Spatial and Spatio-temporal Epidemiology 17 (2016) 59–74

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Original Research

Stepwise and stagewise approaches for spatial cluster detection

Jiale Xu[a], Ronald E. Gangnon[b,*]

[a] *Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, United States*
[b] *Department of Biostatistics and Medical Informatics and Department of Population Health Sciences, University of Wisconsin-Madison, Madison,*
WI 53726, United States

ARTICLE INFO

Article history:
Received 8 April 2015
Revised 5 April 2016
Accepted 12 April 2016
Available online 3 May 2016

Keywords:
Bias adjustment
Cluster detection
Permutation test
Spatial scan statistic
Stagewise
Stepwise

ABSTRACT

Spatial cluster detection is an important tool in many areas such as sociology, botany and public health. Previous work has mostly taken either a hypothesis testing framework or a Bayesian framework. In this paper, we propose a few approaches under a frequentist variable selection framework for spatial cluster detection. The forward stepwise methods search for multiple clusters by iteratively adding currently most likely cluster while adjusting for the effects of previously identified clusters. The stagewise methods also consist of a series of steps, but with a tiny step size in each iteration. We study the features and performances of our proposed methods using simulations on idealized grids or real geographic areas. From the simulations, we compare the performance of the proposed methods in terms of estimation accuracy and power. These methods are applied to the the well-known New York leukemia data as well as Indiana poverty data.

1. Introduction

Spatial cluster detection is a fundamental and challenging problem in spatial epidemiology. The term ‘clustering’ is a vaguely defined concept in the medical literature. A broad definition of clustering is the spatial aggregation of disease events. As the observed spatial pattern may simply be a function of distribution of the population at risk or of some other risk factors, Wakefield et al. (2000) proposed a more robust definition, which describes clustering as residual spatial variation in risk after accounting for known influences. The main goal of disease clustering is to evaluate whether a disease is randomly distributed or has a tendency to cluster over time or space after adjusting for known confounding factors. The identification of clusters may provide clues when studying the etiology of a disease, or when conducting disease surveillance programmes. On the one hand, false identification of a cluster may lead to wasted resources, but on the other hand, failing to detect a genuine disease cluster may cause serious consequences. For instance, underestimation of spatial extent and severity of an infectious disease may discourage necessary public concern and lead to wider spread of disease.

Spatial cluster detection problems have been typically approached under a frequentist hypothesis testing framework. The spatial scan statistic method (Kulldorff, 1997; Kulldorff and Nagarwalla, 1995) and its many variants (Kulldorff et al., 2006; Shu et al., 2012; Tango and Takahashi, 2005) are based on the simultaneous evaluation, via Monte Carlo hypothesis testing, of the statistical significance of the maximum likelihood ratio test statistic across a large collection of potential clusters. The scan statistic approach is typically based on the comparison of a no

[*] Corresponding author. Tel.: +16082650688.
E-mail addresses: zhjxujiale@gmail.com (J. Xu), ronald@biostat.wisc.edu (R.E. Gangnon).

http://dx.doi.org/10.1016/j.sste.2016.04.007
1877-5845/© 2016 Elsevier Ltd. All rights reserved.

clustering null hypothesis against a single cluster alternative. Development of scan statistics has focused on assessment of the no cluster null hypothesis against the single cluster alternative with ad hoc assessments of secondary clusters. Some recent methods more rigorously account for multiple clusters in the detection process. Zhang et al. (2010) propose assessing secondary clusters after sequential deletion of observed data inside the previously detected clusters, essentially a variant of more traditional forward stepwise variable selection. Li et al. (2011) propose a modified scan statistic that evaluates the most likely two (or more) clusters rather than the single most likely cluster. Beyond the requirement of pre-specifying the number of clusters to be evaluated, this approach also greatly increases the size of the search space and hence the computational burden.

As an alternative, a number of authors Gangnon and Clayton (2000, 2003, 2007), Clark and Lawson (2002), Yan and Clayton (2006), and Wakefield and Kim (2013) have developed Bayesian models for cluster detection. All of these methods utilize essentially the same Poisson or binomial likelihood function, which incorporates explicit clusters with distinctive, either elevated or lowered, risks. All of these methods require prior specifications for the number of clusters and for the risk parameters associated with the background and the clusters. The major substantive differences between these methods are differences in prior specifications for these parameters, which also lead to differences in computation. Here, we consider penalized likelihood approaches based on forward stepwise and forward stagewise (Hastie et al., 2007, 2001) algorithms, which do not require prior specifications for these parameters, as an alternative approach to inference for multiple clusters.

