mathematics

MDPI

Article

Spatial Cluster Detection Under Dependent Random Environmental Effects

Walguen Oscar [1,2] and Jean Vaillant [1,*]

[1] Laboratoire de Mathématiques et Informatique et Applications (LAMIA), Université des Antilles, 97157 Pointe-à-Pitre CEDEX, France; walguen.oscar@univ-antilles.fr
[2] Ecole Normale Supérieure, Université d'Etat d'Haiti, Port-au-Prince HT6110, Haiti
* Correspondence: jean.vaillant@univ-antilles.fr

Abstract: This paper develops a new approach for the detection of spatial clusters in the presence of random environmental effects and covariates when the observed data consist of counts over a regular grid. Such data are frequently overdispersed and spatially dependent. Overdispersion and spatial dependence must be taken into account in the modeling, otherwise the classical scan statistics method may lead to the detection of false clusters. Therefore, we consider that the observed counts are generated by a Cox process, allowing for overdispersion and spatial correlation. The environmental effects here represents unobserved covariates, as opposed to observed covariates whose observations are used via the link function in the model. These random effects are modeled by means of spatial copula with margins distributed according to a Gamma distribution. We then prove that the counts are dependent and negative binomial and propose a spatial cluster detection test based on data augmentation techniques. It is worth noting that our model also includes the case of independent effects for which the counts are independent and negative binomial. An illustration of these spatial scan techniques is provided by a Black Leaf Streak Disease (BLSD) dataset from Martinique, French West Indies. The comparison of our model with Poisson models, with or without covariates, demonstrates the importance of our approach in avoiding false clusters.

Keywords: Cox process; scan statistics; likelihood ratio tests; MCMC; copula; spatial cluster; data augmentation

MSC: 62M30; 62P12; 62E10

check for updates

Academic Editors: Raydonal Ospina, Victor Leiva and Cecília Castro

Received: 28 December 2024
Revised: 22 January 2025
Accepted: 23 January 2025
Published: 27 January 2025

Citation: Oscar, W.; Vaillant, J. Spatial Cluster Detection Under Dependent Random Environmental Effects. *Mathematics* **2025**, *13*, 430. https://doi.org/10.3390/math13030430

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The spatial scan statistic proposed by [1,2] is a relevant and widely used tool in several fields of study for the detection and analysis of spatial clusters. Many applications of this statistical procedure mainly concern count data [2], ordinal variables [3], or multinomial models [4]. These spatial scan statistic methods are of great interest when the focus is the detection of spatial clusters of individuals or locations with high or low values for certain attributes. Thus, [5,6] developed models for analyzing survival data, based on exponential and Weibull distributions, and [7] developed a scan statistic for continuous data, based on the normal probability model. The latter model was applied to identify geographical clusters of low birth weight in New York City. [8] proposed a spatial scan statistic based on a beta regression model for geographical cluster detection and continuous data in the interval $(0,1)$, such as rates, proportions, and indices, or data restricted in an interval $(a,b), a < b$. An extension of the scan statistic to multidimensional and spatio-temporal

point processes [9] was also proposed. Ref. [10] introduced a free-distribution spatial scan statistic for labeled point processes based on a concentration index other than the usual likelihood ratio test. In the case of multivariate continuous data, [11] developed a scan statistic that takes into account correlations between variables. The statistical significance of clusters is obtained using Monte Carlo permutations. In order to optimize the efficiency of these methods, some authors have taken several features into account. For example, the authors in [12] proposed a modified scan statistic to examine the behavior of the spatial scan statistic in the presence of overdispersion and spatial correlation. From a theoretical point of view, they showed that when overdispersion or spatial correlation are not taken into account, the classic Kulldorff spatial scan statistic tends to produce more false positive clusters with respect to the nominal significance level, which [13] called false alarms. They confirmed this theory with simulation studies showing that the classical Kulldorff spatial scan statistic identifies too many clusters in the presence of overdispersion or spatial correlation. An approach is presented in [14] to explicitly handle spatial contiguity in cluster detection without using arbitrarily shaped scan windows.

The main objective of this paper is to develop a method for detecting spatial clusters in the presence of unobserved effects and observed covariates from counts observed on a regular grid associated with a spatial Cox process, taking into account overdispersion and spatial correlation. The class of Cox processes (see [15,16] for time processes and [17] for spatio-temporal ones) can be viewed as two-step randomization modeling. A Cox process, also known as a doubly stochastic Poisson process, is in fact a Poisson point process in which the intensity function (describing the expected number of points per unit of time or space) is itself a stochastic process. In a wide range of scientific applications, such as seismology, epidemiology, or ecology, it provides interesting models for describing event distributions (e.g., earthquakes, disease outbreaks, or pest infestations). For example, the log-Gaussian Cox processes have been described and studied by several authors (such as [18,19]). Note also that [20] used, as a working model, a spatio-temporal Cox process for describing the distribution of rat sightings in an urban environment. It should be noted that, for many authors (such as [20]), the expression *environmental variables* refers to known and observed variables. For our purposes, we distinguish between environmental effects that are unknown or unobserved and covariates that are observed and included in the model prior to significance testing. Moreover, the spatial correlation of effects is taken into account in our model, contrary to what is generally proposed in the literature on spatial counting processes ([21], chapter 8). On the other hand, [22] developed a model based on a Cox process associated with a spatial copula [23], modeling the correlation between effects. This model takes into account overdispersion and spatial correlation. For basic information on spatial copulas, we refer readers to [24].

In this paper, we focus on the construction and implementation of statistical tests for cluster detection with respect to a spatial Cox process associated with a spatial copula and unobserved correlated environmental effects. Both cases of independent and dependent effects are analyzed. The model developed and its distributional properties are presented in Section 2. Firstly, the case of independent effects is studied, and then the dependent case is introduced, with a particular focus on joint Gamma distributed effects. The proposed scan test procedures are introduced in Section 3, and an illustration of these techniques is provided in Section 4 from a dataset regarding the Black Leaf Streak Disease (BLSD), collected in Martinique island, French West Indies.

