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/

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 $\text{logit}(p_{i0})$, 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

by the data, $\sum_{j=1}^{m} \theta_j \mathbb{1}_{\{d(\mathbf{z}_i, \mathbf{c}_j) \le r_j\}}$ is the spatial clustering component, where θ_j is the log relative risk associated with potential cluster j.

2.3. Potential clusters

In cluster detection, we consider a large collection of subsets of the study region as potential clusters. When applying hypothesis testing methods or the model-based approaches, a natural choice of window shape is the circle, as it is the most compact shape that can be obtained. To make the discussion more concrete, we consider a collection of potential circular clusters centered at the cell centroids as potential clusters. The radii of the circles varies continuously from zero up to a user-specified maximum radius, r_{max}. For a particular cell, say cell i, the potential clusters centered at its centroid are chosen as follows. Let $0 = d_{i,1} < d_{i,2} < \cdots < d_{i,m_i} \le r_{max}$ be the unique ordered distances from the centroid of cell i to the centroids of all cells, truncated at r_{max}. Then the distinct potential clusters centered at cell i are circles of radii $d_{i,1}, d_{i,2}, \ldots, d_{i,m_i}$. The number of potential clusters is $m = \sum_{i=1}^{N} m_i$. Kulldorff et al. (2006) and Tango and Takahashi (2005) discussed the scan statistics methods using other scanning window shapes such as ellipses, squares, triangles or even more flexible shapes. Although we use a set of circular potential clusters, we note that our methods can be easily adapted to any discrete set of potential clusters.

The number of clusters k may be treated either as a parameter to be estimated or as a fixed constant. Some Bayesian approaches (Gangnon and Clayton, 2003; Lawson, 2000) implement a reversible jump Markov Chain Monte Carlo (RJMCMC) algorithm to account the varying numbers of clusters. Several other models (Gangnon, 2006; Gangnon and Clayton, 2007) define k as a user-specified constant, which is usually chosen as a upper bound on the true number of clusters, say k_0. If k is greater than k_0, the underlying model is correct, albeit possibly overparameterized. Under the variable selection framework, we aim to select a parsimonious set of non-zero components such that $\theta_{k_0+1}, \ldots, \theta_m \approx 0$.

3. Methods

We introduce two novel methods for spatial cluster detection: forward stepwise method and forward stagewise method. Both methods involve sequential updates starting from a null model of no clustering effects. Various stopping criteria are applied to select the optimal solution from the solution path.

3.1. Forward stepwise method

The standard spatial scan statistic method is based on the evaluation, via Monte Carlo hypothesis testing, of the statistical significance of the maximum likelihood ratio test for a large collection of potential clusters. The potential cluster with the maximum likelihood ratio is called *most likely cluster*. When the statistical test of the most likely cluster is significant, sometimes it is of interest to know if there exist any additional clusters. A typical method for detecting an additional cluster is to compare the likelihood ratios of secondary clusters with the maximum likelihood ratios from the simulated data under null hypothesis. This method may lead to a loss of statistical power because the likelihood ratio from the observed data is lower than the maximum likelihood ratio while it is compared with the maximum likelihood ratios from Monte carlo simulations. An alternative approach is to compare the likelihood ratios from secondary clusters with the likelihood ratios from the corresponding secondary clusters from the simulated data. This method is also problematic since the simulation is carried out under the null hypothesis, which does not take into account the existence of one cluster already present in the map. Some existing sequential scan statistic methods for detecting multiple clusters (Li et al., 2011; Zhang et al., 2010) extend the standard spatial scan statistic by removing the shadow effect of detected stronger clusters when identifying secondary weaker clusters. These methods do not allow overlapping clusters because these testing-based procedures typically need a constant disease risk for cells inside each cluster under the alternative hypothesis.

