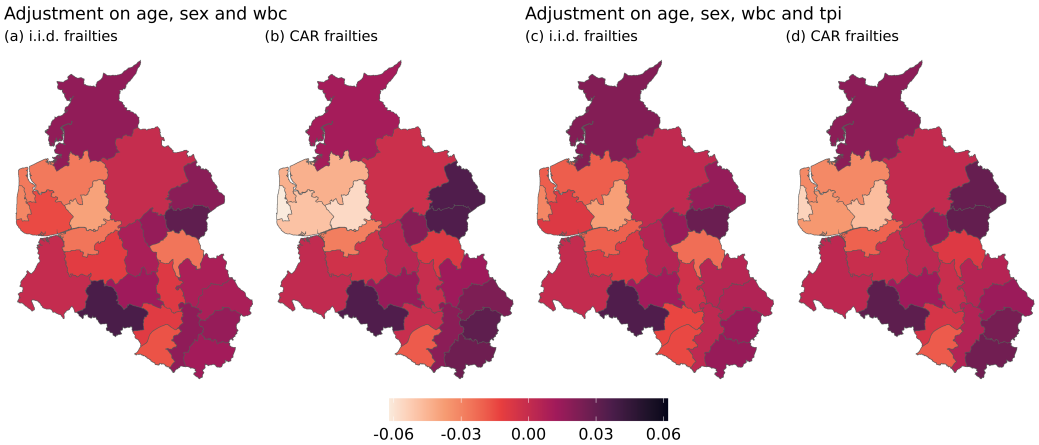

Figure 2. Estimated frailties φ_k^* with the i.i.d. and the CAR models when adjusting cluster detection on age, sex, and wbc (panels (a) and (b)) and when adjusting on age, gender, wbc, and tpi (panels (c) and (d)).

The MLC, presented in Figure 3, is the same for all four models (exponential, Cox without and with i.i.d. or CAR frailties) and both adjustments considered. Tables 1 and 2, respectively, describe the MLC and its statistical significance for the four models and the two covariate adjustments. Similarly to Frévent et al. [17], the hazard ratio in Table 2 was estimated in a conventional Cox model adjusted for the covariates.

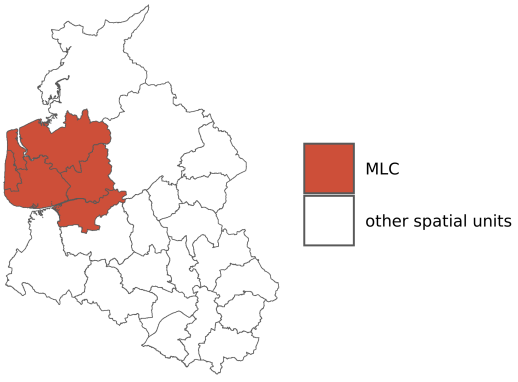

Figure 3. MLC detected by the exponential model, the Cox model without shared frailties, and the Cox model with i.i.d. and CAR shared frailties. The MLC is identical whatever the covariate adjustment or the approach considered.

Table 1. Description of the MLC detected by the exponential model, the Cox model without shared frailties, and the Cox model with i.i.d. and CAR shared frailties. The MLC is identical whatever the covariate adjustment or the approach considered.

	Inside the MLC	Outside the MLC
Number of spatial units	5	19
Number of individuals	234	809
Number of events	193	686
Average patient age (years)	65.5	59.3
Percentage of men	55.1%	51.7%
Average patient wbc	33.3	40.1
Average tpi	-0.75	0.66

Table 2. Estimated p-value and hazard ratio for the MLC detected with each model and each covariate adjustment.

	Adjustment on age, sex, wbc	Adjustment on age, sex, wbc, tpi
Exponential model	0.001	0.004
Cox model without shared frailties	0.001	0.001
Cox model with i.i.d. shared frailties	0.025	0.067
Cox model with CAR shared frailties	0.032	0.065
Hazard ratio	0.65	0.67

Although the MLC is identical whatever the method, it should be noted that when we take into account the correlation of observations (with the i.i.d. and the CAR shared frailties approaches), the MLC becomes less statistically significant. When the Townsend score is included in the adjustment, the cluster is no longer statistically significant with these two models (see Table 2).

5. Conclusions

This article presents a review of the literature on scan statistics for survival data. The first approach developed is based on an exponential model. This has the disadvantage of assuming a constant hazard rate over time, which is rather simplistic. The parametric approaches developed later avoid this problem, as does the approach based on the Cox model, which is even more flexible. However, these scan statistics assume the strong and rather unrealistic, albeit popular, hypothesis of independence of the observations. A more recent approach that does not require this assumption and takes account of the potential correlation, in the data is then presented.

Most applications of spatial scan statistics for survival data require the adjustment of cluster detection on covariates. This is therefore also detailed and illustrated on the `LeukSurv` dataset.

Several drawbacks to the current spatial scan statistics approaches can be mentioned. The estimation of the p-value associated with the MLC is carried out using Monte Carlo simulations as the distribution of the spatial scan statistic is intractable under $\mathcal{H}_0$. This leads to high computation times, which limit the practical application on large datasets. A solution is to approximate the p-value using the method

proposed by [19]. Briefly, this approach consists in estimating the p-value accurately from only a small number of Monte Carlo simulations. Further work would involve obtaining the distribution of the scan statistic under $\mathcal{H}_0$.

