Journal of Multivariate Analysis 122 (2013) 393–408

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Hierarchical Poisson models for spatial count data

Victor De Oliveira

Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX 78249, USA

ARTICLE INFO

Article history:
Received 11 January 2013
Available online 23 August 2013

AMS subject classifications:
60G10
60G60
62M30

Keywords:
Copula
Fréchet–Hoeffding upper bound
Gaussian random field
Generalized linear mixed model
Geostatistics
Poisson–Gamma model
Poisson–Lognormal model

ABSTRACT

This work proposes a class of hierarchical models for geostatistical count data that includes the model proposed by Diggle et al. (1998) [13] as a particular case. For this class of models the main second-order properties of the count variables are derived, and three models within this class are studied in some detail. It is shown that for this class of models there is a close connection between the correlation structure of the counts and their overdispersions, and this property can be used to explore the flexibility of the correlation structures of these models. It is suggested that the models in this class may not be adequate to represent data consisting mostly of small counts with substantial spatial correlation. Three geostatistical count datasets are used to illustrate these issues and suggest how the results might be used to guide the selection of a model within this class.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Spatial *count* data are routinely collected in many earth and social sciences, such as ecology, epidemiology, demography and geography. For instance, death counts due to different causes are collected on a regular basis by government agencies throughout the entire US and classified according to different demographic variables, such as age, gender and race. Among the most common goals for the analysis of this kind of data are determining the effects on mortality of spatially varying risk factors (regression problems) and estimation of unobserved spatially varying quantities of interest (prediction problems). In this work I consider models for geostatistical count data.

Unlike for *continuous* data, few models have been proposed in the literature for geostatistical count data. This is due, in part, to the scarcity of families of multivariate discrete distributions, and the fact that none of the available ones have the flexibility and mathematical tractability comparable to that of some of the families of multivariate continuous distributions (such as the Gaussian family of distributions). This scarcity is reflected by the fact that even the most influential spatial statistics textbooks either lack a treatment of models for geostatistical count data or have a very scant one, with the book by Diggle and Ribeiro [12] as a notable exception. Early works that analyze geostatistical count data include Gotway and Stroup [16] and McShane et al. [25], who proposed using generalized linear models and generalized estimating equations. But the statistical basis and validity of these approaches to model geostatistical data are somewhat questionable. In addition, prediction methodology in these works is either lacking or ad-hoc, with no measures of prediction uncertainty.

Most models of current use for geostatistical count data use Gaussian random fields as building blocks. The prime example is the hierarchical model proposed by Diggle et al. [13], which can be viewed as a generalized linear mixed model. This is also known as the Poisson–Lognormal model, which was initially proposed for the analysis of correlated count data by

E-mail address: victor.deoliveira@utsa.edu.

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.08.015

Aitchison and Ho [1], and used for applications in non-spatial contexts by Chan and Ledolter [4], Chib and Winkelmann [6] and Hay and Pettitt [20]. Applications of this model for the analysis of geostatistical count data appeared in Diggle et al. [13], Christensen and Waagepetersen [8], Zhang [33], Royle and Wikle [29], Christensen et al. [7], Guillot et al. [19] and Eidsvik et al. [14]. Extensions to model geostatistical zero inflated count data and multivariate count data were given, respectively, by Recta et al. [28] and Chagneau et al. [3]. Fitting this kind of hierarchical geostatistical models is a challenging task requiring numerical methods, such as EM or MCMC algorithms, and research efforts to date have almost entirely focused on fitting and computation.

As is apparent from the above, the aforementioned Poisson–Lognormal model has generated a fair amount of attention and research, and currently seems to be (arguably) the 'state-of-the-art' for modeling geostatistical count data. Nevertheless, some of the basic properties of this model are not well understood yet, and in fact the study of its main second-order properties and its adequacy to describe a variety of geostatistical count datasets have been somewhat neglected, with the work by Madsen and Dalthorp [24] as a notable exception. One salient property of this model is that the mean function of the counts may exert a substantial influence on their correlation function. This makes the model to have two features that may be undesirable for the modeling of some datasets. First, the regression parameters might be difficult to interpret and estimate because of their influence on two different aspects of the model. Second, the regression parameters induce a 'whitening effect' on the correlation structure of the counts, which in some cases renders the model unable to represent substantial spatial correlation present in some datasets. This effect is specially severe when the data consist mostly of small counts, which is precisely the case when a model accurately describing the discreteness of the data is most needed (as opposed to a model with continuous distributions intended to approximate discrete data).