2. The Model and Its Distributional Properties

Let $\mathcal{H}$ be a metric space associated with its Borelian σ-algebra $\mathcal{B}$. We assume that the region of interest $\mathcal{H}$ is partitioned into n subsets, $\mathcal{H}_j, j$ in $J = \{1, \ldots, n\}$. A Cox process on

the measured space $(\mathcal{H}, \mathcal{B}, \nu)$ is considered, and we denote by N its associated counting measure. We write $Y_j = N(\mathcal{H}_j)$, which stands for the number of occurrences in $\mathcal{H}_j$. The model is described as follows: for any j in J, the count Y_j follows a conditional Poisson distribution, $\mathcal{P}\big(A_j S_{\nu,j}(\theta, Z)\big)$:

$$\forall j \in J, \quad Y_j \mid A_j, Z \sim \mathcal{P}\big(A_j S_{\nu,j}(\theta, Z)\big) \tag{1}$$

with $S_{\nu,j}(\theta, Z) = \int_{\mathcal{H}_j} f_\theta\big(Z(x)\big)\nu(dx)$, where

- ν denotes a reference measure on $(\mathcal{H}, \mathcal{B})$ (in practice, either the Lebesgue measure or the counting measure),
- A_j, for j in J, is a positive random effect linked to $\mathcal{H}_j$,
- $Z = (Z(x))_{x\in\mathcal{H}}$ is a random field taking values in a q-dimensional space where $Z(x)$ denotes the random vector of q covariates at point x,
- $f_\theta(.)$ is a link function from $\mathbb{R}^q$ onto $\mathbb{R}_+$ parameterized by the unknown vector θ, with $\theta \in \mathbb{R}^{q'}$ and $q' \geq q$,
- $S_{\nu,j}(.)$ is the integral of $f_\theta(Z)$ over $\mathcal{H}_j$ with respect to ν.

For example, if $q = q'$ and ν is the counting measure and $f_\theta(Z(x)) = e^{<\theta, Z(x)>}$, then $S_{\nu,j}$ is the sum $\sum_{x\in\mathcal{H}_j} e^{<\theta,Z(x)>}$.

Since (1) is a Poisson mixture model, overdispersion exists and the variance-to-expectation ratio of Y_j is greater than 1:

$$\frac{V(Y_j|Z)}{E(Y_j|Z)} = 1 + S_{\nu,j}(\theta, Z)\frac{V(A_j|Z)}{E(A_j|Z)} > 1.$$

The following Proposition 1 gives us the likelihood function of $(y_j)_{j\in J}$ conditional on Z and $(a_j)_{j\in J}$, with respect to (θ, β), where β is the joint density parameter of effects.

Proposition 1. *For the model defined by Equation (1), if the joint distribution of $(A_j)_{j\in J}$ admits a probability density function g parametrized by β, then the likelihood of $(y_j)_{j\in J}$ conditional on Z and $(a_j)_{j\in J}$ with respect to (θ, β) is equal to*

$$\mathcal{L}\Big((\theta, \beta); (y_j)_{j\in J}\Big) = \left(\prod_{j=1}^{n} \frac{(a_j S_{\nu,j}(\theta, Z))^{y_j}}{y_j!} \exp(-a_j S_{\nu,j}(\theta, Z))\right) g\big((a_j)_{j\in J}\big). \tag{2}$$

Proof. Conditional on A_j and Z, Y_j are independent and have a Poisson distribution with parameter $A_j S_{\nu,j}(\theta, Z)$. Therefore, the probability

$$\mathbb{P}\left(\bigcap_{j=1}^{n}(Y_j = y_j)|(a_j)_{j\in J}, Z\right) = \prod_{j=1}^{n} \frac{(a_j S_{\nu,j}(\theta, Z))^{y_j}}{y_j!} \exp(-a_j S_{\nu,j}(\theta, Z))$$

The final result is obtained by multiplying this probability by the joint density of $(a_j)_{j\in J}$. □

In the sequel, two situations are considered: the effects are either independent or dependent. Most often, practitioners assume the independence of the effects, as the model is relatively easy to implement. This is no longer true in case-dependent effects.

In paragraph 2.1 (respectively paragraph 2.2), we present distributional results in the case of independent (respectively dependent) random effects. We focus on effects

distributed according to a Gamma distribution, as this is the most widely used distribution for positive random effects in the literature [25,26], along with the lognormal distribution.

2.1. Case of Independent Effects

Proposition 2 provides the joint distribution of $(Y_j)_{j\in J}$, in the case where the effects A_j are independent and distributed according, $\Gamma(m_j, \gamma_j)$, which stands for the Gamma distribution with two parameters: $m_j > 0$ (location parameter) and $\gamma_j > 0$ (shape parameter).

Proposition 2. *Consider the model defined by Equation (1). If we assume that A_j are independent and distributed according to a Gamma distribution, $\Gamma(m_j, \gamma_j)$, of density function f_{A_j} given by*

$$f_{A_j}(a_j) = \frac{a_j^{\gamma_j - 1}}{\Gamma(\gamma_j)} \left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} \exp\left(-a_j \frac{\gamma_j}{m_j}\right) \tag{3}$$

then the Y_j are distributed according to a Negative Binomial, such that, for all $j \in J$, we have:

$$\mathbb{P}(Y_j = y_j|Z) = \frac{\Gamma(y_j + \gamma_j)}{y_j!\Gamma(\gamma_j)} \left(\frac{\gamma_j}{m_j S_{v,j}(\theta, Z) + \gamma_j}\right)^{\gamma_j} \left(\frac{m_j S_{v,j}(\theta, Z)}{m_j S_{v,j}(\theta, Z) + \gamma_j}\right)^{y_j} \tag{4}$$

Proof. The probability of the event $Y_j = y_j$ is:

$$\mathbb{P}(Y_j = y_j|Z) = \int_0^{+\infty} \mathbb{P}(Y_j = y_j|A_j, Z) f_{A_j}(a_j; m_j, \gamma_j) da_j$$

where $Y_j|A_j, Z \sim \mathcal{P}\big(A_j S_{v,j}(\theta, Z)\big)$ and $f_{A_j}(a_j; m_j, \gamma_j)$ is the density of the gamma distribution with parameters m_j and γ_j. We obtain:

$$\begin{aligned}
\mathbb{P}(Y_j = y_j|Z) &= \int_0^{+\infty} \exp(-a_j S_{v,j}(\theta, Z)) \frac{(a_j S_{v,j}(\theta, Z))^{y_j}}{y_j!} \times \frac{a_j^{\gamma_j - 1}}{\Gamma(\gamma_j)} \exp\left(-a_j \frac{\gamma_j}{m_j}\right) \left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} da_j \\
&= \left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} \frac{(S_{v,j}(\theta, Z))^{y_j}}{\Gamma(\gamma_j) y_j!} \int_0^{+\infty} a_j^{y_j + \gamma_j - 1} \exp\left(-a_j \left(S_{v,j}(\theta, Z) + \frac{\gamma_j}{m_j}\right)\right) da_j.
\end{aligned}$$

Writing $s = a_j \left(S_{v,j}(\theta, Z) + \frac{\gamma_j}{m_j}\right)$, then $a_j = \frac{m_j s}{m_j S_{v,j}(\theta, Z) + \gamma_j}$ and $da_j = \frac{m_j}{m_j S_{v,j}(\theta, Z) + \gamma_j} ds$.

