Statistica Sinica **25** (2015), 225-241
doi:http://dx.doi.org/10.5705/ss.2013.220w

SPATIAL SCAN STATISTICS FOR MODELS WITH OVERDISPERSION AND INFLATED ZEROS

Max S. de Lima[1], Luiz H. Duczmal[2], José C. Neto[1] and Letícia P. Pinto[2]

[1]*Federal University of Amazonas and* [2]*Federal University of Minas Gerais*

Abstract: The Spatial Scan Statistic is one of the most important methods for detecting and monitoring spatial disease clusters. Usually it is assumed that disease cases follow a Poisson or Binomial distribution. In practice, however, case count datasets frequently present an excess of zeroes and/or overdispersion, resulting in the violation of those commonly used models, increasing type I error occurrence. This paper describes a modification of the Spatial Scan Statistic with the Zero Inflated Double Poisson (ZIDP) model to reduce type I error, accommodating simultaneously an excess of zeroes and overdispersion. The null and alternative model parameters are estimated by the Expectation-Maximization algorithm and the p-value is obtained through the Fast Double Bootstrap Test. An application is presented for Hanseniasis data in the Brazilian Amazon.

Key words and phrases: Double Poisson, EM-algorithm, overdispersion, spatial scan statistics, zero inflated.

1. Introduction

The Spatial Scan Statistics (Kulldorff (1997)) is a popular method for the detection and inference of spatial disease clusters. Recently, several extensions have been devised to accommodate correlation (Loh and Zhu (2007)), covariate adjustment (Jung (2009)), log-linear modeling (Zhang and Lin (2009)), overdispersion (Zhang, Zhang, and Lin (2012)) and zero inflation (Cançado, da-Silva, and da Silva (2011, 2014)). In public health surveillance, the disease count variability is often greater than allowed by the Poisson model, which assumes that the mean and variance have the same value. This variability excess is called overdispersion and has been widely discussed in the literature. Disregarding the presence of overdispersion in the model may lead to the inflation of type I error and consequent erroneous inference for the model parameters. In the presence of overdispersion, the Generalized Poisson (Consul and Jain (1973)) and the Double Poisson (Efron (1986)) are more adequate data models. Another commonly occurring problem in count data, unexpected from the employed model, is that the dataset exhibits an excess of zeroes, or zero inflation. Overdispersion may sometimes occur as a consequence of zero inflation; in this case the Zero-Inflated

Poisson (**ZIP**) model offers a good adjustment to data. However, when overdispersion still persists, after adjusting for zero inflation modeling, a more robust model must be considered to accommodate additional overdispersion in positive count values.

Zero inflated models have been used in many areas (Hall (2000); Cheung (2002); Yau, Lee, and Carrivick (2004)). The estimation of parameters employing **ZIP** may also be severely biased when the positive counts exhibit significantly larger variability than expected. Then, good alternatives, modeling simultaneously zero inflation and overdispersion, are the Zero-Inflated Generalized Poisson (**ZIGP**), Double Poisson (**ZIDP**), or Negative Binomial (**ZINB**) models. In the context of spatial cluster detection, a common cause for overdispersion is spatial correlation (Houssian and Lawson (2006)); on the other hand, zero inflation occurs due to underreporting or absence of disease risk exposure for some groups of individuals.

Excessive false alarm may occur due to the simultaneous presence of zero inflation and overdispersion. In a simulated study, Perumean-Chaneya et al. (2012) verified that the Poisson based model estimates are inefficient, and statistically significant results may be lost when zero inflation is neglected. Likewise, when overdispersion is ignored, type I error estimates are inflated.

In the non-spatial context, a score test was proposed (Xiang et al. (2007)) to detect overdispersion based on a mixed **ZINB** model. The same type of score test was used through **ZIGP** (Yang, Harding, and Addyb (2010)). Another score test considered zero inflation and overdispersion simultaneously (Deng and Paul (2005)) in regression models (**ZINB**).

In the spatial context, a Spatial Scan Statistic for zero-inflated models **ZIP** was proposed (Cançado, da-Silva, and da Silva (2011, 2014)). Further, a Spatial Scan Statistic developed for overdispersion models was presented (Zhang, Zhang, and Lin (2012)), based on a Poisson-Gamma mixture.

In this paper, a modified Spatial Scan Statistics is developed, based on the **ZIDP** model, incorporating simultaneously zero inflation and overdispersion. The null and alternative model parameters are estimated by the EM (Expectation-Maximization) algorithm and the p-value is obtained through the Fast Double Bootstrap Test (Davidson and MacKinnon (2001)).

The paper is organized as follows. Section 2 reviews the Zero-Inflated Overdispersed Poisson model and the Spatial Scan Statistics. Section 3 presents the modified Spatial Scan Statistic with overdispersion and inflated zeros. Numerical studies with simulated data are reported in Section 4. Section 5 shows an application for Hanseniasis data in the Brazilian Amazon. Final remarks are in Section 6.

2. Background

2.1. Zero inflated overdispersed Poisson-ZIOP

Consider L locations with counts given by $\mathbf{Y} = (Y(s_1), \ldots, Y(s_L))'$, where $Y_i \equiv Y(s_i)$ is a random variable representing the number of disease cases at location s_i, with population at risk n_i and observed count value y_i. Zero-inflated models for Y_i are employed when the observed zero counts exceed the zero counts expected by the standard model. A typical example is given by the **ZIP** model, which assumes

$$Y_i \sim \begin{cases} 0 & \text{with probability} \quad p, \\ \mathcal{P}(\mu_i) & \text{with probability} \quad 1-p, \end{cases}$$

where $\mathcal{P}$ denotes the Poisson distribution. The resulting distribution is

$$P(Y_i = y_i) = \begin{cases} p + (1-p)e^{-\mu_i} & y_i = 0, \\ (1-p)\mathcal{P}(\mu_i) & y_i = 1, 2, \ldots. \end{cases}$$

It can be shown generally that

$$\mathbb{E}(Y_i) = (1-p)\mu_i \quad \text{and} \quad \mathbb{V}(Y_i) = (1-p)\sigma_i^2 + p(1-p)\mu_i^2, \tag{2.1}$$

where (μ_i, σ_i^2) denotes, respectively, the mean and variance of the standard model and p is the zero inflated parameter. If the zero inflation is ignored in the model, estimators will be inconsistent with the parameters.