In this work I propose a class of hierarchical models for geostatistical count data that includes the Poisson–Lognormal model as a particular case. The main second-order properties of the models within this class are derived, both for the latent random field in the second level of the hierarchy and for the observable counts. It is shown that for this class of models the correlation structure of the counts strongly depends on their overdispersions, which in turn are determined by the mean–variance relationship of the marginal distributions of the latent random field. An explicit expression is obtained for the correlation function of the counts in terms of their overdispersions and the correlation function of the latent random field. When there is no explicit expression for the correlation function of the latent random field, a series expansion is used to numerically investigate the second-order properties of these models. These properties should help researchers and practitioners judge the possible adequacy of any of these models to describe a particular dataset, and point to the importance of estimating overdispersion in geostatistical count data.

The above general results are used to study in some detail the second-order properties of the Poisson–Lognormal model and two Poisson–Gamma models. It is shown that one of the Poisson–Gamma models has similar second-order properties as the Poisson–Lognormal model, but the properties of the other Poisson–Gamma model are different. The findings shed light into the scope and limitations of this class of hierarchical models, and suggest that this class of models may not be adequate to describe datasets that consist mostly of small counts and display substantial spatial correlation. It is argued that these second-order properties may be used to guide the selection of a model within the proposed class and informally assess its adequacy to describe spatial count data. Some of the issues raised here are illustrated using three geostatistical count datasets that were previously analyzed in the literature.

2. A class of models for geostatistical count data

I describe here a class of models for the variation of spatial count data that generalizes the model proposed by Diggle et al. [13], and study their main second-order properties. Let $\{\Lambda(\mathbf{s}) : \mathbf{s} \in D\}$, with $D \subset \mathbb{R}^d$ and $d \geq 1$, be a *positive* random field describing the spatial variation of a quantity of interest over the domain D, usually a spatially varying intensity or risk, whose values are *not* observable. To learn about this random field, spatial information is collected on random variables $Y_1, \ldots, Y_n$ that take nonnegative integer values and are stochastically related to $\Lambda(\cdot)$. Two examples illustrate this situation. In the Bjertorp Farm dataset analyzed by Guillot et al. [19], $\Lambda(\mathbf{s})$ represents weed intensity at location $\mathbf{s}$ and Y_i is the number of weeds observed within a rectangular frame of area t_i centered at sampling location $\mathbf{s}_i$. In the Rongelap Island dataset analyzed by Diggle et al. [13], $\Lambda(\mathbf{s})$ represents the level of radio-nuclide Caesium (^{137}Cs) at location $\mathbf{s}$ and Y_i is the number of photon emissions collected at a sampling location $\mathbf{s}_i$ by a gamma-ray counter during a period of time t_i. In these examples, for each of a set of sampling locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ within D, count measurements Y_i are taken, together possibly with measurements of location-dependent covariates. The main goal in both examples is the prediction of $\Lambda(\cdot)$ throughout D based on the data $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ and the covariate information, if available, but answering regression questions might be a secondary goal; see Section 5 for further details about these datasets.

The aim is for a class of models constructed to possess the following properties:

(a) The count variables Y_i have overdispersed marginal distributions, a property found to hold in many spatial (and non-spatial) count datasets.
(b) The marginal distributions of the random field $\{\Lambda(\mathbf{s}) : \mathbf{s} \in D\}$ are given by a conjectured parametric family of continuous cdfs $\mathcal{G} = \{G_{\mathbf{s}}(\cdot) : \mathbf{s} \in D\}$, where each cdf has support $[0, \infty)$.

The proposed class of models for the random vector $\mathbf{Y}$ and random field $\{\Lambda(\mathbf{s}) : \mathbf{s} \in D\}$ that satisfies the aforementioned properties is defined in terms of their family of finite-dimensional distributions. It is hierarchically specified as follows:

Download English Version:

https://daneshyari.com/en/article/1145816

Download Persian Version:

https://daneshyari.com/article/1145816

Daneshyari.com