Then,

$$\begin{aligned}
\mathbb{P}(Y_j = y_j|Z) &= \left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} \frac{(S_{v,j}(\theta, Z))^{y_j}}{\Gamma(\gamma_j) y_j!} \int_0^{+\infty} \left(\frac{m_j s}{m_j S_{v,j}(\theta, Z) + \gamma_j}\right)^{y_j + \gamma_j - 1} e^{-s} \frac{m_j}{m_j S_{v,j}(\theta, Z) + \gamma_j} ds \\
&= \left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} \frac{(S_{v,j}(\theta, Z))^{y_j}}{\Gamma(\gamma_j) y_j!} \left(\frac{m_j}{m_j S_{v,j}(\theta, Z) + \gamma_j}\right)^{y_j + \gamma_j} \int_0^{+\infty} s^{y_j + \gamma_j - 1} e^{-s} ds \\
&= \left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} \frac{(S_{v,j}(\theta, Z))^{y_j}}{\Gamma(\gamma_j) y_j!} \left(\frac{m_j}{m_j S_{v,j}(\theta, Z) + \gamma_j}\right)^{y_j + \gamma_j} \Gamma(y_j + \gamma_j) \\
&= \frac{\Gamma(y_j + \gamma_j)}{y_j!\Gamma(\gamma_j)} \left(\frac{\gamma_j}{m_j S_{v,j}(\theta, Z) + \gamma_j}\right)^{\gamma_j} \left(\frac{m_j S_{v,j}(\theta, Z)}{m_j S_{v,j}(\theta, Z) + \gamma_j}\right)^{y_j}
\end{aligned}$$

which leads to the final result. □

Consequently, with independent and gamma-distributed random effects, the counts follow a negative binomial distribution with parameters $m_j S_{v,j}(\theta, Z)$ and γ_j, such that

$$E(Y_j|Z) = m_j S_{v,j}(\theta, Z) \quad \text{and} \quad V(Y_j|Z) = E(Y_j|Z)\left(1 + \frac{m_j S_{v,j}(\theta, Z)}{\gamma_j}\right)$$

Remark 1.

1. $\dfrac{V(Y_j|Z)}{E(Y_j|Z)} = 1 + \dfrac{m_j S_v(\theta, Z)}{\gamma_j}$ *is an overdispersion index ([17,27,28]). When* $\gamma_j \to \infty$*, this index tends to 1 and we get the Poisson model conditional on Z.*
2. *Our statistical tests for cluster detection will be based on the expected values,* m_j*.*
3. *For a fixed* γ_j*, an increase in* m_j *leads to an increase in the overdispersion index, and for a fixed* m_j*, an increase in* γ_j *leads to a decrease in the overdispersion index.*
4. *Note that Chang and Rosychuk [29] considered a compound Poisson representation of the negative binomial distribution.*

2.2. Case of Dependent Effects

Now, we consider that the random effects, $(A_j)_{j\in J}$, are dependent and distributed according to a multivariate distribution whose margins are univariate gamma distributions. The dependence structure is modeled by a spatial copula (see [24] for elements on the general copula theory and spatial copula modeling). The following Proposition 3 provides the likelihood conditional on Z, for $(y_j)_{j\in J}$:

Proposition 3. *For the model defined by Equation (1), if* $(A_j)_{j\in J}$ *follows a multivariate joint density distribution* g *associated with a spatial copula density* c *with spatial correlation parameter* ρ *and margins distributed according to the gamma distribution* $\Gamma(m_j, \gamma_j)$*, then the likelihood conditional on* Z*, for* $(y_j)_{j\in J}$ *with respect to* θ *and* $\beta = \left((m_j, \gamma_j)_{j\in J}, \rho\right)$*, is*

$$\mathcal{L}\left((\theta, \beta); (y_j)_{j\in J}\right) =$$

$$\int_{\mathbb{R}_+^n} \left(\prod_{j=1}^{n} \frac{(a_j S_{v,j}(\theta, Z))^{y_j}}{y_j!} \exp(-a_j S_{v,j}(\theta, Z))\right) c\left(\left(F_{A_j}(a_j)\right)_{j\in J}\right) \prod_{j=1}^{n} f_{A_j}(a_j) da_j \tag{5}$$

where F_{A_j} *and* f_{A_j} *are the distribution function and density of* A_j*, respectively.*

Proof. By Sklar's theorem ([30]), we deduce that the joint density g of the $(A_j)_{j\in J}$ is the product of the marginal densities multiplied by the density of the copula c. Consequently, the joint density of the random effects A_j is given by

$$g\left((a_j)_{j\in J}\right) = c\left(\left(F_{A_j}(a_j)\right)_{j\in J}\right) \prod_{j=1}^{n} f_{A_j}(a_j). \tag{6}$$

Proposition 1 leads to the final result by integrating over $\mathbb{R}_+^n$ with respect to $(a_j)_{j\in J}$. □

To model the spatial dependence between random effects, we used a multivariate spatial Gaussian copula, where the correlation between two effects, A_j and $A_{j'}$, from two regions, j and j', denoted $cor(A_j, A_{j'})$, depends on the distance $d(C_j, C_{j'})$ between the centroids C_j and $C_{j'}$ of the two regions. The density of this copula is defined, for all $u = (u_1, \ldots, u_n) \in [0, 1]^n$, by:

$$c(u) = |W|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(u^*)^t (W^{-1} - I_n) u^*\right) \tag{7}$$

with W being the correlation matrix defined as follows:

$$W = \left(\rho^{d(C_j, C_{j'})}\right)_{(j,j') \in J^2} \tag{8}$$

where

- $u^* = (\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_n))$, with $u_j = F_{A_j}(a_j), j \in \{1, \ldots, n\}$, and Φ^{-1} is the inverse of the standard normal probability distribution function,
- $\rho \in [0, 1]$ is the spatial correlation parameter and $|W|$ is the determinant of W.

