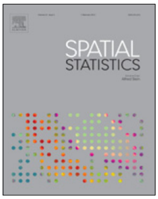
Spatial Statistics 18 (2016) 444–454

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



A spatial scan statistic for beta regression

Max Sousa de Lima [a,1], Vanessa Souza dos Santos [a,1],
Luiz H. Duczmal [b,*], Diego da Silva Souza [a,1]

[a] Department of Statistics, Universidade Federal do Amazonas, 69077-000, Manaus, AM, Brazil
[b] Statistics Department, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 5 April 2016
Accepted 6 September 2016
Available online 20 September 2016

Keywords:
Cluster detection
Spatial scan
Beta regression
Bootstrap *p*-value

ABSTRACT

Spatial Scan Statistics have been developed for geographical cluster detection in different types of models, for example, Bernoulli, multinomial, Poisson, Exponential, Weibull and Normal. However, some data are continuous in the interval $(0, 1)$ such as rates, proportions and indices, or are limited in the interval (a, b), $a < b$. In this paper, we propose a spatial scan statistic for a Beta regression model. The test statistic is based on a likelihood ratio test and evaluated using Bootstrap *p*-value. The proposed method is illustrated using index of basic education and infant mortality in the Brazilian Amazon. The statistical power, sensitivity and positive predicted value of the test are examined through a simulation study.

1. Introduction

The spatial distribution of rates and proportions received considerable attention in literature. In particular, spatial cluster detection is widely regarded as a powerful tool to search for possible agglomerations of events which are statistically significant (Murray et al., 2014). Those events may be individual cases of a certain disease, mortality rate from some known cause or local inequality index (e.g., social, economic, educational). When those events are aggregated by areas, a spatial or geographic cluster is represented by a localized group of areas with event rate (or proportions and index) well above (or well below) the expected rate, which constitutes an anomaly in the study

---

[*] Corresponding author.
E-mail address: duczmal@est.ufmg.br (L.H. Duczmal).

[1] Fax: +55 92 33052840.

http://dx.doi.org/10.1016/j.spasta.2016.09.003
2211-6753/

area. Some applications of spatial cluster detection are found in crime analysis, astronomy, biology and genetics (Glaz et al., 2009). Current methods of cluster detection usually employ hypotheses testing. The Spatial Scan (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) is the most popular method for the detection and inference of spatial clusters, both for discrete and continuous models, Bernoulli (Kulldorff and Nagarwalla, 1995), Poisson (Kulldorff, 1997), Multinomial (Jung et al., 2010), Exponential (Huang et al., 2007), Normal (Huang et al., 2010) and Weibull (Bhatt and Tiwari, 2014). Extensions of the spatial scan were devised to take into account spatial correlation (Meng and Zhu, 2015), covariate adjustment (Jung, 2009; Zhang et al., 2014), log-linear models (Zhang and Lin, 2009), multivariate data (Kulldorff et al., 2007; Neill et al., 2013), repeated events (Rosychuk and Chang, 2013), overdispersion and inflated zeros (Zhang et al., 2012; Cançado et al., 2014; Lima et al., 2015), multiple cluster (Wan et al., 2009), marked point processes (Cucala, 2014). Power comparisons and approximations of the Scan statistic are described in Kulldorff et al. (2003) and Read et al. (2013). Its relative risk bias was discussed in Prates et al. (2014). Also, Almeida et al. (2011) provided a correction for the statistical significance bias due to different cluster sizes.

The Spatial Scan is used to detect and test the statistical significance of the cluster, without the previous knowledge of its size and location, adjusted for the implicit multiple testing. In computational terminology, the method scans the study area imposing a window of arbitrary geometric shape (Duczmal and Assunção, 2004; Duczmal et al., 2006; Assuncao et al., 2006; Duczmal et al., 2007). The Circular Scan is a popular version of the Spatial Scan Statistic. It employs circularly shaped windows, centered on the centroids of each administrative area of the map. It is usual to form a collection $\mathcal{Z}$ of clusters whose populations never exceed 50% of the total population of the map. The Spatial Scan test statistic is evaluated for each cluster candidate in the collection $\mathcal{Z}$ to find the most likely cluster, which has the highest test value. A Monte Carlo procedure is usually employed to obtain the test p-value and compute the significance of the most likely cluster. For simplicity, the circular scan will be used in this work.

In some of the previously cited Spatial Scan probabilistic models, it is assumed that the random variable has unbounded support. However, there are situations where the variable of interest is continuous and limited in the real interval (a, b). For rates, proportions or index numbers, the interval $(0, 1)$ is a convenient support, for which the Beta distribution is adequate, with density function

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1 \tag{1}$$

where $p > 0, q > 0$ and $\Gamma(.)$ is the Gamma function. The distribution Y has average and variance, given respectively by

$$\mathbb{E}(Y) = \frac{p}{p+q}, \quad \text{and} \quad \mathbb{V}\text{ar}(Y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

The uniform distribution is a particular case of (1). It is usual to define $\mu = \frac{p}{p+q}$ and $\phi = p + q$ (see Ferrari and Cribari-Neto, 2004). Then $0 < \mu < 1$ and $\phi > 0$, resulting in $\mathbb{E}(Y) = \mu$ and $\mathbb{V}\text{ar}(Y) = V(\mu)/(1+\phi)$, where $V(\mu) = \mu(1-\mu)$. When ϕ is large, the variance of Y is small, and ϕ is interpreted as a parameter of precision or dispersion in the model. Using this new parametrization, a regression model was built elsewhere (Ferrari and Cribari-Neto, 2004), which is, in many aspects, similar to the class of generalized linear models, although the Beta distribution does not belong to this class. In this work, a Regression Beta Spatial Scan Statistic is developed, for data in the interval (a, b), where the expected value of the model is adjusted by covariates. A Newton–Raphson routine is used to estimate the parameters, the log-likelihood ratio is the test statistic, and the most likely cluster's significance is evaluated by bootstrap p-value. The method is illustrated using index of basic education and infant mortality in the municipalities of Amazonas state in Brazil.

The paper is organized as follows. In Section 2, the Spatial Scan Statistic for Beta Regression is defined. Numerical studies are presented in Section 3, and a real data application is illustrated in Section 4. Finally, Section 5 presents the concluding remarks.