Overdispersion appears when data variance is greater than predicted by the probabilistic model. Two mechanisms can cause overdispersion: data is generated by a process consisting of a mixture of two or more distributions; the observed data are not independent, but positively correlated. To treat overdispersion, Negative Binomial (**BB**), Generalized Poisson (**GP**) and Double Poisson (**DP**) models are utilized. Within the zero inflation context, **ZIGP** and **ZIDP** can be used to accommodate overdispersion in the **ZIP** model. Consider here the overdispersion **DP** model, with probability function

$$\tilde{f}_{DP}(y_i|\mu_i, \phi) = c(\mu_i, \phi) f_{DP}(y_i|\mu_i, \phi), \tag{2.2}$$

where the normalization constant satisfies the relation

$$\frac{1}{c(\mu_i, \phi)} = 1 + \frac{1-\phi}{12\mu_i\phi}\left(1 + \frac{1}{\mu_i\phi}\right),$$

and

$$f_{DP}(y_i|\mu_i, \phi) = (\phi^{1/2} e^{-\phi\mu_i})\left(\frac{e^{-y_i} y_i^{y_i}}{y_i!}\right)\left(\frac{e\mu_i}{y_i}\right)^{\phi y_i}. \tag{2.3}$$

(By convention, $0^0 = 1$ and $0\log(0) = 0$). Efron (1986) shows that

$$\mathbb{E}(Y_i) \doteq \mu_i \quad , \quad \mathbb{V}(Y_i) \doteq \frac{\mu_i}{\phi}, \tag{2.4}$$

and (2.3) is an approximation for (2.2). The approximate distribution has been used with success in temporal series modeling under overdispersion (Heinen (2003); Xu et al. (2012)) and easily accommodates covariate adjustment. In (2.4), it can be seen that ϕ is the parameter controlling overdispersion when $0 < \phi < 1$. If $\phi = 1$, then **DP** is the Poisson distribution.

To model simultaneously the zeroes excess and overdispersion in data, we propose the use of **ZIDP**(μ_i, ϕ, p) with probability function

$$P(Y_i = y_i|p, \mu_i, \phi) = \begin{cases} p + (1-p)f_{DP}(0|\mu_i, \phi) & y_i = 0, \\ (1-p)f_{DP}(y_i|\mu_i, \phi) & y_i = 1, 2, \ldots. \end{cases} \tag{2.5}$$

with $\mu_i = \theta n_i$. Combining (2.1) with (2.4),

$$\mathbb{E}(Y_i) = (1-p)\mu_i \quad \mathbb{V}(Y_i) = \mathbb{E}(Y_i)\left(p\mu_i + \frac{1}{\phi}\right). \tag{2.6}$$

Clearly, ϕ measures the overdispersion in the Zero-Inflated Poisson model. When $p = 0$ and $\phi = 1$, the model **ZIDP**$(\mu_i, 1, 0)$ is the standard Poisson $\mathcal{P}(\mu_i)$; when $p \neq 0$ and $\phi = 1$, the model **ZIDP**$(\mu_i, 1, p)$ is the **ZIP** model.

2.2 Spatial scan statistics

Given a study region represented by a geographic map divided into areas, each with an assigned population at risk and number of disease cases, the Spatial Scan Statistic (Kulldorff (1997)) is a test devised to identify a cluster (subset of the study area) with elevated incidence of cases compared to the rest of the map. This is a likelihood ratio test and makes use of a scanning procedure (the spatial scan) to search for the most likely cluster among the many candidate clusters in space or space-time. The simplest spatial version imposes circularly or elliptically shaped moving windows over the study region looking for compact clusters (Duczmal, Kulldorff, and Huang (2006), Duczmal et al. (2011)).

Specifically, let $\mathcal{S}$ be a study region projected in the Cartesian plane with L areas $\{s_1, \ldots, s_L\}$, population at risk $n(s_i) = n_i$. It is usual to determine, in the interior of each area s_i, a point (or *centroid*) a_i in the plane. Under the assumption of completely random distribution of cases (the null hypothesis H_0), let $Y_i \sim \mathcal{P}(\theta n_i)$ for every $s_i \in \mathcal{S}$. Let Z be a candidate cluster. Under the alternative hypothesis H_1, let $Y_i \sim \mathcal{P}(\theta_1 n_i)$ for every $s_i \in Z$ and $Y_i \sim \mathcal{P}(\theta_2 n_i)$ for every $s_i \notin Z$ with $\theta_1 > \theta_2$. The likelihood function for Z is given by

$$\mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y}) = \Big(\prod_{i=1}^{L} \frac{n_i^{y_i}}{y_i!}\Big)\theta_1^{y_z} e^{-\theta_1 n_z} \theta_2^{(y_+ - y_z)} e^{-\theta_2(n_+ - n_z)}, \tag{2.7}$$

where

$$y_+ = \sum_{i=1}^{L} y_i, \quad y_z = \sum_{s_i \in Z} y_i, \quad n_+ = \sum_{i=1}^{L} n_i \quad \text{e} \quad n_z = \sum_{s_i \in Z} n_i.$$

The likelihood ratio function for $H_0 : \theta_1 = \theta_2 = \theta$ versus $H_1 : \theta_1 > \theta_2$ is (Kulldorff (1997)):

$$\Lambda_Z = \frac{\max\limits_{\theta_1 > \theta_2} \mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y})}{\max\limits_{\theta_1 = \theta_2} \mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y})} = \left(\frac{y_z/n_z}{y_+/n_+}\right)^{y_z} \left(\frac{(y_+ - y_z)/(n_+ - n_z)}{y_+/n_+}\right)^{(y_+ - y_z)},$$

if $y_z/n_z > (y_+ - y_z)/(n_+ - n_z)$ and $\Lambda_Z = 1$ otherwise. With $\mathcal{Z}$ the collection of all cluster candidates Z, the Spatial Scan Statistics is defined as

$$\Lambda = \max_{Z \in \mathcal{Z}} \Lambda_Z, \tag{2.8}$$

and the *most likely cluster* is $\hat{Z} = \arg\big(\max\limits_{Z \in \mathcal{Z}} \Lambda_Z\big)$. A Monte Carlo procedure is usually employed to obtain the test p-value. The Circular Scan is the most popular variant of the Spatial Scan Statistic (Kulldorff (1999)): given the area $s_{i_1} = s_i$ with centroid $a_{i_1} = a_i$, consider the L areas $(s_{i_1}, \ldots, s_{i_L})$ with the respective centroids $(a_{i_1}, \ldots, a_{i_L})$ sorted by their increasing order of distance from the centroid a_i. The candidate clusters $z_{i_m} = \{s_{i_1}, \ldots, s_{i_m}\}$, $i = 1, \ldots, L$, $m = 1, \ldots, S$ (not all distinct) form the collection of circular clusters of maximum size S, $S = 1, \ldots, L$.