We define the matrix W in such a way that the level of partial dependence between the effects of two subregions is a function of the distance between them. Other choices of correlation functions are possible, particularly in the case of isotropy [31].

When W is the identity matrix $\mathbb{I}_n$, we have $C(u_1, \ldots, u_n) = \prod_{j=1}^{n} u_j, \forall u_j \in [0, 1]^n$, i.e., the n-dimensional independence copula of constant density and equal to 1, so the random effects A_j are independent and the Proposition 2 is verified.

3. Statistical Test Construction

In this section, we present statistical tests we have developed for the detection of spatial clusters in which predicted rates within the cluster are significantly different from predicted rates elsewhere, while there is overdispersion and spatial correlation between effects. The null hypothesis denoted by H_0 is defined by:

$$H_0 : \forall j \in \{1, \ldots, n\}, \quad m_j = m, \quad \text{and} \quad \gamma_j = \gamma. \tag{9}$$

For a candidate cluster $C \subset \{1, \ldots, n\}$, the alternative hypothesis denoted by $H_1(C)$ is defined by:

$$H_1(C) : \quad (m_j, \gamma_j) = \begin{cases} (m_C, \gamma_C) \neq (m, \gamma) & \text{if} \quad j \in C \\ (m, \gamma) & \text{if} \quad j \notin C. \end{cases} \tag{10}$$

3.1. Cases of Independent Effects

According to Proposition 2, under H_0, the likelihood function conditional on Z is:

$$\mathcal{L}_{H_0}(m, \gamma, \theta; (y_j)_{j \in J}) = \prod_{j=1}^{n} \frac{\Gamma(y_j + \gamma)}{y_j! \Gamma(\gamma)} \left(\frac{\gamma}{mS_{v,j}(\theta, Z) + \gamma}\right)^{\gamma} \left(\frac{mS_{v,j}(\theta, Z)}{mS_{v,j}(\theta, Z) + \gamma}\right)^{y_j}. \tag{11}$$

The likelihood function under $H_1(C)$ denoted by $\mathcal{L}_{H_1(C)}(m, m_C, \gamma, \gamma_C, \theta; y_j)$ is:

$$\mathcal{L}_{H_1(C)}(m, m_C, \gamma, \gamma_C, \theta; (y_j)_{j \in J}) = \prod_{j=1}^{n} \mathbb{P}(Y_j = y_j | Z) = \prod_{j \in C} \mathbb{P}(Y_j = y_j | Z) \times \prod_{j \notin C} \mathbb{P}(Y_j = y_j | Z)$$

$$= \prod_{j \in C} \frac{\Gamma(y_j + \gamma_C)}{y_j! \Gamma(\gamma_C)} \left(\frac{\gamma_C}{m_C S_{v,j}(\theta, Z) + \gamma_C}\right)^{\gamma_C} \left(\frac{m_C S_{v,j}(\theta, Z)}{m_C S_{v,j}(\theta, Z) + \gamma_C)}\right)^{y_j}$$
$$\times \prod_{j \notin C} \left(\frac{\Gamma(y_j + \gamma)}{y_j! \Gamma(\gamma)} \left(\frac{\gamma}{mS_{v,j}(\theta, Z) + \gamma}\right)^{\gamma} \left(\frac{mS_{v,j}(\theta, Z)}{mS_{v,j}(\theta, Z) + \gamma}\right)^{y_j}\right) \tag{12}$$

Therefore, the likelihood ratio regarding $H_1(C)$ and H_0 is:

$$\frac{\mathcal{L}_{H_1(C)}(m, m_C, \gamma, \gamma_C, \theta; (y_j)_{j\in J})}{\mathcal{L}_{H_0}(m, \gamma, \theta; (y_j)_{j\in J})} = \frac{\prod_{j\in C} \frac{\Gamma(y_j+\gamma_C)}{y_j!\Gamma(\gamma_C)} \left(\frac{\gamma_C}{m_C S_{\nu,j}(\theta,Z)+\gamma_C}\right)^{\gamma_C} \left(\frac{m_C S_{\nu,j}(\theta,Z)}{m_C S_{\nu,j}(\theta,Z)+\gamma_C}\right)^{y_j}}{\prod_{j\in C} \frac{\Gamma(y_j+\gamma)}{y_j!\Gamma(\gamma)} \left(\frac{\gamma}{mS_{\nu,j}(\theta,Z)+\gamma}\right)^{\gamma} \left(\frac{mS_{\nu,j}(\theta,Z)}{mS_{\nu,j}(\theta,Z)+\gamma}\right)^{y_j}}$$

where $\frac{\Gamma(y_j+\gamma_C)}{\Gamma(\gamma_C)} = \prod_{k=0}^{y_j-1}(\gamma_C+k)$.

Consequently, the log likelihood ratio is:

$$\sum_{j\in C}\left(\sum_{k=0}^{y_j-1}\log\left(\frac{\gamma_C+k}{\gamma+k}\right) - \log\left(\frac{(m_C S_{\nu,j}(\theta,Z)+\gamma_C)^{\gamma_C}}{(mS_{\nu,j}(\theta,Z)+\gamma)^{\gamma}}\right)\right) + n_C \log\left(\frac{\gamma_C^{\gamma_C}}{\gamma^{\gamma}}\right) + \log\left(\frac{m_C}{m}\right)\sum_{j\in C} y_j + \sum_{j\in C} y_j \log\left(\frac{mS_{\nu,j}(\theta,Z)+\gamma}{m_C S_{\nu,j}(\theta,Z)+\gamma_C}\right) \tag{13}$$

where n_C stands for the number of $\mathcal{H}_j$ in cluster C.

