STATISTICAL ANALYSIS OF SPATIAL COUNT DATA

Mark S. Kaiser
Department of Statistics, Iowa State University, USA

Keywords: disease mapping, epidemiology, hierarchical models, spatial quadrats, Markov random fields

Contents

Summary

Data in the form of counts that are also geo-referenced may arise in a wide variety of situations. This chapter discusses a number of spatial processes that may be used to model such data. The fundamental categorization used here is to consider processes as having either random or fixed spatial indices. The discussion focuses on the types of models that are commonly used in the analysis of spatial count data, and a number of issues that are relevant for those modeling approaches. Relatively little coverage is given to statistical techniques for estimation and inferential procedures, other than those that are intimately connected with the objectives of model formulation. This emphasis on modeling issues, at the expense of technical considerations in estimation and inference is purposeful. It is not possible, or at least not advisable, to discuss details of how a model for spatial counts might be analyzed without first considering how various models arise in the first place. Model formulation and specification are closely connected with the scientific phenomena that are the focus of investigation, that is, the question of why spatial counts are to be analyzed.

Spatial counts may result from the aggregation of event data, such as occurs in the analysis of point processes using quadrat summaries, or the occurrence of a relevant disease within small geographic units or areas. Alternatively, counts may result from observation of a set of discrete random variables, each of which has possible values in

the set of non-negative integers. Many models for such situations are formulated for continuous Gaussian random variables, and a common analysis of counts is to first perform a transformation that renders models for Gaussian random variables more appropriate than they would be for raw counts. While this approach is still popular in the analysis of counts, statistical advances over the past 10 years have made models for discrete, integer-valued random variables more accessible to the scientific community. These include Markov random field models and hierarchical models for conditionally independent Poisson data with spatially dependent latent processes. Although the analysis of these models may require computationally-intensive statistical techniques such as Monte Carlo maximum likelihood and Markov Chain Monte Carlo, their use allows inference and prediction to be made directly on the scale of interest, namely on the scale of spatial counts.

1. Introduction

Spatial count data may be considered to be any collection of observations with possible values in the set of non-negative integers {0,1…} and for which each observation is geo-referenced, that is, corresponds to a particular spatial location. Such data may arise in a number of ways, either from direct observation of quantities with integer values or as the result of the aggregation of observed quantities which represent presence or absence of some characteristic. A convenient way to organize the possible scenarios leading to spatial count data is to consider situations in which the spatial locations corresponding to observed quantities are determined by where the objects of such observation happen to occur, and situations in which the spatial locations are fixed *a priori.* Call these two categories situations with random spatial indices and non-random spatial indices, respectively.

A random field representation for spatial processes will be used throughout this chapter. Let $\boldsymbol{s}$ denote a spatial location variable. For example, $\boldsymbol{s}$ might represent the latitude u and longitude v of any point within a given geographic region D. This presentation will consider primarily two-dimensional spatial locations so that $D \subset \Re^2$, but most of what is contained here generalizes to the case in which D is a subset of the *d*-dimensional real numbers. A *spatial process* is represented as,

$$\mathbf{Z} \equiv \{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}, \qquad (1)$$

where $Z(\boldsymbol{s})$ may be a random variable at the location $\boldsymbol{s}$, and D may be a random set on which $Z(\cdot)$ is defined. This chapter will consider only univariate $Z(\cdot)$, but the definition of equation (1) is easily extended to deal with vector-valued random variable. In the case of random spatial indices, D in (1) is taken to be a random set (i.e., D may vary among realizations of the process $\mathbf{Z}$) and $Z(\cdot)$ is either non-random (with fixed value 1) or as also random. In the case of non-random spatial indices D is considered a fixed set and $Z(\cdot)$ a random variable.

2. Random Spatial Indices

Spatial count data may arise from the analysis of *spatial point patterns,* which is concerned with the spatial pattern generated by observing locations at which a particular event occurs, such as the locations at which a particular species of plant is found. Consider a situation in which objects exist at a finite set of locations and fail to exist at all other locations, which may be formulated as in (1) with $Z(\boldsymbol{s}) \equiv 1$ and D consisting of a random *point process,* that is, a collection of points at which (the now non-random) $\mathbf{Z}$ is defined. This formulation allows a unified framework within which to extend the basic definition of point processes to include *marked* processes in which both the set of locations of objects D and an associated value (or 'mark') $Z(\cdot)$ are random, although that possibility will not be considered in the sequel. Spatial point processes lead to data represented as counts from an aggregation process in which the number of events occurring in a set of sub-regions is tallied, but the specific locations may not be. The analysis of spatial point processes is covered elsewhere in this encyclopedia and will not be discussed here, but the following is included to indicate the way that point processes lead to data in the form of spatial counts.