3.1.1. Stepwise testing method

In this section, we present a forward stepwise method for spatial cluster detection. Our method allows for overlapping clusters and provides a frequentist solution to generalized linear models defined in (2) and (4). We start with a null cluster model, i.e., a cluster model with constant risk for each cell, and then iteratively add currently most likely cluster to the cluster model. For simplicity, we will mostly illustrate the proposed methods for Poisson model. All these methods can be used for binomial model in a similar way. Specifically, the procedure consists of the following steps.

1. Start with a null model: $\theta_1, \theta_2, \ldots, \theta_m = 0$.
2. Find the most likely cluster by $\arg\max_j \mathrm{LLR}_{A_j}$, where LLR_{A_j} represents the log-likelihood ratio of the potential cluster A_j for $j = 1, \ldots, m$.
3. Test the significance of the currently most likely cluster.
4. Update and normalize the expected number of cases (or baseline probability if under the binomial framework).
5. Repeat steps 2–4 until the test is not significant.

We now discuss each step of the stepwise testing method.

First, we obtain a null model with a common background risk for all cells in the study area. The expected number of cases $E_i, i = 1, \ldots, N$ is proportional to the population at risk for each cell, or can be calculated after adjusting for some confounding variables. If we consider a binomial model, we can fit a logistic regression model to estimate the effects of confounding variables and use the fitted value of probability as the baseline probability.

We next follow the same way as the standard scan statistic procedure to search for the most likely cluster and perform a statistical test. Let A be a set of cells within a circular window (potential cluster) in the study area, the evidence in favor of A as a cluster is given by the log-likelihood ratio test statistic for H_0: $\rho_i \equiv \rho$ for $\forall i$ versus

$H_A : \rho_i = \rho_{in}$ for $i \in$ A and $\rho_i = \rho_{out}$ for $i \in \mathrm{A}^c$.

$$\mathrm{LLR_A} = y(\mathrm{A}) \log\left[\frac{y(\mathrm{A})}{E(\mathrm{A})}\right] + [y_{tot} - y(\mathrm{A})] \log\left[\frac{y_{tot} - y(\mathrm{A})}{E_{tot} - E(\mathrm{A})}\right],$$

where $y(\mathrm{A}) = \sum_{i=1}^{N} y_i \mathbb{1}_{\{i \in \mathrm{A}\}}$ is the number of cases inside A, $E(\mathrm{A}) = \sum_{i=1}^{N} E_i \mathbb{1}_{\{i \in \mathrm{A}\}}$ is the expected number of cases inside A, $Y_{tot} = \sum_{i=1}^{N} y_i$, $E_{tot} = \sum_{i=1}^{N} E_i$. Without loss of generality, we assume that E_i have been internally standardized so that $Y_{tot} = E_{tot}$. Then we can search through the collection of potential clusters, $A_1, A_2, \ldots, A_p$, for the most likely cluster which maximizes the likelihood ratio test statistic. The spatial scan statistic, i.e., the maximum likelihood ratio test statistic over all potential clusters $\mathrm{LR}_{max} = \max_j \mathrm{LR}_{A_j}$, serves as a global cluster detection test statistic. The global p-value is calculated by comparing LR_{max} with its simulated values under null hypothesis H_0. Under H_0 and the assumption that Y_{tot} is a known constant, the distribution of $(y_1, y_2, \ldots, y_N)$ is multinomial and free of unknown parameters. For Binomial model with covariates, the null distribution, conditional on Y_{tot}, is difficult to simulate, the simulations are drawn unconditionally from multinomial with success probabilities $p_{10}, p_{20}, \ldots, p_{N0}$.

