Spatial Statistics 7 (2014) 1–20

Contents lists available at ScienceDirect

Spatial Statistics

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

Poisson kriging: A closer investigation

Victor De Oliveira *

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

A R T I C L E   I N F O

Article history:
Received 15 July 2013
Accepted 6 November 2013
Available online 12 November 2013

Keywords:
Geostatistics
Nugget effect
Ordinary kriging
Semivariogram
Spatial count data

A B S T R A C T

This work revisits a simple geostatistical model for the analysis of spatial count data and describes some of its main second-order properties. This geostatistical model is simpler than an alternative hierarchical model, also used for the analysis of spatial count data, so it may be more appealing to practitioners and spatial data analysts. Geostatistical methods for trend parameter estimation, semivariogram estimation and prediction of the latent process are reviewed, and new estimators and predictors are proposed. Finally, a designed simulation experiment is carried out to investigate and compare the sampling properties of the different estimators and predictors.

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 U.S. 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 a model for *geostatistical* count data.

Early attempts to model geostatistical count data include Gotway and Stroup (1997) and McShane et al. (1997) who proposed the use of, respectively, generalized linear models and generalized estimating equations. But the statistical basis of these approaches to model geostatistical count data is somewhat questionable. In addition, prediction methodology in these works is either lacking or ad-hoc, with no measures of prediction uncertainty. As an alternative, many models of current use

* Tel.: +1 210 4586592.
E-mail address: victor.deoliveira@utsa.edu.

http://dx.doi.org/10.1016/j.spasta.2013.11.001

for geostatistical count data use Gaussian random fields as building blocks. The prime example is the hierarchical model proposed by Diggle et al. (1998), which can be viewed as a generalized linear mixed model; see Christensen and Waagepetersen (2002), Royle and Wikle (2005), and Guillot et al. (2009) for applications of this model, and De Oliveira (2013) for a study of its properties. Although currently this model seems to be (arguably) the 'state-of-the-art' for modeling geostatistical count data, fitting this kind of hierarchical model is a challenging task requiring computationally intensive numerical methods, such as the EM or MCMC algorithms. This complexity is likely to preclude the use of this model by most practitioners and spatial data analysts, so it is desirable to also have alternative simpler models that can be fitted using traditional geostatistical methods. This work studies one such model.

In this work I consider a model for geostatistical count data proposed by Monestiez et al. (2006). A similar model was previously proposed by Zeger (1988) for the analysis of time series of count data and later adapted by McShane et al. (1997) for the analysis of spatial count data. The main goal in these works was to perform regression analysis, i.e., to assess the effect of covariates on the mean response. Monestiez et al. (2006) proposed a geostatistical approach for inference about a latent (unobserved) process that 'drives' the observed count data, for the case when this latent process has constant mean function and the data collection may involve unequal sampling efforts. Later, Bellier et al. (2010) extended the methodology for the case when the latent process has a non-constant mean function, but the sampling efforts are all equal. These works adapted traditional geostatistical methods for semivariogram estimation and optimal linear unbiased prediction (kriging) to deal with the more challenging context of count data: the latent process of interest is not observed and the observed data are non-stationary due to heteroscedasticity. The methodology has been applied and extended, among others, by Goovaerts (2005, 2006) and Krivoruchko et al. (2011a,b), where the latter used the software ArcGIS Geostatistical Analyst.

Although several applications of the aforementioned model have appeared in the literature, the study of the model properties as well as properties of the parameter estimators and latent process predictors have, to the best of my knowledge, not been carried out. In this work I revisit the aforementioned model and undertake a detailed study of its properties and properties of the statistical methods used to make inference about the model. First, the assumptions under which the model is constructed are made explicit and some of its main second-order properties are derived. Second, methods for estimation of the model components (trend and semivariogram) are described in some detail. I propose estimating the trend parameters using one of two variants of maximum pseudo likelihood. Also, the semivariogram estimator proposed by Monestiez et al. (2006) is reviewed, and two other estimators are proposed. Third, the method proposed by Monestiez et al. (2006) to predict the latent process is reviewed, and an alternative predictor is proposed that parallels simple kriging. Finally, the main sampling properties of the different trend parameter estimators, semivariogram estimators and latent process predictors are investigated using a designed simulation experiment. The findings of this investigation are summarized in the last section.

2. A model for geostatistical count data

Let $\{\Lambda(\mathbf{s}) : \mathbf{s} \in D\}$, with $D \subset \mathbb{R}^2$, 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 non-negative integer values and whose mean values are related to $\Lambda(\cdot)$. Three examples illustrate this situation. In the Rongelap Island dataset analyzed by Diggle et al. (1998), $\Lambda(\mathbf{s})$ represents the level of the radionuclide Cesium (^{137}Cs) at location $\mathbf{s}$ and Y_i is the number of photon emissions collected at the sampling location $\mathbf{s}_i$ during a period of time t_i by a gamma-ray counter. In the Bjertorp Farm dataset analyzed by Guillot et al. (2009), $\Lambda(\mathbf{s})$ represents the weed intensity at location $\mathbf{s}$ and Y_i is the number of weeds observed within a rectangular frame centered at the sampling location $\mathbf{s}_i$ and having area t_i. Finally, in the New England cancer mortality dataset analyzed by Goovaerts (2005), $\Lambda(\mathbf{s})$ represents the risk of breast cancer for white females at location