From expression (13), it results that the likelihood ratio statistic for testing the null hypothesis H_0 against the alternative hypothesis $H_1(C)$ is:

$$\widehat{LR}_C = \sum_{j\in C}\left(\sum_{k=0}^{y_j-1}\log\left(\frac{\hat\gamma_{C,1}+k}{\hat\gamma_0+k}\right) - \log\left(\frac{(\hat m_{C,1} S_{\nu,j}(\hat\theta_1,Z)+\hat\gamma_{C,1})^{\hat\gamma_{C,1}}}{(\hat m_0 S_{\nu,j}(\hat\theta_0,Z)+\hat\gamma_0)^{\hat\gamma_0}}\right)\right) + n_C \log\left(\frac{\hat\gamma_{C,1}^{\hat\gamma_{C,1}}}{\hat\gamma_0^{\hat\gamma_0}}\right) + \log\left(\frac{\hat m_{C,1}}{\hat m_0}\right)\sum_{j\in C} y_j + \sum_{j\in C} y_j \log\left(\frac{\hat m_0 S_{\nu,j}(\hat\theta_0,Z)+\hat\gamma_0}{\hat m_{C,1} S_{\nu,j}(\hat\theta_1,Z)+\hat\gamma_{C,1}}\right) \tag{14}$$

where $\hat m_1$, $\hat m_{C,1}$, $\hat\gamma_1$, $\hat\gamma_{C,1}$, and $\hat\theta_1$ are the conditional maximum likelihood estimates under $H_1(C)$ of the parameters m, m_C, γ, γ_C, and θ for a given cluster C; then, $\hat m_0$, $\hat\gamma_0$, and $\hat\theta_0$ are the estimates of m, γ, and θ under H_0.

The spatial scan statistic LR^* associated with the independent gamma-distributed effect Cox process is defined as a maximized index of the likelihood ratio statistic over all potential clusters:

$$LR^* = \arg\max_{C\in\mathcal{C}}(\widehat{LR}_C) \tag{15}$$

where $\mathcal{C}$ is the set of all candidate clusters.

3.2. Cases of Dependent Effects

We now assume that the A_j in (1) are dependent and distributed according to a multivariate distribution of density, g, associated with a Gaussian copula and distributed gamma margins.

The conditional likelihood function of Equation (5) is then written as

$\mathcal{L}\left((\theta,\beta);(y_j)_{j\in J}\right) =$

$$\int_{\mathbb{R}_+^n}\left(\prod_{j=1}^n \frac{(a_j S_{\nu,j}(\theta,Z))^{y_j}}{y_j!}\exp(-a_j S_{\nu,j}(\theta,Z))\right) c\left(\left(F_{A_j}(a_j)\right)_{j\in J}\right)\prod_{j=1}^n \frac{a_j^{\gamma_j-1}}{\Gamma(\gamma_j)}\left(\frac{\gamma_j}{m_j}\right)^{\gamma_j}\exp\left(-a_j\frac{\gamma_j}{m_j}\right)da_j$$

$$= \int_{\mathbb{R}_+^n}\left(\prod_{j=1}^n \frac{(S_{\nu,j}(\theta,Z))^{y_j}}{\Gamma(\gamma_j)y_j!}\left(\frac{\gamma_j}{m_j}\right)^{\gamma_j}\right)\prod_{j=1}^n a_j^{y_j+\gamma_j-1}\exp\left(-a_j\left(S_{\nu,j}(\theta,Z)+\frac{\gamma_j}{m_j}\right)\right)c\left(\left(F_{A_j}(a_j)\right)_{j\in J}\right)da_j$$

$$= \prod_{j=1}^{n} \frac{(S_{\nu,j}(\theta,Z))^{y_j}}{\Gamma(\gamma_j)y_j!}\left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} \int_{\mathbb{R}_+^n} \prod_{j=1}^{n} a_j^{y_j+\gamma_j-1} \exp\left(-a_j\left(S_{\nu,j}(\theta,Z)+\frac{\gamma_j}{m_j}\right)\right) c\Big(\big(F_{A_j}(a_j)\big)_{j\in J}\Big) da_j$$

where

$$c(\mathbf{u}) = |W|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(u^*)^t(W^{-1}-I_n)u^*\right) \tag{16}$$

with $u^* = (\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_n))$ and $\mathbf{u} = (u_j)_{j\in J}$, where $u_j = F_{A_j}(a_j)$ and W given by Equation (8).

In expression (16), $c(u)$ depends on m_j, γ_j, and ρ, which are omitted for commodity and easy reading.

Due to the complexity of the above conditional likelihood function, it is preferable to adopt a Bayesian approach and use a Monte Carlo algorithm to draw conclusions from our assumptions. The augmented likelihood with respect to $(a_j)_{j\in J}$ is:

$$\mathcal{L}\Big((\theta,\beta),(a_j)_{j\in J};(y_j)_{j\in J}\Big) =$$

$$\left(\prod_{j=1}^{n} \frac{(a_j S_{\nu,j}(\theta,Z))^{y_j}}{y_j!}\exp(-a_j S_{\nu,j}(\theta,Z))\right) c\Big(\big(F_{A_j}(a_j)\big)_{j\in J}\Big) \prod_{j=1}^{n} \frac{a_j^{\gamma_j-1}}{\Gamma(\gamma_j)}\left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} \exp\left(-a_j\frac{\gamma_j}{m_j}\right)$$

$$= c\Big(\big(F_{A_j}(a_j)\big)_{j\in J}\Big) \prod_{j=1}^{n} \frac{(S_{\nu,j}(\theta,Z))^{y_j}}{\Gamma(\gamma_j)y_j!}\left(\frac{\gamma_j}{m_j}\right)^{\gamma_j} a_j^{y_j+\gamma_j-1} \exp\left(-a_j\left(S_{\nu,j}(\theta,Z)+\frac{\gamma_j}{m_j}\right)\right). \tag{17}$$

Thus, Equation (17) provides the augmented likelihood under both H_0 and $H_1(C)$. In particular, we have, under H_0,

$$\mathcal{L}_{H_0}\Big((\theta,m,\gamma,\rho),(a_j)_{j\in J};(y_j)_{j\in J}\Big) =$$

$$c_0\Big(\big(F_{A_j}(a_j)\big)_{j\in J}\Big)\left(\frac{\gamma}{m}\right)^{n\gamma}\Gamma(\gamma)^{-n}\prod_{j=1}^{n}\frac{(S_{\nu,j}(\theta,Z))^{y_j}}{y_j!}a_j^{y_j+\gamma-1}\exp\Big(-a_j\Big(S_{\nu,j}(\theta,Z)+\frac{\gamma}{m}\Big)\Big)$$