For each iteration, we update E_i's, the expected number of cases in Poisson model, or p_{i0}'s, the baseline probability in binomial model so that the effect of previously detected clusters are taken into account when we test an additional cluster. Our algorithm is similar to the forward stepwise regression that is frequently used in multiple regression problems. However, instead of adding a new variable in each step, the estimated parameters are increased or decreased in a direction to reduce the disparity between the estimated risks inside the currently detected cluster and outside this cluster. Suppose A is the newly detected cluster, we can first multiply the E_i or p_{i0} for $\forall i \in A$ by a factor of $\hat{r}_A$, which is the estimated ratio of risk inside the cluster and outside the cluster for the model induced by the detected cluster A

$$\hat{r}_A = \frac{y(A)/E(A)}{[y_{tot} - y(A)]/[E_{tot} - E(A)]}. \tag{5}$$

Then we normalize the expected number of cases or the baseline probabilities for all cells so that the E_{tot} remains unchanged, i.e., $E_{tot} = Y_{tot}$ for all iterations. This procedure is iterated until the cluster with maximum likelihood ratio becomes insignificant.

3.1.2. Stepwise bias-corrected method

The above algorithm uses hypothesis testing as the stopping rule. Alternatively, we propose a modified version of stepwise approach using bias-corrected log-likelihood to terminate the iteration. We outline the modified stepwise method below.

1. Start with a null model: $\theta_1, \theta_2, \ldots, \theta_m = 0$.
2. Find the most likely cluster by $\arg\max_j \mathrm{LLR}_{A_j}$.
3. Update and normalize the expected number of cases (or baseline probability if under the binomial framework).
4. Compute the bias-corrected log-likelihood via simulations.
5. Repeat steps 2–4 many times.
6. Find the optimal solution by comparing the penalized log-likelihood for different iterations.

Steps 1–3 are exactly the same as step 1, 2 and 4 in stepwise testing method. We skip the hypothesis testing step in this modified stepwise method because we use a different stopping rule for this method. The stepwise testing method evaluates the statistical significance of currently most likely cluster after absorbing the effect of previously detected clusters into the expected number of cases (or baseline probability) for each iteration and stop immediately when a test is not significant. The modified stepwise method also requires a series of iterations; however, it iterates a user-specified number of times to obtain a series of penalized log-likelihood scores.

Now we focus on the method of estimating the bias-correction term for the log-likelihood. Kullback–Leibler information, a measure of the difference between two probability distributions, is widely used in numerous model selection procedures which are based on the likelihood principle. The model selection criterion of the modified stepwise method uses an approximation of the Kullback–Leibler information formula defined as

$$\begin{aligned} & I(g(\cdot); f(\cdot|\hat{\boldsymbol{\rho}}(\mathbf{y}))) \\ &= \int g(\mathbf{x}) \log \frac{g(\mathbf{x})}{f(\mathbf{x}|\hat{\boldsymbol{\rho}}(\mathbf{y}))} d\mathbf{x} \\ &= \int g(\mathbf{x}) \log g(\mathbf{x}) d\mathbf{x} - \int g(\mathbf{x}) \log f(\mathbf{x}|\hat{\boldsymbol{\rho}}(\mathbf{y})) d\mathbf{x}, \end{aligned} \tag{6}$$

where g is the probability density function of an unknown true model G, f is the density function of a parametric model F, which aims at approximating G, $\mathbf{x} = (x_1, \ldots, x_N)^T$ is a running variable, $\mathbf{y} = (y_1, \ldots, y_N)^T$ is the vector of observations, $\boldsymbol{\rho}$ is a vector of parameters and $\hat{\boldsymbol{\rho}}(\mathbf{y})$ is the estimator under the model F. Since the first term of the right hand side of the last equation in (6) is independent of the model F, minimizing the Kullback–Leibler information is equivalent to maximizing a target variable, which we denote by

$$T(\mathbf{y}) = \int g(\mathbf{x}) \log f(\mathbf{x}|\hat{\boldsymbol{\rho}}(\mathbf{y})) d\mathbf{x} = E_{\mathbf{x}}\{\log f(\mathbf{x}|\hat{\boldsymbol{\rho}}(\mathbf{y}))\}. \tag{7}$$