Consider a simple point process in which $Z(\boldsymbol{s}) \equiv 1$ for all $\boldsymbol{s} \in D$, with D random. Suppose that a given study area A has been divided into a set of sub-regions or *quadrats* $\boldsymbol{Q} \equiv \{Q_k : k = 1,...,K\}$. For $D \in A$, a description of the spatial pattern of events (or occurrences) results from counting the number of events in each quadrat. Formally, the equivalence of information provided by knowing D and knowing the number of events in each element of $\boldsymbol{Q}$ requires that we be able to count the number of events in *every* possible division of A into quadrats (technically, all bounded sets contained in the Borel σ-algebra of the area $A \subset \Re^2$). The equivalence of information provided by knowing the location of each point in D and knowing counts in collections of quadrats leads to the comparison of counts with what would be expected under a model or 'complete spatial randomness', that is , a model in which quadrat counts are realizations of a homogeneous Poisson process. This comparison underlies some of the common indices of spatial pattern, such as David and Moore's 'index of clumping', Douglas' 'index of cluster frequency', and the 'patchiness' index of Morisita. The use of such techniques will not be discussed further here as they do not truly constitute methods for the spatial analysis of count data, although they are valuable exploratory tools in the description of point pattern in spatial settings.

Aside from the production of various indices of spatial pattern, methodologies for the analysis of spatial counts that have originated from the analysis of spatial point patterns involve fitting models of various stochastic processes to data in the form of quadrat counts. The stochastic process models considered are typically models for point processes rather than models that describe the distributions of counts directly. The connection between a stochastic model of a point process and spatial count is made on the basis of the "intensity function" of the stochastic process under consideration. The intensity of a process may be defined as λ, where, for a bounded area B, E (y points in B) $= \lambda$ (area of B). The better-known stochastic process models include the clustering processes of Neyman-Scott and Cox, and processes to describe spatial inhibition such as the Strauss process. The Neyman-Scott process is the best-known example of a larger class of models known as Poisson cluster processes.

3. Non-random Spatial Indices

Spatial count data arise naturally from the observations, at a set of fixed spatial locations, of quantities that have possible values in the set of non-negative integers. In this setting, the spatial domain D of the general spatial process (1) is taken to be a fixed set, and the random variable $Z(\boldsymbol{s})$ to have possible values {0,1…}. Even with non-random spatial indices we distinguish between two possibilities, those being continuous spatial indices and discrete spatial indices. Models used with problems conceptualized as having continuous spatial index are typically categorized as *geostatistical* models, while models with discrete spatial indices are commonly called *lattice* or *Markov random field* models. The distinction lies in the range of possible values for the spatial index $\boldsymbol{s}$ in (1). In the case of a continuous index we have that $\boldsymbol{s}$ can vary continuously within D, while for a discrete index $\boldsymbol{s} \equiv \{\boldsymbol{s}_1, \boldsymbol{s}_2, ...\} \subset D$. It is possible to simply define D as a set of discrete elements in the first place, but the notation used here may be preferred since unobserved locations are often of interest. With either continuous or discrete index cases, interest focuses on describing the probability distribution of the random field $\mathbf{Z}$.

3.1 Models with Continuous Spatial Index

While models having continuous spatial index are conceptualized as the random field $\boldsymbol{Z} \equiv \{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ with continuously varying $\boldsymbol{s}$, any application will depend on using observations at a finite number of locations $\{Z(\boldsymbol{s}_i) : i = 1,...,n\}$. Theoretical model quantities are defined in terms of the process $\mathbf{Z}$, but estimation of those quantities must be accomplished using the observed values of $\{\mathrm{Z}(\boldsymbol{s}_i) : i = 1,...,n\}$. Such models are typically called *geostatistical* models, since their origin was in the geological sciences.
Underlying the geostatistical approach is an implicit model of the form

$$Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \delta(\boldsymbol{s}), \quad \boldsymbol{s} \in D, \tag{2}$$

Where $\mu(\boldsymbol{s}) = E\{Z(\boldsymbol{s})\}$ and $\delta(\boldsymbol{s})$ is a zero-mean spatial process. Various assumptions are made about the behavior of the mean process $\{\mu(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ and the spatially dependent error process $\{\delta(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ to facilitate estimation of the mean at any location and prediction of the observable random variable $Z(\boldsymbol{s}_0)$ at any location s_0 that is not included in the set of observed locations. Prediction, in particular, is conducted using a set of techniques known as *kriging.* A kriging predictor is a linear function of the data,

$$p(Z(\boldsymbol{s}_0)) = \sum_{i}^{n} \lambda_i Z(\boldsymbol{s}_i), \tag{3}$$

where the coefficients $\{\lambda_i : i = 1,...,n\}$ are chosen so that the predictor is optimal in the sense that it minimizes the mean squared prediction error among all linear predictors. Restrictions are often placed on the coefficients of (5) to insure that the predictor is also unbiased, that is, $E\{p(Z(\boldsymbol{s}_0))\} = \mathbf{Z}(\boldsymbol{s}_0)$.