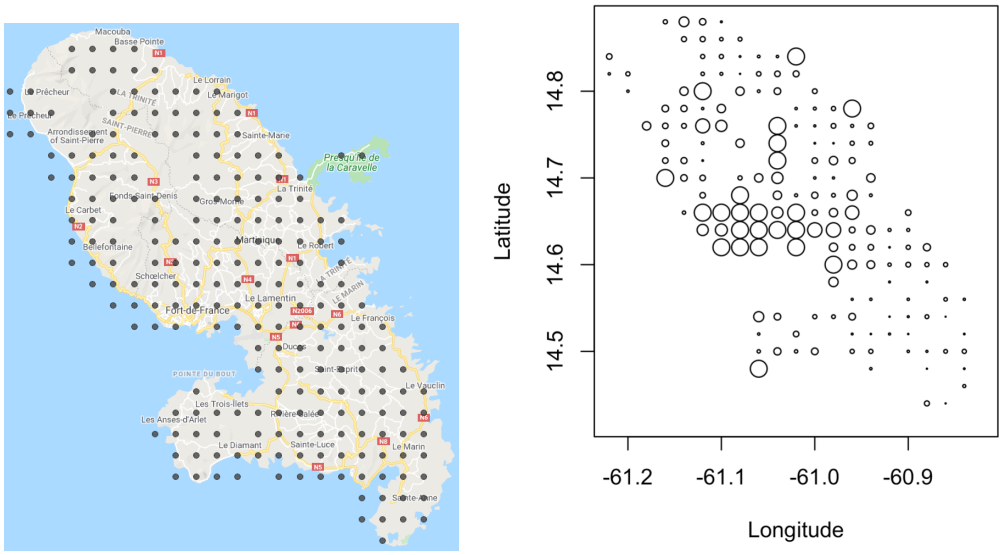
where c_0 stands for the copula density under H_0 and is given by Equations (8) and (16). Therefore, c_0 is a function of m, γ, and ρ. The spatial scan statistic associated with the dependent gamma-distributed effect Cox process is a maximized index of the posterior likelihood ratio statistic over all potential clusters:

$$LR^* = \arg\max_{C\in\mathcal{C}}\left(\frac{\mathcal{L}_{H_1(C)}(\widetilde{m}_1,\widetilde{m}_{C,1},\widetilde{\gamma}_1,\widetilde{\gamma}_{C,1},\widetilde{\theta}_1;(y_j)_{j\in J})}{\mathcal{L}_{H_0}(\widetilde{m}_0,\widetilde{\gamma}_0,\widetilde{\theta}_0;(y_j)_{j\in J})}\right) \tag{18}$$

where $\mathcal{C}$ is the set of all candidate clusters and $\widetilde{m}_1$, $\widetilde{m}_{C,1}$, $\widetilde{\gamma}_1$, $\widetilde{\gamma}_{C,1}$, and $\widetilde{\theta}_1$ are the maximum posterior likelihood estimates under $H_1(C)$ of the parameters m, m_C, γ, γ_C, and θ for a given cluster C, and $\widetilde{m}_0$, $\widetilde{\gamma}_0$, and $\widetilde{\theta}_0$ are the estimates of m, γ, and θ under H_0. The general principle of data augmentation algorithms for estimating parameters by maximum likelihood is given in [32]. Bayesian inference approaches for such models are proposed in [22] to estimate the effect and the covariate link function parameters.

4. Application to Black Leaf Streak Disease Dataset

We considered a dataset from Martinique (French West Indies) consisting of spatial counts of the Black Leaf Streak Disease (BLSD), a major foliar disease of bananas. Using a regular spatial counting grid (Figure 1a), preliminary results were obtained concerning the proportion of plants with a positive diagnosis for this disease (Figure 1b).



(**a**) (**b**)

Figure 1. (**a**) The dots on the map of Martinique indicate the lower left corners of the cells in the observation grid; (**b**) the size of the circle corresponds to the proportion of positive plants in each cell.

In addition to count data for plants diagnosed with BLSD, the following variables were available from Météo France: humidity, number of rainy days, maximum temperature, rainfall. The statistical inference about the model parameters was provided by Markov Chain Monte Carlo (MCMC) methods, as in [33,34] or [35]. These techniques consist of generating samples of the posterior distribution of $(\beta, \theta, (a_j)_{j\in J})$ and then calculating any statistics concerning parameters and unobserved effects. This posterior distribution is proportional to:

$$\mathcal{L}\Big((\theta,\beta),(a_j)_{j\in J};(y_j)_{j\in J}\Big) \times \pi(\theta)\pi(\beta)\pi((a_j)_{j\in J}|\beta) \tag{19}$$

where $\pi(\theta)$ and $\pi(\beta)$ are the prior distributions for θ and β, the joint distribution of $(a_j)_{j\in J}$ being denoted by $\pi((a_j)_{j\in J}|\beta)$.

The proposed hybrid Gibbs–Metropolis–Hasting algorithm is based on the acyclic directed graph in Figure 2, which provides posterior distribution samples for parameters and unobserved effects. From these samples, we could estimate and test the nullity of the link function coefficients and then select statistically significant covariates. The obtained results (Table 1) show that the covariates Humidity (p-value = 0.046) and Maximum Temperature (p-value = 0.032) are significantly correlated with the number of BLSD occurrences. On the other hand, the posterior distribution results for the spatial copula parameter ρ given in Table 2 indicate that ρ is significantly different from zero. It is therefore important to include Humidity and Maximum Temperature in the modeling, as well as spatial dependence, for better detection of spatial clusters.

Table 1. Estimation and test of nullity of the link function coefficients.

Covariate	Coefficient Estimate	Standard Error	z-Value	p-Value
Constant	50.21	19.57	2.57	0.010 *
Humidity	−21.86	10.95	−2.00	0.046 *
Number of raining days	−0.07	0.084	−0.84	0.400
Max temperature	−1.21	0.56	−2.15	0.032 *
Precipitation	0.0168	0.0143	1.17	0.241

* stands for significant covariate.