The bias of the log-likelihood with respect to the target variable $T(\mathbf{y})$ can be written as

$$E_{\mathbf{y}}\{T(\mathbf{y}) - \log f(\mathbf{y}, \hat{\boldsymbol{\rho}}(\mathbf{y}))\} = E_{\mathbf{y}} E_{\mathbf{x}} \log \frac{f(\mathbf{x}, \hat{\boldsymbol{\rho}}(\mathbf{y}))}{f(\mathbf{y}, \hat{\boldsymbol{\rho}}(\mathbf{y}))}. \tag{8}$$

Suppose we run a total of S stepwise iterations and derive a series of updated expected number of cases for these iterations, the estimated relative risks in a particular iteration is equal to the ratio of current expected number of cases to the initial expected number of cases. We denote the estimated relative risks in these iterations by $\hat{\boldsymbol{\rho}}^{(1)}(\mathbf{y}), \hat{\boldsymbol{\rho}}^{(2)}(\mathbf{y}), \ldots, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y})$, which corresponds to a series of models $F^{(1)}, F^{(2)}, \ldots, F^{(s)}$. An optimum solution is the one that maximizes the target variables,

i.e., $\arg\max_s T^{(s)}(\mathbf{y})$. Suppose we consider the sth iteration, our goal is to evaluate the expected value of the target variable $T(\mathbf{y})$, denoted $E(T^{(s)}(\mathbf{y}))$, which equals the sum of the log-likelihood and the bias term expressed in (8). The bias term can be estimated by replacing $\mathbf{x}$ and $\mathbf{y}$ with Monte Carlo simulations under the model $F^{(s)}$. In practice, we simulate data $\mathbf{x}^{(s)}$ and $\mathbf{y}^{(s)}$ from a multinomial distribution with parameters y_{tot} and $(\hat{\rho}_1^{(s)}E_1/y_{tot}, \hat{\rho}_2^{(s)}E_2/y_{tot}, \ldots, \hat{\rho}_N^{(s)}E_N/y_{tot})$. Then we derive the estimated parameters $\hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^{(s)})$ after one iteration, starting from the simulated data $\mathbf{y}^{(s)}$ and the expected number of cases in the $(s-1)$th iteration
$(\hat{\rho}_1^{(s-1)}E_1, \hat{\rho}_2^{(s-1)}E_2, \ldots, \hat{\rho}_N^{(s-1)}E_N)$. The bias adjustment of log-likelihood in the sth iteration is estimated by

$$B^{(s)}(\mathbf{y}) = E_* \log \frac{f(\mathbf{x}^{(s)}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^{(s)}))/f(\mathbf{x}^{(s)}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))}{f(\mathbf{y}^{(s)}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^{(s)}))/f(\mathbf{y}^{(s)}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))},$$

where E_* stands for an average of the Monte Carlo simulation results. The bias-corrected log-likelihood for the estimated parameters in the sth iteration can be calculated by $\log[f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))] + B^{(s)}(\mathbf{y})$. The model with maximum bias-corrected log-likelihood is then selected from the solution path. In simulation studies, we find that the calculated bias terms in different iterations do not change dramatically. From a simulation study we found little variations in bias terms for different iterations. To reduce the computational complexity, we propose to estimate bias terms only in iteration 1–3, and use the estimated bias in iteration 3 for the remaining iterations.