3. Spatial Scan Statistics with Overdispersion and Inflated Zeros

3.1. Spatial scan statistics for ZIDP models

In order to accommodate simultaneously an excess of zeroes and overdispersion, suppose that the data $\mathbf{Y} = (Y(s_1), \ldots, Y(s_L))'$ are modeled by the $\mathbf{ZIDP}(\mu_i, \phi, p)$ model, with distribution given by (2.5). Following Kulldorff's (1997) cluster model, assume that $\mu_i = \theta_1 n_i$ when $s_i \in Z$, and $\mu_i = \theta_2 n_i$ when $s_i \notin Z$. Consider testing $H_0 : \theta_1 = \theta_2 = \theta$ against $H_1 : \theta_1 > \theta_2$. For a given Z, under H_1, the likelihood function is

$$\begin{aligned}
&\mathcal{L}_Z(p, \theta_1, \theta_2, \phi; \mathbf{y}) \\
&= \prod_{s_i \in Z} \left(p + (1-p) f_{DP}(0|\theta_1 n_i, \phi)\right)^{1 - I(y_i > 0)} \left((1-p) f_{DP}(y_i|\theta_1 n_i, \phi)\right)^{I(y_i > 0)} \\
&\quad \times \prod_{s_i \notin Z} \left(p + (1-p) f_{DP}(0|\theta_2 n_i, \phi)\right)^{1 - I(y_i > 0)} \left((1-p) f_{DP}(y_i|\theta_2 n_i, \phi)\right)^{I(y_i > 0)},
\end{aligned}$$

where $I(y_i > 0)$ is the indicator function of positive value occurrence. Under H_0 the likelihood function is

$$\mathcal{L}_0(p,\theta,\phi;\mathbf{y}) = \prod_{i=1}^{L}\left(p+(1-p)f_{DP}(0|\theta n_i,\phi)\right)^{1-I(y_i>0)}\left((1-p)f_{DP}(y_i|\theta n_i,\phi)\right)^{I(y_i>0)}.$$

Let $(\hat{p}_1,\hat{\theta}_1,\hat{\theta}_2,\hat{\phi}_1)$ and $(\hat{p}_0,\hat{\theta}_0,\hat{\phi}_0)$ be respectively the maximum likelihood estimators for the parameters of the model under H_1 and H_0. Then the likelihood ratio statistic and the Spatial Scan Statistics for the ZIDP model are, respectively,

$$\hat{\Lambda}_Z = \frac{\mathcal{L}_Z(\hat{p}_1,\hat{\theta}_1,\hat{\theta}_2,\hat{\phi}_1;\mathbf{y})}{\mathcal{L}_0(\hat{p}_0,\hat{\theta}_0,\hat{\phi}_0;\mathbf{y})} \quad \text{and} \quad \hat{\Lambda} = \max_{Z\in\mathcal{Z}}\hat{\Lambda}_Z, \tag{3.1}$$

with estimated cluster $\hat{Z} = \arg\big(\max_{Z\in\mathcal{Z}}\hat{\Lambda}_Z\big)$. By inspecting $\mathcal{L}_Z(.;\mathbf{y})$ and $\mathcal{L}_0(.;\mathbf{y})$ it may be noted that there is no independence between the parameter p and the remaining parameters. This fact complicates the maximization of the likelihood function, especially when there are covariates involved. Thus, the inclusion of a latent vector of variables is necessary to factorize the likelihood to facilitate the maximization process, making use of the EM (Expectation-Maximization) algorithm. Let $\mathbf{U} = (U_1,\ldots,U_L)$, where $U_i = 1$ when Y_i occurs due to a zero state, and $U_i = 0$ when Y_i occurs due to a **DP** model. Assume that $U_i \sim$ Bernoulli(p). Then the augmented likelihood is

$$\mathcal{L}_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u}) = \prod_{s_i\in Z} p^{u_i}\left[(1-p)f_{DP}(y_i|\theta_1 n_i,\phi)\right]^{1-u_i} \times \prod_{s_i\notin Z} p^{u_i}\left[(1-p)f_{DP}(y_i|\theta_2 n_i,\phi)\right]^{1-u_i}.$$

Marginally, $Y_i \sim \mathbf{ZIDP}(\mu_i,\phi,p)$. The logarithm of the likelihood ratio for the ZIDP model under H_1 is

$$\begin{aligned} l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{L}\left(u_i\log p+(1-u_i)\log(1-p)\right)+\sum_{s_i\in Z}(1-u_i)\log f_{DP}(y_i|\theta_1 n_i,\phi) \\ &\quad+\sum_{s_i\notin Z}(1-u_i)\log f_{DP}(y_i|\theta_2 n_i,\phi) \\ &= l_Z^a(p;\mathbf{u})+l_Z^a(\theta_1,\phi;\mathbf{y},\mathbf{u})+l_Z^a(\theta_2,\phi;\mathbf{y},\mathbf{u}), \end{aligned} \tag{3.2}$$

and under H_0 is

$$\begin{aligned} l_0^a(p,\theta,\phi;\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{L}\left(u_i log p+(1-u_i)\log(1-p)\right)+\sum_{i=1}^{L}(1-u_i)\log f_{DP}(y_i|\theta n_i,\phi) \\ &= l_0^a(p;\mathbf{u})+l_0^a(\theta,\phi;\mathbf{y},\mathbf{u}). \end{aligned} \tag{3.3}$$

Here the likelihood is easily maximized and the estimators $(\hat{p}_1, \hat{\theta}_1, \hat{\theta}_2, \hat{\phi}_1)$ and $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$ may be independently obtained. The estimator for ϕ in H_1 is obtained by maximizing $l_Z^a(\phi; \mathbf{y}, \mathbf{u}) = l_Z^a(\hat{\theta_1}, \phi; \mathbf{y}, \mathbf{u}) + l_Z^a(\hat{\theta_2}, \phi; \mathbf{y}, \mathbf{u})$, and for H_0 it is obtained by maximizing $l_0^a(\hat{\theta_0}, \phi; \mathbf{y}, \mathbf{u})$. To maximize (3.2) and (3.3) the EM algorithm is used. In this case the logarithm of the likelihood function is maximized iteratively in two steps until convergence. The maximization of $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u})$ is obtained as follows.