Geostatistical analysis and the construction of kriging predictors may be considered ‘distribution free’ in the same sense that ordinary least squares estimators of regression parameters in simple linear regression are free of distributional assumptions. But, just as in the regression case, an implicit near-Gaussian property is invoked for the purposes of inference and, in particular, to form prediction intervals. Indeed, the entire concept of restricting attention to linear estimators and predictors has a connection with a Gaussian distribution for which the conditional mean

$E[Z(s_0)|\{Z(s_i):i=1,\ldots,n\}]$ is linear in the values $\{Z(s_i):i=1,\ldots,n\}$. In this case the optimal (minimum mean squared error) linear predictor is also the optimal predictor; this will not be true for non-Gaussian processes.

The connection of standard geostatistical analysis to spatial processes that are Gaussian has led to the practice of transforming spatial count data to produce Gaussian-like behavior on the transformed scale. Similar to transformations used in linear regression analysis the focus is often on producing homogeneity of variance, with symmetry in distribution frequently occurring as a concomitant effect. Perhaps the most common transformation of count data is some form of square root transformation, although logarithmic transformations are also frequently used. These transformations have been employed, for example, in the analysis of cancer mortalities in Scotland and in the analysis of data on sudden infant death syndrome.

-
-
-

Bibliography

Clayton, D. and Kaldor, J. (1987), Empirical Bayes estimates of age-Standardized relative risks for use in disease mapping, *Biometrics* **43,** 671-681. [This article makes use of a hierarchical Poisson-log Gaussian model to analyze lip cancer in Scotland].

Cressie, N.A.C. (1993), *Statistics for Spatial Data,* revised ed., New York: Wiley. [This work contains a complete coverage of spatial modeling and data analysis up to the mid 1990’s including many of the topics covered in this article].

Diggle, P.J. (1990), A point process modeling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point, *Journal of the Royal Statistical Society* **153,** 349-362. [This discusses several issues related to spatial epidemiology].

Griffith, D.A. and Layne, L.J. (1990), *A Casebook for Spatial Statistical Data Analysis,* New York: Oxford University Press. [This presents a large number of examples of spatial problems and data analysis emphasizing geostatistical and conditional autoregressive techniques.].

Haining, R. (1990), *Spatial Data Analysis in the Social and Environmental Sciences,* Cambridge: Cambridge University Press. [This work focuses on overall modeling strategies for spatial problems and presents many examples from the social and environmental sciences].

`Kaiser, M.S. and Cressie, N. (2000), The construction of multivariate distributions from Markov random fields, *Journal of Multivariate Analysis* **73,** 199-220. [This extends the Markov random field approach to model specification, including multiparameter exponential families].

Lawson, A.B. (2001), *Statistical Methods in Spatial Epidemiology*, Chichester: Wiley. [This presents models appropriate for both point-event and regional count data].

Marshall, R.J. (1991), A review of methods for statistical analysis of spatial patterns of disease, *Journal of the Royal Statistical Society* A, **154,** 421-441. [This is a review article of approaches toward disease mapping].

Muirhead, C. and Darby, S. (eds)(1989), Cancer near nuclear installations, *Journal of the Royal Statistical Society A* **152,** 305-381. [This presents a collection of papers devoted to the issue of spatial epidemiology and disease mapping].

Stern, H.S. and Cressie, N. (1999), Inference for extremes in disease mapping, In Lawson, A., Biggeri, A., Bohning, D., Lesaffre, E., Viel, J-F. and Bertonllini, R. (eds), *Disease Mapping and Risk Assessment for Public Health*, Wiley, Chichester, pp. 63-84. [This discusses issues in the development of models for disease mapping and uses a Gaussian-Gaussian hierarchical model].

Biographical Sketch

Mark Kaiser is an Associate Professor of Statistics at Iowa State University, Ames, Iowa, USA. His interest in the development of statistical methods appropriate for the analysis of environmental and ecological studies stems from early graduate work in these fields. Current research efforts focus on Markov random field models, dynamic models, and hierarchical models that contain complex dependence structures. Among his primary areas of application are the analysis of water quality data, the assessment of information from marine fisheries, and spatio-temporal modeling of environmental processes.