3.1.3. Stepwise information criterion method

Stepwise Bias-Correction Method uses penalized log-likelihood as the model selection criterion and relies on the estimation of bias correction. Alternatively, we consider several information criteria as means for model selection. The Akaike Information Criterion (AIC) is a popular way of selecting a model from a set of models. Akaike (1974) defined it as:

$$\mathrm{AIC} = -2\ln(\mathrm{L}) + 2(k+1),$$

where $k+1$ is the number of free parameters in the model and k is the number of clusters when a cluster model is considered, and L is the maximized value of the likelihood function for the model. Hurvich and Tsai (1989) proposed a corrected version of AIC, named AICc:

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2(k+1)(k+2)}{n-k-2},$$

where n denotes the sample size. BIC, the Bayesian information criterion, was introduced by Schwarz (1978) as a competitor to AIC. The formula of BIC is

$$\mathrm{BIC} = -2\ln(\mathrm{L}) + (k+1)\cdot\ln(n).$$

These criteria are composed of a goodness of fit component, i.e., the log-likelihood, and a complexity component, that is a function of number of parameters and number of observations. The most prominent advantage of using these criteria is they are computationally cheap compared with the other two stepwise methods. However, as the correction term in these criteria is a simple minded bias adjustment to the log-likelihood, there is no assurance that the bias correction yields a good estimate of Kullback–Leibler information.

3.2. Stagewise method

In this section, we propose a stagewise method for spatial cluster detection. We obtain the entire solution path in a stagewise fashion via a series of tiny steps. Then we propose some possible stopping rules to search for the optimal solution.

3.2.1. Generalized monotone forward stagewise algorithm

In model (2) and (4), $\mathbb{1}_{\{d(\mathbf{z}_i,\mathbf{c}_j)\le r_j\}}$ is the dummy variable indicating whether cell i belongs to the potential cluster j. For simplicity, we denote it by x_{ij}. In practical applications, the number of potential clusters p can be extremely large. Consider the well-known New York leukemia data that consists of 789 cells, there are 191, 129 potential clusters when we use circular windows with a maximum radius of 20 miles. In such a high-dimensional setting ($m \gg N$), regular regression methods may encounter overfitting issues. Consequently, we need a variable selection procedure to select a parsimonious set of covariates from the huge amount of potential clusters. LASSO is an appealing method for variable selection due to its property of shrinking some of the model coefficients to exactly zero. Hastie et al. (2001) showed that the incremental forward stagewise algorithm solves a version of the LASSO problem that enforces monotonicity. We hope to leverage the shrinkage and selection properties of the forward stagewise regression to select a parsimonious set of potential clusters. The algorithm of our proposed forward stagewise method is as follows.

1. Start with a null model: $\theta_1, \theta_2, \ldots, \theta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ with largest absolute value of gradient element $\frac{\partial L}{\partial \theta_j}$ evaluated at the current model, where L is the loss function.
3. Given a fixed step size $\epsilon > 0$, update the coefficient estimate by $\theta_j \leftarrow \theta_j + \epsilon \cdot sign(\frac{\partial L}{\partial \theta_j})$.
4. Update and normalize the expected number of cases (or baseline probability if under the binomial framework).
5. Repeat steps 2–4 many times.

We now elaborate on these steps.

First, we start with a null model with the intercept α equal to overall rate across the study area, and clustering component equal to zero. We standardize the covariates so that they have mean 0 and unit length.

Second, the algorithm identifies the best direction by looking for the most extreme gradient element, i.e., $\arg\max_j |\frac{\partial L}{\partial \theta_j}|$. We use log-likelihood as the loss function in our procedure. The Poisson log-likelihood is given by

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} [y_i \log(\rho_i E_i) - \rho_i E_i - \log(y_i!)].$$

So the gradient is

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j}\left\{\sum_{i=1}^{N}[y_i \log(\rho_i E_i) - \rho_i E_i - \log(y_i!)]\right\} \\ &= \frac{\partial}{\partial \theta_j}\left\{\sum_{i=1}^{N}[y_i(\alpha + \sum_{j=1}^{p}\theta_j \tilde{x}_{ij}) - E_i e^{\alpha+\sum_{j=1}^{p}\theta_j \tilde{x}_{ij}}]\right\} \\ &= \sum_{i=1}^{N}[y_i \tilde{x}_{ij} - (E_i \rho_i)\tilde{x}_{ij}] \\ &= \sum_{i=1}^{N}[(y_i - \mu_i)\tilde{x}_{ij}],\end{aligned}$$

where $\tilde{x}_{ij}$ is the standardized ith element of jth covariate, and $\mu_i = \rho_i E_i$ is the expected number of cases in cell i. The binomial log-likelihood is given by