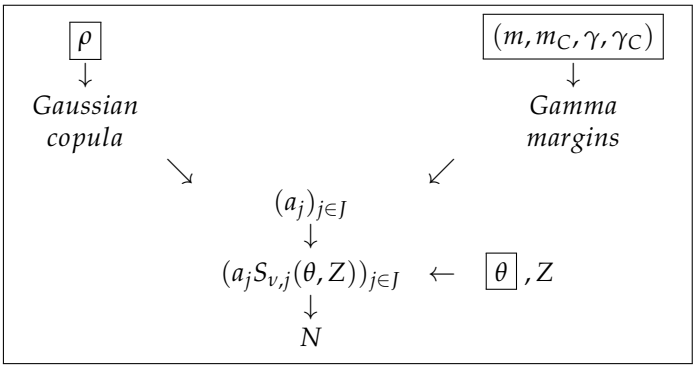


Figure 2. Acyclic directed graph of the hierarchical model defined by Proposition 3 under $H_1(C)$.

Table 2. MCMC posterior distribution results for the spatial copula parameter ρ.

Posterior Mean	Posterior Standard Deviation	95% Posterior Confidence Interval
0.486	0.28	[0.03, 0.96]

To assess how well the proposed model fits the BLSD data, we performed a residual analysis of spatial point processes [36]. Figure 3 shows Q–Q plots for our proposed model and for the Poisson model fitted to these data. Figure 3b suggests good agreement between the proposed model and the BLSD data. Figure 3a shows clear disagreement between the Poisson model and the data. In addition, the Akaike Information Criterion (AIC) is smaller for the proposed model, which supports the results provided by the Q–Q plots. Thus, the proposed model, which includes spatial correlation, overdispersion, and significant covariates, compares much more favorably with the Poisson models. It avoids falsely detecting clusters that would only be statistically significant if there were no overdispersion or spatial dependency, as shown in Figure 4.

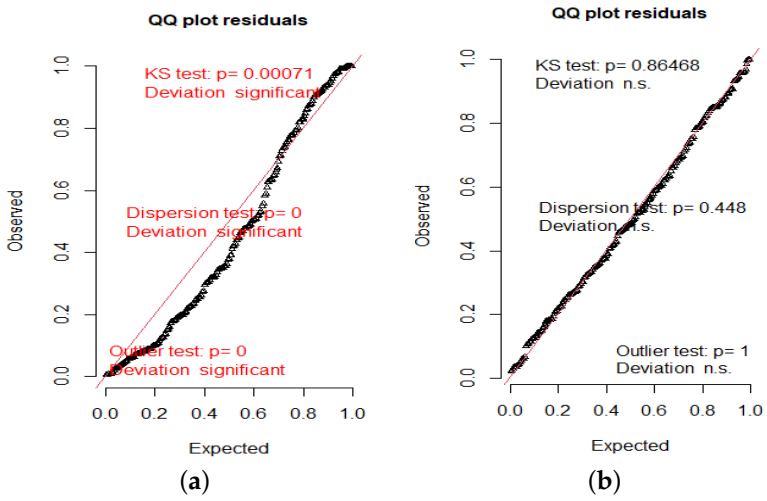


Figure 3. Goodness of fit for (**a**) the Poisson model (AIC = 895), (**b**) the proposed model (AIC = 823).



Figure 4. Statistically significant clusters for the (**a**) Poisson model, (**b**) proposed model. Red dots indicate spatial cluster locations.

5. Conclusions

Maximizing the likelihood ratio over all possible areas of the study zone is the fundamental goal of scan-based procedures. In case of overdispersion and spatial correlation, it is important that the detection of spatial clusters remains reliable, i.e., that no erroneous clusters are reported. The various approaches proposed in the literature for the detection of spatial clusters sometimes take into account observed covariates or observed spatial effects, but they do not address the case where covariates are observed and at the same time spatial effects are unobserved. In many scientific fields, however, it is well known that there are unobserved underlying factors influencing the observed phenomenon. For example, this is often the case in ecology, where numerous constraints make it impossible to obtain data on these environmental factors. Consequently, a model taking this type of situation into account was required to assess the ability of spatial scan procedures to correctly detect clusters, compared with other working models. In this paper, we considered a Cox process with covariates and spatial effects that takes into consideration overdispersion and spatial correlation. The joint distribution of effects are modeled by means of a spatial copula. When the effects are gamma-distributed, the associated spatial counts are correlated and negative binomial. Therefore, a spatial scan method based on data augmentation techniques is developed in the case of overdispersion and spatial correlation, with the aim of avoiding false alarms in cluster detection, i.e., the erroneous detection of spatial clusters as significant. In fact, some studies have shown that taking overdispersion into account can substantially reduce the risk of Type I error. In order to verify this presumed quality of the method we have developed, a BLSD dataset was studied. Our data analysis shows that our model is a good fit to the BLSD data. When applying our scan procedure to this dataset, fewer spatial clusters were detected compared to the scan procedure based on the Poisson models. This clearly indicates that by taking into account overdispersion and spatial correlation we could avoid false clusters. This study shows that the model on which the scan statistic is based is of great importance, since the quality of spatial cluster detection depends on it. It may be considered to use other spatial copulas, link functions, and choice of covariates, depending on the type of phenomenon studied. The robustness, with respect to the working model, of the test procedure used to locate the most likely significant spatial clusters and the presence of missing data are areas for further research.

Author Contributions: W.O. and J.V. contributed equally to this work in terms of conceptualization, methodology, software, and original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from FREDON Martinique.