- Step E: Initialize the iterative process with $\boldsymbol{\gamma}^{(0)} = (p_1^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}, \phi_1^{(0)})$. At the $(k+1)$th iteration the estimate of $u_i^{(k)}$ is the conditional mean over $\mathbf{y}$ and the current estimates $\boldsymbol{\gamma}^{(k)} = (p_1^{(k)}, \theta_1^{(k)}, \theta_2^{(k)}, \phi_1^{(k)})$. Thus compute $\mathbb{E}\{l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}) | \mathbf{y}, \boldsymbol{\gamma}^{(k)}\}$ with respect to the conditional distribution of $\mathbf{u}$. As $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u})$ is linear in $\mathbf{u}$, this is $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$, where $\mathbf{u}^{(k)} = \mathbb{E}_{H_1}(\mathbf{u} | \mathbf{y}, \boldsymbol{\gamma}^{(k)})$, with the ith element

$$u_i^{(k)} = P_{H_1}(u_i = 1 | y_i, \boldsymbol{\gamma}^{(k)})$$
$$= \frac{P_{H_1}(Y_i = y_i | u_i = 1, \boldsymbol{\gamma}^{(k)}) P_{H_1}(u_i = 1 | p_1^{(k)})}{P_{H_1}(Y_i = y_i | u_i = 1, \boldsymbol{\gamma}^{(k)}) P_{H_1}(u_i = 1 | p_1^{(k)}) + P_{H_1}(Y_i = y_i | u_i = 0, \boldsymbol{\gamma}^{(k)}) P_{H_1}(u_i = 0 | p_1^{(k)})}$$

and

$$u_i^k = \begin{cases} \left(1 + \exp\{-\log(\frac{p_1^{(k)}}{1-p_1^{(k)}}) - \phi_1^{(k)} \theta_1^{(k)} n_i + \frac{1}{2} \log \phi_1^{(k)}\}\right)^{-1} & \text{if} \quad y_i = 0, s_i \in Z, \\ \left(1 + \exp\{-\log(\frac{p_1^{(k)}}{1-p_1^{(k)}}) - \phi_1^{(k)} \theta_2^{(k)} n_i + \frac{1}{2} \log \phi_1^{(k)}\}\right)^{-1} & \text{if} \quad y_i = 0, s_i \notin Z, \\ 0 & \text{if} \quad y_i > 0. \end{cases}$$

- Step M: Maximize $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$.
  1. Step M for p : In the $(k+1)$th iteration maximize $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ with respect to p, equivalently maximize $l_Z^a(p; \mathbf{u})$ as (3.3) considering $\mathbf{u} = \mathbf{u}^{(k)}$. Analytically, $p_1^{(k+1)} = \sum_{i=1}^{L} u_i^{(k)} / L$ and $\hat{p}_1$ is the value $p_1^{(k+1)}$ satisfying $|p_1^{(k+1)} - p_1^{(k)}| < \epsilon$.
  2. Step M for θ_1: In the $(k+1)$th iteration maximize $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ with respect to θ_1, equivalently to maximize $l_Z^a(\theta_1, \phi; \mathbf{y}, \mathbf{u})$ as (3.3) considering $\mathbf{u} = \mathbf{u}^{(k)}$. Analytically, ${\theta_1}^{(k+1)} = \sum_{s_i \in Z} (1 - u_i^{(k)}) y_i / \sum_{s_i \in Z} (1 - u_i^{(k)}) n_i$ and $\hat{\theta}_1$ is the quantity ${\theta_1}^{(k+1)}$ satisfying $|{\theta_1}^{(k+1)} - {\theta_1}^{(k)}| < \epsilon$.
  3. Step M for θ_2: Similar to Step M for θ_1 subtitute $l_Z^a(\theta_1, \phi; \mathbf{y}, \mathbf{u})$ by $l_Z^a(\theta_2, \phi; \mathbf{y}, \mathbf{u})$. Then ${\theta_2}^{(k+1)} = \sum_{s_i \notin Z} (1 - u_i^{(k)}) y_i / \sum_{s_i \notin Z} (1 - u_i^{(k)}) n_i$ and $\hat{\theta}_2$ is the quantity ${\theta_2}^{(k+1)}$ satisfying $|{\theta_2}^{(k+1)} - {\theta_2}^{(k)}| < \epsilon$.
  4. Step M for ϕ: In the $(k+1)$th iteration maximize $l_Z^a(\theta_1^{(k+1)}, \phi; \mathbf{y}, \mathbf{u}) +$

$l_Z^a(\theta_2^{(k+1)}, \phi; \mathbf{y}, \mathbf{u})$ with respect to ϕ considering $\mathbf{u} = \mathbf{u}^{(k)}$. Analytically,

$$\phi_1^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})}{2\left\{\sum_{s_i\in Z}(1-u_i^{(k)})y_i\log(\theta_i/\theta_1^{(k+1)}) + \sum_{s_i\notin Z}(1-u_i^{(k)})y_i\log(\theta_i/\theta_2^{(k+1)})\right\}},$$

where $\theta_i = y_i/n_i$ and $\hat{\phi}_1 = min\{1, \phi_1^{(k+1)}\}$ with $\phi_1^{(k+1)}$ satisfying $|\phi_1^{(k+1)} - \phi_1{}^{(k)}| < \epsilon$.

The maximization of $l_0^a(p, \theta, \phi; \mathbf{y}, \mathbf{u})$ is processed similarly to the maximization of $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ with the following modification. At step E, under H_0, use

$$u_i^k = \begin{cases} \left(1+\exp\{-\log(\frac{p_0^{(k)}}{1-p_0^{(k)}}) - \phi_0^{(k)}\theta_0^{(k)}n_i + \frac{1}{2}\log\phi_0^{(k)}\}\right)^{-1} & \text{if } y_i = 0, i = 1, \ldots, L, \\ 0 & \text{if } y_i > 0. \end{cases}$$

Now maximize $l_0^a(p, \theta, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ with respect to the parameters, obtaining at the $(k+1)$th iteration,

$$p_0^{(k+1)} = \frac{\sum_{i=1}^{L} u_i^{(k)}}{L}, \quad \theta_0^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})y_i}{\sum_{i=1}^{L}(1-u_i^{(k)})n_i},$$

$$\phi_0^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})}{2\left\{\sum_{i=1}^{L}(1-u_i^{(k)})y_i\log(\theta_i/\theta_0^{k+1})\right\}}.$$

After the convergence of the algorithm, denote the estimates via the EM algorithm by $(\hat{p}_1, \hat{\theta}_1, \hat{\theta}_2, \hat{\phi}_1)$, $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$ and compute $(\hat{\Lambda}_Z, \hat{\Lambda})$ given in (3.1). Now, using $\hat{\Lambda}$, the spatial cluster may be identified under an excess of zeroes and overdispersion.