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N}[y_i \log(p_i) + (n_i - y_i)\log(1 - p_i)].$$

Similarly, we can derive that the gradient is

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j} &= \sum_{i=1}^{N}[y_i \tilde{x}_{ij} - (n_i p_i)\tilde{x}_{ij}] \\ &= \sum_{i=1}^{N}[(y_i - \mu_i)\tilde{x}_{ij}],\end{aligned}$$

where $\mu_i = n_i p_i$ is the expected number of cases in cell i.

Next, we increment the coefficient for the covariate with the most extreme gradient by an amount $\pm\,\epsilon$ with the sign determined by the sign of gradient. Occasionally, the selected predictor may correspond to a potential cluster with an extremely large population at risk. In such situation, step size is not small enough and the increment will be too big. The same potential cluster will be selected constantly and ϵ will be added to and subtracted from the corresponding coefficient repeatedly. To avoid this endless loop, we can replace the fixed step size with an adaptive step size. That is, we reduce the current step size by a factor of 2 whenever above situation occurs. Let S be the number of iterations, which is prespecified and usually very big, and ϵ_s be the step size in the sth iteration. This algorithm generates a forward stagewise path, indexed by the total distance stepped $d = \sum_{s=1}^{S}\epsilon_s$. Under certain conditions, the limiting version of forward stagewise paths coincide with the lasso paths (Efron et al., 2004). However, for most problems the forward stagewise paths and lasso paths are different. The predictors can drop out in lasso, but the corresponding predictors may go flat instead of turning back towards zero in stagewise method. The forward stagewise procedure behaves like a monotone version of lasso, which tends to slow down the search, not allowing the sudden changes of direction that can occur with the lasso. For problems with large number of correlated predictors, the forward stagewise procedure will produce similar coefficient profiles in the early stages as the lasso method. For the later stage, the forward stagewise paths will be much smoother and takes longer to overfit. So the forward stagewise might be preferable to the lasso method when there is a large number of correlated predictors.

3.2.2. Stopping rules of stagewise algorithm

Now we consider three categories of stopping rules: information criteria, bias estimation via Monte Carlo simulation under the null model with constant risk, and bootstrap estimation of log-likelihood bias. All these rules are closely related to the Kullback–Leibler information (6).

Information criteria. Similar to stepwise information criterion method, we consider AIC, AICc and BIC as the candidate model selection methods for stagewise algorithm. These simple bias adjustment methods do not necessarily provide exceptionally good estimates of Kullback–Leibler information, but they are favorable in terms of computational efficiency.

Bias estimation via Monte Carlo simulation under the null model. The bias in (8) is difficult to estimate because the distribution G is unknown and the running variable $\mathbf{x}$ cannot be simulated. A null model with common relative risk, conditional on y_{tot} and observed locations, is multinomial and free of unknown parameters. Under the null model, vector of observations, denoted by $\mathbf{y}^*$, and vector of running variable, denoted by $\mathbf{x}^*$, can be generated according to a multinomial distribution with parameters y_{tot} and $(E_1/y_{tot}, E_2/y_{tot}, \ldots, E_N/y_{tot})$. We run our forward stagewise algorithm using $\mathbf{y}^*$ in place of observed data $\mathbf{y}$ and denote the estimated relative risk in the sth iteration by $\hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*)$. It is expected that after replacing $\mathbf{x}$ and $\mathbf{y}$ by $\mathbf{x}^*$ and $\mathbf{y}^*$, the bias estimate for the sth iteration

$$B_0^{(s)} = E_* \log \frac{f(\mathbf{x}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))}{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))}$$

will be close to the bias in (8). Here, E_* stands for average of results of Monte Carlo simulations.