In this paper, we develop two alternative approaches to detection of multiple clusters. First, we consider two novel approaches based on traditional forward stepwise selection. In contrast with Zhang et al. (2010), we retain all observations in the original dataset and instead absorb the effects of previously detected clusters into the offset term for the binomial or Poisson model. In addition to sequential hypothesis tests, we consider penalized likelihood approaches using either bootstrap bias corrections or traditional information criteria. Second, we recognize spatial cluster detection as a special case of high-dimensional variable selection in generalized linear models and propose the use of incremental forward stagewise regression (Hastie et al., 2007), a variation of the LASSO. We evaluate a number of different optimality criteria, including bootstrap-based bias corrections and traditional information criteria, to select a single model from the solution path.

The paper is organized as follows. In Section 2, we describe the spatial cluster models for Poisson and binomial data. In Section 3, we propose a stepwise method based on sequential permutation test, a modified stepwise method based on penalized likelihood, as well as a forward stagewise procedure. In Section 4, we conduct simulation studies. In Section 5, we present analysis of the New York leukemia data set and the Indiana Poverty data set. In Section 6, we present some concluding remarks.

2. Statistical models

The spatial data in disease clustering studies usually fall into two categories: point location (case-control) data and aggregated (cell count) data. Point location data contains the exact location of each study subject. In spatial epidemiology, the process of aggregation involves summing up counts of disease events within a defined area (or cell) to yield the total number of disease cases in each area. For confidentiality reasons, a majority of disease clustering studies use cell count data. With cell count data, an entire study region is divided into N cells. For each cell i, we observe y_i, the number of cases, $\mathbf{z}_i = (z_{1i}, z_{2i})$, the vector of co-ordinates of the geographic centroid, and n_i, the population at risk in cell i. We consider two probabilistic models for count data: a Poisson model and a binomial model.

2.1. Binomial model

Typically, the underlying statistical model assumes that the observed number of cases y_i, $i = 1, 2, \ldots, N$, are independently and identically distributed as

$$y_i \sim \text{binomial}(n_i, p_i), \tag{1}$$

where the unknown parameter p_i is the probability of the events for cell i and is modeled as

$$\text{logit}(p_i) = \text{logit}(p_{i0}) + \alpha + \sum_{j=1}^{m} \theta_j \mathbb{1}_{\{d(\mathbf{z}_i,\mathbf{c}_j)\le r_j\}}. \tag{2}$$

The non-spatial effect components include the intercept α and logit(p_{i0}), where p_{i0} is the baseline probability and can be estimated by a logistic regression model with some predictor variables such as demographic variables (race, ethnicity, gender, age, and etc.), or other non-spatial effect factors. The spatial clustering component of the model is $\sum_{j=1}^{m} \theta_j \mathbb{1}_{\{d(\mathbf{z}_i,\mathbf{c}_j)\le r_j\}}$, where p is the number of potential clusters, $\mathbf{c}_j$, r_j are the center and radius of potential circular cluster j (in metric d) associated with log odds ratio θ_j, $j = 1, 2, \ldots, p$, and $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

2.2. Poisson model

For a rare disease, we can approximate y_i, $i = 1, 2, \ldots, N$ by the Poisson distribution

$$y_i \sim \text{Poisson}(\rho_i E_i), \tag{3}$$

where the parameter ρ_i is the relative risk for cell i and E_i is the expected number of cases in cell i (based on internal or external standardization). When a confounding variable is of concern, let n_{il} be the population at risk in cell i with covariate value l and λ_l be the disease rate for people with covariate value l, the standardized expected number of cases in cell i is calculated as $E_i = \sum_l \lambda_l n_{il}$, where λ_l can be estimated internally or externally. A log-linear model for the relative risk ρ_i is modeled as

$$\log(\rho_i) = \alpha + \sum_{j=1}^{m} \theta_j \mathbb{1}_{\{d(\mathbf{z}_i,\mathbf{c}_j)\le r_j\}}, \tag{4}$$

where α is the background component which is related to the overall rate across the study area and is well-identified