3.2. Fast Double Bootstrap-EM for the p-value computation

As the distribution of $\hat{\Lambda}$ is not available analytically, the statistic p-value is computed using the Fast Double Bootstrap Test (Davidson and MacKinnon (2001)), jointly with the application of the EM algorithm for each new dataset generated under the null hypothesis. The Fast Double Bootstrap procedure is necessary in this situation because the parameters of the $\hat{\Lambda}$ distribution are unknown under the null hypothesis.

Under H_0, Y_i is a Bernoulli(p)-$\mathbf{DP}(\theta n_i, \phi)$ mixture. By Efron (1986),

$$X_i \sim \mathcal{P}(\theta n_i \times \phi) \Longrightarrow \left(\frac{X_i}{\phi}\right) \dot{\sim} \mathbf{DP}(\theta n_i, \phi).$$

Given (p_0, θ_0, ϕ_0), Y_i is generated from the **ZIDP**$(n_i\theta_0, \phi_0, p_0)$ model as follows.

- Algorithm **ZIDP**$(n_i\theta_0, \phi_0, p_0)$
  1. Generate $x_i \sim \mathcal{P}(\theta_0 n_i \times \phi_0)$ and $v_i \sim Uniform(0,1)$.
  2. If $v_i \leq p_0$ let $y_i = 0$. Else $y_i = x_i/\phi_0$.

The p-value is computed as follows.

- Fast Double Bootstrap-EM algorithm for $\hat{\Lambda}$.
  1. Based on data $\mathbf{y} = (y_1, \ldots, y_L)$, use the EM algorithm and compute $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$. Derive the observed value $\hat{\Lambda}$ and denote it by $\hat{\lambda}$.
  2. Generate $\mathbf{y}_b^* = (y_{1,b}^*, \ldots, y_{L,b}^*)$ using the EM-algorithm **ZIDP** with (p_0, θ_0, ϕ_0) substituted by $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$.
  3. Based on generated data $\mathbf{y}_b^*$, use the EM algorithm and compute the pseudo-estimators $(\hat{p}_{0,b}^*, \hat{\theta}_{0,b}^*, \hat{\phi}_{0,b}^*)$ for (p_0, θ_0, ϕ_0). Derive the pseudo-value of $\hat{\Lambda}_b^*$ and denote it by $\hat{\lambda}_b^*$.
  4. Repeat Steps 2 and 3 for $b = 1, \ldots, B$, compute the usual p-value for $\hat{\Lambda}$ as
$$p_{value}^* \doteq p_{value}^*(\hat{\Lambda}) = \sum_{b=1}^{B+1} \frac{I(\hat{\lambda} \geq \hat{\lambda}_b^*)}{(B+1)}, \quad \text{with} \quad \hat{\lambda}_{B+1}^* = \hat{\lambda}.$$
  5. Generate $\mathbf{y}_b^{**} = (y_{1,b}^{**}, \ldots, y_{L,b}^{**})$ using the **ZIDP** algorithm with (p_0, θ_0, ϕ_0) substituted by $(\hat{p}_{0,b}^*, \hat{\theta}_{0,b}^*, \hat{\phi}_{0,b}^*)$. Using Steps 3 and 4, derive $\hat{\Lambda}_b^{**}$ and denote it by $q_{1-p_{value}^*}^{**}$, the $1 - p_{value}^*$-quantile of the empirical distribution of $\hat{\Lambda}_b^{**}$. This quantile is the solution of the equation
$$\frac{1}{B}\sum_{b=1}^{B} I(\hat{\Lambda}_b^{**} > q_{1-p_{value}^*}^{**}) = p_{value}^*.$$
  6. Compute the fast double bootstrap $p - value$ for $\hat{\Lambda}$ by
$$p_{value}^{**} \doteq p_{value}^{**}(\hat{\Lambda}) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\Lambda}_b^* > q_{1-p_{value}^*}^{**}).$$

The convergence of the ZIDP EM algorithm is studied through simulations, and a proof of the convergence is also given (see the Supplementary Materials Section). A program implementing the ZIOP algorithm was written in C language, and can be requested from the corresponding author.

4. A Simulation Study

The zero inflation and overdispersion effects on type I error probability and power of detection for the four Poisson based Spatial Scan Statistic models are

evaluated in this section, namely the Poisson (**ScanP**), Zero Inflated Poisson (**ScanZIP**), Overdispersed Poisson (**ScanOP**), and Zero Inflated Overdispersed Poisson (**ScanZIOP**). The **ScanZIOP** is represented by the **ZIDP** model and the **ScanOP** is obtained from the **ZIDP** model by using $p = 0$.

The study region is the Amazonas state in Brazil with $L = 62$ municipalities (Figure 1). The populations at risk consist of children under 15 years living in 2010. Alternative hypotheses models with artificial clusters were simulated to evaluate the power of detection, and null hypothesis model maps were simulated to evaluate the type I error. For each model, 1,000 Monte Carlo replications were generated. An artificial circularly shaped (Kulldorff (1999)) spatial cluster $Z = \{\text{Anori, Coari, Codajás, Tefé, Tapauá}\}$ is located in the central part of the study region (Figure 1(D)).

Under null hypothesis, $\mu_i = n_i\lambda_0$, where $\lambda_0 = 0.001$ is a global rate reference for the disease; under the alternative model, $\mu_i = n_i\lambda_0(1+\theta)$ for every $s_i \in Z$ and $\mu_i = n_i\lambda_0$ otherwise, where $\theta > 0$ indicates the cluster intensity. Note that $\theta = 0$ under the null model.

The simulation procedure was given by

- (1) Generate 1,000 Monte Carlo replications under H_0, with data generated by $\mathcal{P}(n_i\lambda_0)$ and estimate the upper 5% quantile for each one of the four empirical distributions of the methods **ScanP**, **ScanZIP**, **ScanOP**, and **ScanZIOP**.
- (2) Generate 1,000 Monte Carlo replications under the null ($\theta = 0$) and alternative ($\theta = \{0.5, 1.0, 2.0\}$) models with overdispersion $1/\phi = \{1, 1.5, 2.0, 3.0\}$, zero inflation $p = \{0, 0.1, 0.2, 0.3\}$; estimate empirically the type I error and power of detection using the critical value given by the previously obtained upper 5% quantile.
  Let the detected most likely cluster $\hat{Z}^{(q)}$ obtained in the qth simulation be the estimator of the artificial cluster Z (#{A} indicates the cardinality of the set A).
- The precision for the cluster detection was evaluated by the following measures:
  - Sensitivity-(**SS**)= the average ratio of the number of locations correctly detected by the number of locations belonging to the artificial cluster:

$$\mathbf{SS} = \frac{1}{1,000}\sum_{q=1}^{1,000}\left(\frac{\#\{\hat{Z}^{(q)} \cap Z\}}{\#Z}\right),$$

  - Positive Predicted Value-(**PPV**)= the average ratio of the number of locations correctly detected by the number of locations belonging to the detected cluster:

$$\mathbf{PPV} = \frac{1}{1,000} \sum_{q=1}^{1,000} \left(\frac{\#\{\hat{Z}^{(q)} \cap Z\}}{\#\{\hat{Z}^{(q)}\}} \right),$$

The measures **SS** and **PPV** evaluate the performance of the methods according to their ability to locate the cluster, when it exists.