Moreover, in practice, it is sometimes necessary to detect secondary clusters, i.e., other clusters that are also statistically significant. Several approaches have been suggested in the literature. For example, Kulldorff [2] proposed to perform statistical inference on the other potential clusters in exactly the same way as for the MLC, while other authors suggest removing the MLC from the data [20], before repeating the scan procedure. However, these approaches do not maintain the type I error. This is a challenging subject that requires further work.

Use of Generative-AI tools declaration

The author declare that she has not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The author is grateful to the reviewers for their helpful comments, which improved the quality of the paper. The author would also like to thank Sophie Dabo-Niang and Michaël Genin, thanks to whom she developed an expertise in spatial scan statistics during her PhD.

Conflict of interest

The author declares no conflict of interest in this paper

References

1. M. Kulldorff, N. Nagarwalla, Spatial disease clusters: detection and inference, *Stat. Med.*, **14** (1995), 799–810. http://dx.doi.org/10.1002/sim.4780140809
2. M. Kulldorff, A spatial scan statistic, *Commun. Stat.-Theory Methods*, **26** (1997), 1481–1496. http://dx.doi.org/10.1080/03610929708831995
3. I. Jung, M. Kulldorff, A. C. Klassen, A spatial scan statistic for ordinal data, *Stat. Med.*, **26** (2007), 1594–1607. http://dx.doi.org/10.1002/sim.2607
4. M. Kulldorff, L. Huang, L. Pickle, L. Duczmal, An elliptic spatial scan statistic, *Stat. Med.*, **25** (2006), 3929–3943. http://dx.doi.org/10.1002/sim.2490
5. L. Cucala, C. Demattei, P. Lopes, A. Ribeiro, A spatial scan statistic for case event data based on connected components, *Comput. Stat.*, **28** (2013), 357–369. http://dx.doi.org/10.1007/s00180-012-0304-6
6. T. Tango, K. Takahashi, A flexibly shaped spatial scan statistic for detecting clusters, *Int. J. Health Geogr.*, **4** (2005), 11. http://dx.doi.org/10.1186/1476-072X-4-11
7. P. S. Lin, Y. H. Kung, M. Clayton, Spatial scan statistics for detection of multiple clusters with arbitrary shapes, *Biometrics*, **72** (2016), 1226–1234. http://dx.doi.org/10.1111/biom.12509

8. M. Kulldorff, L. Huang, K. Konty, A scan statistic for continuous data based on the normal probability model, *Int. J. Health Geogr.*, **8** (2009), 58. http://dx.doi.org/10.1186/1476-072X-8-58
9. L. Cucala, A distribution-free spatial scan statistic for marked point processes, *Spat. Stat.*, **10** (2014), 117–125. http://dx.doi.org/10.1016/j.spasta.2014.03.004
10. I. Jung, H. J. Cho, A nonparametric spatial scan statistic for continuous data, *Int. J. Health Geogr.*, **14** (2015), 30. http://dx.doi.org/10.1186/s12942-015-0024-6
11. L. Huang, M. Kulldorff, D. Gregorio, A spatial scan statistic for survival data, *Biometrics*, **63** (2007), 109–118. http://dx.doi.org/10.1111/j.1541-0420.2006.00661.x
12. V. Bhatt, N. Tiwari, A spatial scan statistic for survival data based on Weibull distribution, *Stat. Med.*, **33** (2014), 1867–1876. http://dx.doi.org/10.1002/sim.6075
13. V. Bhatt, N. Tiwari, A spatial scan statistic for survival data based on generalized life distribution, *Commun. Stat.-Theory Methods*, **45** (2016), 5730–5744. http://dx.doi.org/10.1080/03610926.2014.948207
14. I. Usman, R. J. Rosychuk, A log-Weibull spatial scan statistic for time to event data, *Int. J. Health Geogr.*, **17** (2018), 20. http://dx.doi.org/10.1186/s12942-018-0137-9
15. A. J. Cook, D. R. Gold, Y. Li, Spatial cluster detection for censored outcome data, *Biometrics*, **63** (2007), 540–549. http://dx.doi.org/10.1111/j.1541-0420.2006.00714.x
16. W. R. Tobler, A computer movie simulating urban growth in the Detroit region, *Econ. Geogr.*, **46** (1970), 234–240. http://dx.doi.org/10.2307/143141
17. C. Frévent, M. S. Ahmed, S. Dabo-Niang, M. Genin, A shared-frailty spatial scan statistic model for time-to-event data, *Biometrical J.*, **66** (2024), e202300200. http://dx.doi.org/10.1002/bimj.202300200
18. R. Henderson, S. Shimakura, D. Gorst, Modeling spatial variation in leukemia survival data, *J. Am. Stat. Assoc.*, **97** (2002), 965–972. http://dx.doi.org/10.1198/016214502388618753
19. A. M. Abrams, K. Kleinman, M. Kulldorff, Gumbel based p-value approximations for spatial scan statistics, *Int. J. Health Geogr.*, **9** (2010), 61. http://dx.doi.org/10.1186/1476-072X-9-61
20. Z. Zhang, R. Assunção, M. Kulldorff, Spatial scan statistics adjusted for multiple clusters, *J. Probab. Stat.*, 2010. http://dx.doi.org/10.1155/2010/642379

AIMS Press