Acknowledgments: We would like to thank DAAF and FREDON Martinique for providing the BLSD spatial count data, and Météo France for providing Martinique's rainfall data.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike Information Criterion
BLSD	Black Leaf Streak Disease
DAAF	Direction de l'Agriculture, de l'Agro-alimentaire et de la Forêt
FREDON	Fédération Régionale de Défense contre les Organismes Nuisibles
MCMC	Markov Chain Monte Carlo
Q–Q	Quantile–Quantile

References

1. Kulldorff, M.; Nagarwalla, N. Spatial disease clusters: Detection and inference. *Stat. Med.* **1995**, *14*, 799–810. [CrossRef] [PubMed]
2. Kulldorff, M. A spatial scan statistic. *Commun.-Stat.-Theory Methods* **1997**, *26*, 1481–1496. [CrossRef]
3. Jung, I.; Kulldorff, M.; Klassen, A.C. A spatial scan statistic for ordinal data. *Stat. Med.* **2007**, *26*, 1594–1607. [CrossRef] [PubMed]
4. Jung, I.; Kulldorff, M.; Richard, O.J. A spatial scan statistic for multinomial data. *Stat. Med.* **2010**, *29*, 1910–1918. [CrossRef]
5. Huang, L.; Kulldorff, M.; Gregorio, D. A spatial scan statistic for survival data. *Biometrics* **2007**, *63*, 109–118. [CrossRef]
6. Bhatt, V.; Tiwari, N. A spatial scan statistic for survival data based on Weibull distribution. *Stat. Med.* **2014**, *33*, 1867–1876. [CrossRef]
7. Kulldorff, M.; Huang, L.; Konty, K. A scan statistic for continuous data based on the normal probability model. *Int. J. Health Geogr.* **2009**, *8*, 58. [CrossRef]
8. de Lima, M.S.; dos Santos, V.S.; Duczmal, L.H.; da Silva Souza, D. A spatial scan statistic for beta regression. *Spat. Stat.* **2016**, *18*, 444–454. [CrossRef]
9. Kulldorff, M.; Heffernan, R.; Hartman, J.; Assunçao, R.; Mostashari, F. A space–time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2005**, *2*, e59. [CrossRef]
10. Cucala, L. A distribution-free spatial scan statistic for marked point processes. *Spat. Stat.* **2014**, *10*, 117–125. [CrossRef]
11. Cucala, L.; Genin, M.; Lanier, C.; Occelli, F. A multivariate Gaussian scan statistic for spatial data. *Spat. Stat.* **2017**, *21*, 66–74. [CrossRef]
12. Loh, J.M.; Zhu, Z. Accounting for spatial correlation in the scan statistic. *Ann. Appl. Stat.* **2007**, *1*, 560–584. [CrossRef]
13. Zhang, T.; Zhang, Z.; Lin, G. Spatial scan statistics with overdispersion. *Stat. Med.* **2012**, *31*, 762–774. [CrossRef] [PubMed]
14. Murray, A.T.; Grubesic, T.H.; Wei, R. Spatially significant cluster detection. *Spat. Stat.* **2014**, *10*, 103–116. [CrossRef]
15. Dassios, A.; Jang, J.W. Kalman-Bucy Filtering for Linear Systems Driven by the Cox Process with Shot Noise Intensity and Its Application to the Pricing of Reinsurance Contracts. *J. Appl. Probab.* **2005**, *42*, 93–107. [CrossRef]
16. Bouzas, P.R.; Valderrama, M.J.; Aguilera, A.M. On the characteristic functional of a doubly stochastic Poisson process: Application to a narrow-band process. *Appl. Math. Model.* **2006**, *30*, 1021–1032. [CrossRef]
17. Vaillant, J. Negative binomial distributions of individuals and spatio-temporal Cox processes. *Scand. J. Stat.* **1991**, *18*, 235–248.
18. Diggle, P.J.; Moraga, P.; Rowlingson, B.; Taylor, B.M. Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Stat. Sci.* **2013**, *28*, 542–563. [CrossRef]
19. Møller, J.; Syversveen, A.R.; Waagepetersen, R.P. Log gaussian cox processes. *Scand. J. Stat.* **1998**, *25*, 451–482. [CrossRef]
20. Tamayo-Uria, I.; Mateu, J.; Diggle, P.J. Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spat. Stat.* **2014**, *9*, 192–206. [CrossRef]
21. Cressie, N. *Statistics for Spatial Data*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
22. Oscar, W.; Vaillant, J. Cox Processes Associated with Spatial Copula Observed through Stratified Sampling. *Mathematics* **2021**, *9*, 524. [CrossRef]
23. Nelsen, R.B. *An Introduction to Copulas*, 2nd ed.; Springer: New York, NY, USA, 2006.
24. Ma, X.; Luan, S.; Du, B.; Yu, B. Spatial copula model for imputing traffic flow data from remote microwave sensors. *Sensors* **2017**, *17*, 2160. [CrossRef] [PubMed]

25. Wolpert, R.L.; Ickstadt, K. Poisson/gamma random field models for spatial statistics. *Biometrika* **1998**, *85*, 251–267. [CrossRef]
26. Aguero-Valverde, J. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. *Accid. Anal. Prev.* **2013**, *50*, 289–297. [CrossRef]
27. Taylor, L. Assessing and interpreting the spatial distributions of insect populations. *Annu. Rev. Entomol.* **1984**, *29*, 321–357. [CrossRef]
28. Eastoe, E.F.; Tawn, J.A. Statistical models for overdispersion in the frequency of peaks over threshold data for a flow series. *Water Resour. Res.* **2010**, *46*, W02510. [CrossRef]
29. Chang, H.M.; Rosychuk, R.J. A spatial scan statistic for compound Poisson data, using the negative binomial distribution and accounting for population stratification. *Stat. Sin.* **2015**, *25*, 313–327. [CrossRef]
30. Sklar, M. Fonctions de repartition an dimensions et leurs marges. *Ann. L'Isup* **1959**, *8*, 229–231.
31. Davison, A.C.; Padoan, S.A.; Ribatet, M. Statistical modeling of spatial extremes. *Stat. Sci.* **2012**, *27*, 161–186. [CrossRef]
32. Tanner, M.A.; Wong, W.H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **1987**, *82*, 528–540. [CrossRef]
33. Smith, A.F.; Roberts, G.O. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B (Methodol.)* **1993**, *55*, 3–23. [CrossRef]
34. Landry, C.; Abadie, C.; Bonnot, F.; Vaillant, J. A Spatio-temporal Stochastic Model for an Emerging Plant Disease Spread in a Heterogeneous Landscape. *Int. J. Comput. Appl.* **2021**, *174*, 1–7. [CrossRef]
35. Valmy, L.; Vaillant, J. Bayesian Inference on a Cox Process Associated with a Dirichlet Process. *Int. J. Comput. Appl.* **2014**, *95*, 1–7. [CrossRef]
36. Baddeley, A.; Turner, R.; Møller, J.; Hazelton, M. Residual analysis for spatial point processes (with discussion). *J. R. Stat. Soc. Ser. Stat. Methodol.* **2005**, *67*, 617–666. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.