The simulation results are summarized on Tables 2, 3 and 4 in the Supplementary Materials Section.

In the absence of zero inflation ($p = 0$) and overdispersion ($\phi = 1$), type I error probability is adequate for all four methods (see Table 2 in the Supplementary Materials Section). With zero inflation ($p > 0$) but no overdispersion ($\phi = 1$), the type I error probability for the **ScanZIP** and **ScanZIOP** stay below 5%, whereas the corresponding values for **ScanP** and **ScanOP** are elevated, showing their inefficiency in this situation. In the absence of zero inflation ($p = 0$) and in the presence of overdispersion ($1/\phi > 1$), the **ScanOP** and **ScanZIOP** attain the lowest type I error probability; those values are somewhat larger than 5% due to the fact that their null hypothesis critical values 5% quantiles were obtained under the assumption that the true model is Poisson. However, these probabilities decrease when the overdispersion increases. The **ScanP** and **ScanZIP** attain large type I error probability values, making both of them inadequate for this scenario. When zero inflation and overdispersion occur simultaneously ($p > 0$ and $1/\phi > 1$), the three first methods, **ScanP**, **ScanOP** and **ScanZIP**, exhibit large values of type I error probability; only the **ScanZIOP** method presents an adequate performance.

According to Table 3 of the Supplementary Materials Section, the power of detection is greater in the presence of overdispersion and zero inflation for the **ScanP** and **ScanZIP**, as expected, as these methods attained high values of probability of type I error. The only reliable power estimate in this scenario is the one for the **ScanZIOP**. In the simulations, it was also observed that **ScanZIOP**'s power increases rapidly with small increases in cluster cases intensity ($\theta > 0$). When the cluster intensity and zero inflation remain fixed, power decreases. The same effect is observed when the cluster intensity and overdispersion remain fixed. This is evidence that the **ScanZIOP** is better suited to detect spatial clusters for small values of zero inflation and overdispersion.

From the results in Table 4 of the Supplementary Materials Section, **SS** and **PPV** are low for the **ScanOP** under zero inflation and overdispersion but increase as the cluster intensity increases. The **ScanP** attains low **PPV** values and sensitivity decreases when the cluster intensity increases, an indication that the **ScanP** tends to detect larger clusters than the true cluster. The methods **ScanZIP** and **ScanZIOP** behave similarly in terms of precision: the **SS** and

PPV measures increase when the cluster intensity increases. When cluster intensity is small ($\theta = 0.5$) the **ScanZIP** has more precision than the **ScanZIOP**. However, as the cluster intensity increases, the differences are negligible.

The artificial cluster $Z_1 = \{$Anori, Coari, Codajás, Tefé, Tapauá$\}$ of Figure 1(D) is located in the central part of the map, including about 8% of the total population. On the other hand, the small population artificial cluster $Z_2 = \{$Fonte Boa, Japurá Jutaí Maraã Tonantins$\}$ to the west contains only 3.5% of the total population. The power of detection of **ScanZIOP** was compared for those two population clusters. The results for those alternative model sets, with 1,000 simulations each, are presented in Tables 5 and 6 of the Supplementary Materials Section. The power is almost the same, except for $\theta = 0.5$, when there is a slight reduction of power for Z_2, compared to the Z_1 cluster.

5. Application: Hanseniasis Clusters

This study uses data for new Hanseniasis cases in children under 15 years old in the Amazonas state, Brazil, from 2008 to 2010 for each of their 62 municipalities. The dataset was divided into two periods: 2008/2009 (207 new cases in two years, 0.0000831 cases per child per year) and 2010 (190 new cases, 0.0001525 cases per child per year), see Figure 1 (A and B). Hanseniasis is an endemic contageous disease related to extreme poverty. In the 2008/2009 period, 20 municipalities (32%) registered zero new cases, compared with 30 municipalities (48%) that registered zero new cases in 2010 alone. In the 2008/2009 period, the average of the 62 municipalities' rates of new cases for 10,000 persons was 2.944, with variance equal to 6.776 (in the two years period). In the 2010 period, the corresponding mean and variance values were respectively 2.706 and 7.421 in the one year period. Figure 2 A and B displays the rates for the 62 municipalities. As the variance is substantially greater than the mean for those two scenarios, the **ZIDP** model seems quite plausible.

In this application, the Circular Scan employs the collection of circular clusters with maximum size $S = 15$ (25% of the municipalities), for the four models of Section 3: **ScanP**, **ScanZIP**, **ScanOP** and **ScanZIOP**.

The results are shown in Table 1.

In the 2008/2009 period, **ScanZIOP** and **ScanOP** did not detect significant clusters (p-value=0.114 and 0.112, respectively). The estimated overdispersion by **ScanZIOP** was $1/\hat{\phi}_0 = 2.325$, the zero inflation was below 1% ($\hat{p}_0 = 0.009$), and the cases rate was 1.67 per 10,000 persons ($\hat{\theta}_0 = 0.000167$). However, **ScanZIP** and **ScanP** detected a significant cluster (both with p-value=0.001). The zero inflation estimated by **ScanZIP** was $\hat{p}_1 = 0.013$, with estimated rates inside and outside the cluster given by $\hat{\theta}_1 = 0.000427$, and $\hat{\theta}_2 = 0.000149$, respectively (the estimated relative risk was 2.866). Taking into account that **ScanZIP**