Bootstrap estimation of log-likelihood bias. The bootstrap is a powerful tool for estimating various properties of a given statistic by sampling from an approximating distribution, most commonly empirical distribution of the observed data. It can be used to estimate the bias of log-likelihood evaluated at the parameter estimates in each iteration with respect to the target variable $T(\mathbf{y})$. By changing the position of $\mathbf{y}$ and $\mathbf{x}$ in the formula (8), we can write the bias of the log-likelihood evaluated at the parameter estimates from model in the sth iteration ($s = 1, 2, \ldots$) with respect to the target variable $T(\mathbf{y})$ as

$$\begin{aligned}E_{\mathbf{y}}\{T(\mathbf{y}) - \log f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))\} &= E_{\mathbf{y}}E_{\mathbf{x}} \log \frac{f(\mathbf{x}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))}{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))} \\ &= E_{\mathbf{y}}E_{\mathbf{x}} \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{x}))}{f(\mathbf{x}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{x}))}. \quad (9)\end{aligned}$$

In practice, we can use the parametric bootstrap as the resampling method for estimating the bias term and replace the bootstrap expectation by an average of the results of Monte Carlo simulations. The main idea of bootstrap estimation of bias correction is to replace the unknown running variable $\mathbf{x}$ in (9) by the parametric bootstrap sample $\mathbf{y}^*$, which is drawn from the maximum likelihood model.

The expectation of a bootstrap estimate

$$B_1^{(s)}(\mathbf{y}) = E_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))}{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))}$$

with respect to $\mathbf{y}$ is expected to be quite close to the bias (9), and the same bootstrap estimate is used as the bias term in WIC (Ishiguro and Sakamoto, 1991). Besides B_1, four alternative bootstrap estimates can also be used (Cavanaugh and shumway, 1997; Shibata, 1997). The formulae of these bootstrap estimates are

$$B_2^{(s)}(\mathbf{y}) = 2E_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))}{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))}$$
$$B_3^{(s)}(\mathbf{y}) = 2E_* \log \frac{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))}{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))}$$
$$B_4^{(s)}(\mathbf{y}) = 2E_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))}{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))}$$
$$B_5^{(s)}(\mathbf{y}) = 2E_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))}{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))}.$$

The above bootstrap estimates of the bias share a lot similarities. The difference is mainly about where the bootstrap sample $\mathbf{y}^*$ is placed in the definition of the log likelihood ratio. We note that in B_2, B_3 the observed outcome of likelihood in the denominator is the same as that in the numerator, and in B_1, B_4, and B_5, the observed outcome of likelihood are different in the denominator from that in the numerator. The difference of $*$ position where the bootstrap sample is used may cause deviation from the true bias (9). To handle this problem, we modify B_1, B_4, and B_5 by

$$B_1^{(s)}(\mathbf{y}) = E_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))/f(\mathbf{y}, \boldsymbol{\rho}_0)}{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))/f(\mathbf{y}^*, \boldsymbol{\rho}_0)}$$
$$B_4^{(s)}(\mathbf{y}) = 2E_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))/f(\mathbf{y}, \boldsymbol{\rho}_0)}{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))/f(\mathbf{y}^*, \boldsymbol{\rho}_0)}$$
$$B_5^{(s)}(\mathbf{y}) = 2E_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}))/f(\mathbf{y}, \boldsymbol{\rho}_0)}{f(\mathbf{y}^*, \hat{\boldsymbol{\rho}}^{(s)}(\mathbf{y}^*))/f(\mathbf{y}^*, \boldsymbol{\rho}_0)},$$

where $\boldsymbol{\rho}_0 = (1, 1, \ldots, 1)^T$ is the initial parameter estimate of our algorithm.