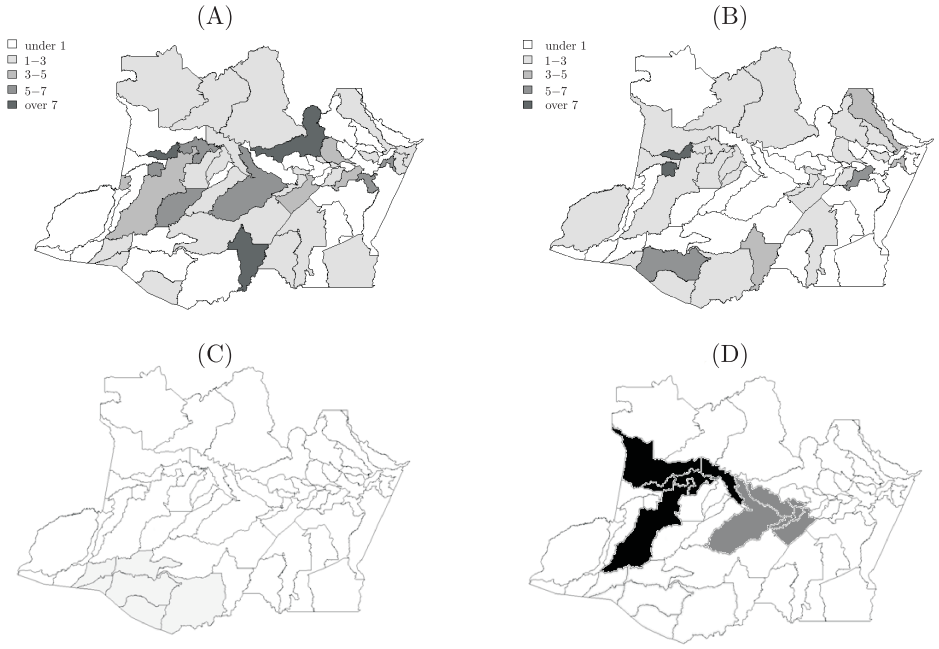

Figure 1. Spatial distribution of Hanseniasis cases: 2008/2009 (A) and 2010 (B). Detected Cluster in 2010 (C). Artificial cluster used in the simulations (section 4) (D).

Table 1. Spatial Clusters of new cases of Hanseniasis for 2008/09 and 2010.

Year	Scan	$\log \hat{\Lambda}$	p-value	$(\hat{p}_0, \hat{\phi}_0, 1000 \times \hat{\theta}_0)$	$(\hat{p}_1, \hat{\phi}_1, 1000 \times \hat{\theta}_1, 1000 \times \hat{\theta}_2)$
	ScanP	12.510	0.001	(0.000, 1.000, 0.166)	(0.000, 1.000, 0.427, 0.148)
2008/09	**ScanZIP**	11.074	0.001	(0.010, 1.000, 0.167)	(0.013, 1.000, 0.427, 0.149)
	ScanOP	5.886	0.122	(0.000, 0.428, 0.166)	(0.000, 0.518, 0.427, 0.148)
	ScanZIOP	5.882	0.114	(0.009, 0.430, 0.167)	(0.013, 0.521, 0.427, 0.149)
	ScanP	16.785	0.001	(0.000, 1.000, 0.152)	(0.000, 1.000, 0.518, 0.134)
2010	**ScanZIP**	15.955	0.001	(0.224, 1.000, 0.171)	(0.013, 1.000, 0.597, 0.154)
	ScanOP	7.696	0.034	(0.000, 0.406, 0.152)	(0.000, 0.520, 0.517, 0.134)
	ScanZIOP	8.849	0.006	(0.224, 0.442, 0.171)	(0.258, 0.689, 0.597, 0.154)

does not accommodate overdispersion in the positive counts, this cluster significance value is doubtful.

In the 2010 period, the four methods detected the same cluster (Figure 1 (C)), with 30 new cases when the expected number was $(1 - \hat{p}_1) n_{\hat{Z}} \hat{\theta}_1 = 25.67$. The zero inflation and overdispersion estimated by **ScanZIOP** was $\hat{p}_1 = 0.258$ and $1/\hat{\phi}_1 = 1.471$ respectively. The cluster is situated in a region well known for

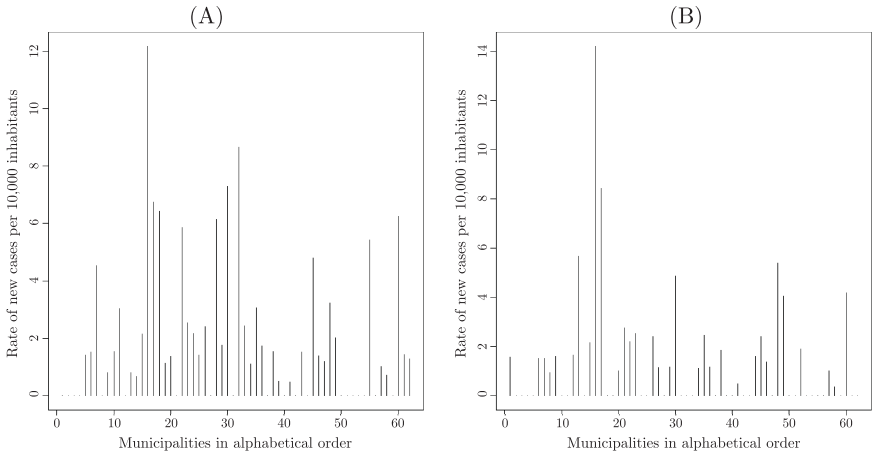

Figure 2. New cases of Hanseniasis in the Brazilian Amazon, 2008/2009 (A) and 2010 (B).

its high social vulnerability index.

6. Final Remarks

A modification of the Spatial Scan Statistics, the Zero Inflated Double Poisson Scan (**ZIPD**), is proposed to accommodate simultaneously an excess of zeroes and overdispersion. It might also be useful in disease surveillance, where the excessive variation for positive counts is frequent.

Sometimes, when the usual scan statistic is used under the null hypothesis of constant rate, a small p-value may result due to the high variability in the number of cases among a reduced number of areas or, alternatively, a small variability among many areas. This may cause in turn the existence of a false positive cluster; this anomaly could be avoided by changing the usual Poisson model by an overdispersed model. This kind of problem was evident in Section 4 (simulations) and Section 5 (applications). The simulations show that accounting for the presence of overdispersion and zero inflation in the **ZIDP** model reduces substantially the probability of type I error, compared to the Poisson, overdispersed Poisson, and zero inflated Poisson, shown here to be inadequate in those scenarios. That means that when a cluster is not detected by the **ZIDP**, and detected by the other methods, it should be carefully analyzed before being recognized as a legitimate cluster.

In the presence of overdispersion for positive count values, the detection of spatial clusters based on the zero-inflated model may be not the best option. In this situation the **ZIDP** Spatial Scan is a more flexible approach, but not the only one. The Binomial Negative (**NB**), Beta-Binomial (**BB**), and Generalized Poisson (**GP**) can also treat overdispersion and, similarly to the **ZIPD**,

it is possible to detect and evaluate spatial clusters based on the **ScanZINB**, **ScanZIBB** and **ScanZIGP** models. The significance of clusters found using those methods may be also assessed using the same strategy based on the Fast Double Bootstrap employed in this paper.

Spatial correlations could also be modeled with the proposed approach. These may be present due to the contagious nature of the disease, heterogeneous distribution of phenotypic traits, environmental causes, or to some latent variables that are related to the disease but not included in the data collection or in the model (Loh and Zhu (2007)). In fact, the objective of the cluster detection process is to see whether the counts from different locations are spatially correlated or not. The existence of a spatial cluster is an indication of the presence of spatial correlation, it signals the presence of a subregion with anomalous counts compared to the rest of the study region. Two approches can be used to tackle this problem, depending on how easily one can identify the spatial correlation factors.

Spatial correlation can be added to the model in order to include some known specific feature related to the population. As example, female population age is known to be strongly related to the occurrence of breast cancer, and a covariate may be added to the model in order to take into account this feature: the usual procedure is to stratify the population of each area by age and recompute the spatial counts, thus reducing the case counts for locations with older than average population. If eventually some breast cancer cluster is found in the modified study region, then it is not due to the age effect (supposing that the stratification was carefully done!). If the study region is not corrected for the age covariate, a cluster may be found that is simply consequence of the concentration of older people in some part of the study region. The ZIOP model allows the introduction of covariates in a straightforward manner, similarly to the other models compared in our work.

When the factors causing the spatial correlation cannot be easily identified, the algorithm of Section 2.3 of (Loh and Zhu (2007)) is a good option. In this case, the number of expected cases in the area i can be rewritten as

$$\mu_i = \exp(\log(n_i) + \theta_i I_{\{s_i \in Z\}} + v_i),$$

where $\log(n_i)$ is the populational adjustment, θ_i is the parameter measuring the intensity of cases in the cluster Z compared to the exterior of Z, and v_i is the random effect used to capture the spatial dependence. The ZIOP model could be adapted to use this modification without additional problems, similarly to the other models compared in our work.

Acknowledgements

The authors thank the Editors and two anonymous reviewers for their comments, which helped to improve the manuscript. The authors were funded with grants from the Brazilian agencies CAPES, UFAM, CNPq and FAPEMIG.

References

Cançado, A. L. F., da-Silva, C. Q. and da Silva, M. F. (2011). A zero-inflated Poisson-based spatial scan statistic. *Emerging Health Threats J.* 4.

Cançado, A. L. F., da-Silva, C. Q. and da Silva, M. F. (2014). A spatial scan statistic for zero-inflated Poisson process. *Envir. Ecol. Stat.* (doi: 10.1007/s10651-013-0272-1).

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Statist. Medicine* **21**, 1461-1469.

Consul, P. C. and Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics* **15**, 791-799.

Davidson, R. and MacKinnon, J. G. (2001). Improving the reliability of bootstrap tests. Queen's University Institute for Economic Research Discussion Paper, No. 995, revised.

Deng, D. and Paul, S. R. (2005). Score test for zero-inflated and over-dispersion in generalized linear models. *Statist. Sinica* **15**, 257-276

Duczmal, L., Kulldorff, M. and Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease clusters. *J. Comput. Graph. Statist.* **15**, 428-442.

Duczmal, L. H., Moreira, G. J. P., Burgarelli, D., Takahashi, R. H. C., Magalhães, F. C. O. and Bodevan, E. C. (2011). Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. *Int. J. Health Geogr.* **10**, 29.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81**, 709-721.

Hall, D. B. (2000). Zero inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030-1039.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional Poisson Model. CORE Discussion Paper, No. 2003-63. University of Louvain. Belgium.

Houssian, M. M. and Lawson, A. B. (2006). Cluster detection diagnostics for small area health data, with reference to evaluation of local likelihood models. *Statist. Medicine* **25**, 771-786.

Jung, I. (2009). A generalized linear models approach to spatial scan statistics for covariate adjustment. *Statist. Medicine* **28**, 1131-1143.

Kulldorff, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* **26**, 1481-1496.

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations and applications. *Scan Statistics and Applications*. (Edited by Glaz and Balakrishnan), 303-322. Birkhauser, Boston.

Loh, J. M. and Zhu, Z. (2007). Accounting for spatial correlation in the scan statistic. *Ann. Appl. Statist.* **1**, 560-584.

Perumean-Chaneya, S. E., Morganb., C., McDowallc., D. and Aband., I. (2012). Zero-inflated and overdispersion: what's one to do? *J. Statist. Comput. Simulation*, 1-13.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statist. Sinica* **15**, 831-840.

Xiang, L., Lee, A.H., Yau, K. K. W. and McLachlan, G. J. (2007). A score test for overdispersion in zero-inflated Poisson mixed regression model. *Statist. Medicine* **26** , 1608-1622.

Xu, H. Y., Xiea, M., Goha, T. N. and Fub, X. (2012). A model for interger-valued time series with conditional overdispersion. *Comput. Statist. Data Anal.* **56**, 4229-4242.

Yang, Z., Harding, J. W. and Addyb, C. L. (2010) . Testing overdispersion in the zero inflated Poisson model. *J. Statist. Plann. Inference* **139**, 3340-3353.

Yau, K. K. W., Lee, A. H. and Carrivick, P. J. W. (2004). Modeling zero-inflated count series with application to occupational health. *Comput. Methods Programs Biomed.* **74**, 47-52.

Zhang, T. and Lin, G. (2009). Spatial scan statistics in loglinear models. *Comput. Statist. Data Anal.* **53**, 2851-2858.

Zhang, T., Zhang, Z. and Lin, G. (2012). Spatial scan statistics with overdispersion. *Statist. Medicine* **2**, 762-774.

Federal University of Amazonas, Manaus, Amazonas, Brazil.

E-mail: maxlima@ufam.edu.br

Federal University of Minas Gerais, Avenida Presidente Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, Brazil.

E-mail: duczmal@est.ufmg.br

Federal University of Amazonas, Manaus, Amazonas, Brazil.

E-mail: jcardoso@ufam.edu.br

Federal University of Minas Gerais, Avenida Presidente Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, Brazil.

E-mail: leticia@dcc.ufmg.br

(Received August 2013; accepted March 2014)