For simplicity, the bias estimation formulas in this section are illustrated only for Poisson models. The parametric bootstrap sample $\mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_N^*)^T$ is drawn from a multinomial distribution with parameter y_{tot} and $(y_1/y_{tot}, y_2/y_{tot}, \ldots, y_N/y_{tot})$. If we consider a binomial model, the bootstrap sample $(y_1^*, n_1^* - y_1^*, \ldots, y_N^*, n_N^* - y_N^*)^T$ can be simulated from a multinomial distribution with parameter n_{tot} and $(y_1/n_{tot}, (n_1 - y_1)/n_{tot}, \ldots, y_N/n_{tot}, (n_N - y_N)/n_{tot})$, where $n_{tot} = \sum_{i=1}^N n_i$. The above formulae of bias estimation $B_0 - B_5$ can be used after we replace $\hat{\boldsymbol{\rho}}^{(s)}$ by $\hat{\mathbf{p}}^{(s)}$ and replace $\boldsymbol{\rho}_0$ by $\mathbf{p}_0 = (p_{10}, p_{20}, \ldots, p_{N0})^T$.

4. Simulation study

In this section, we evaluate the performance of the proposed methods via simulations using an idealized square grid structure and the geographic structure of Indiana. For these simulations we consider the no cluster scenario and single cluster scenarios with a given risk ratio of clustered regions to background regions.

We evaluate the performance of our methods in terms of two quantities: root average mean square error (RAMSE) of the estimated SIR and the probability of belonging to the estimated cluster. The RAMSE of the estimated SIR measures the combined accuracy of all the estimates for SIR. The probability of belonging to the estimated cluster at a certain cell is the percentage of detected clusters containing that cell.

For Poisson models, the RAMSE of the estimated SIR is given by

$$\text{RAMSE} = \sqrt{\frac{\sum_{i=1}^N E_i(\widehat{\rho_i} - \rho_i)^2}{\sum_{i=1}^N E_i}},$$

where $\hat{\rho}_i$ is the estimated SIR, ρ_i is the true SIR and E_i is the expected cases in cell i. For binomial models, the formula becomes

$$\text{RAMSE} = \sqrt{\frac{\sum_{i=1}^N n_i\left(\frac{\hat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)^2}{\sum_{i=1}^N n_i}}.$$

Values of RAMSE may be used for comparative purposes. Smaller RAMSE values indicate more accurate estimates.

4.1. 30 × 30 square grid

We apply our methods to a 30 × 30 square grid, which is a square region divided into 900 cells in a regular grid of 30 rows and 30 columns. The expected number of cases under the null hypothesis, E_i, is assumed to be identical for all cells. Each side of a cell is 1 unit. The set of potential clusters consists of 11,104 circular windows centered at each cell with radii ranging from 0 up to 2 units. We assume that the numbers of cases in each cell are independent Poisson random variables. A total of 900 cases are simulated in the 900 cells for two scenarios: the null model (no clustering) and a model with single cluster with a risk ratio of 2. We generate 1000 random data sets under each scenario. The proposed methods are divided into four categories: stepwise methods, stagewise methods using information criteria, stagewise methods using B_0 for bias estimation, and stagewise methods using bootstrapping for bias estimation ($B_1 \sim B_5$).

Fig. 1 displays the RAMSE of estimated SIR from the simulated data when applying the proposed methods. We evaluate the overall accuracy of the SIR estimates from the distributions of the RAMSE for the 1000 simulations. The results of the stepwise information criteria method are not included in this figure because this method exhibits a dramatic increase in the RAMSE values compared with all other methods under both simulation scenarios. For the the scenario of no clustering, we observe a concentration of RAMSE values near 0 when applying stepwise testing and stepwise bias-corrected method. Among the stagewise methods using information criteria, the RAMSE values by AIC and AICc spread out relatively evenly between 0.05 and 0.25 and the values by BIC are much more concentrated between minimum and upper quartile RAMSE

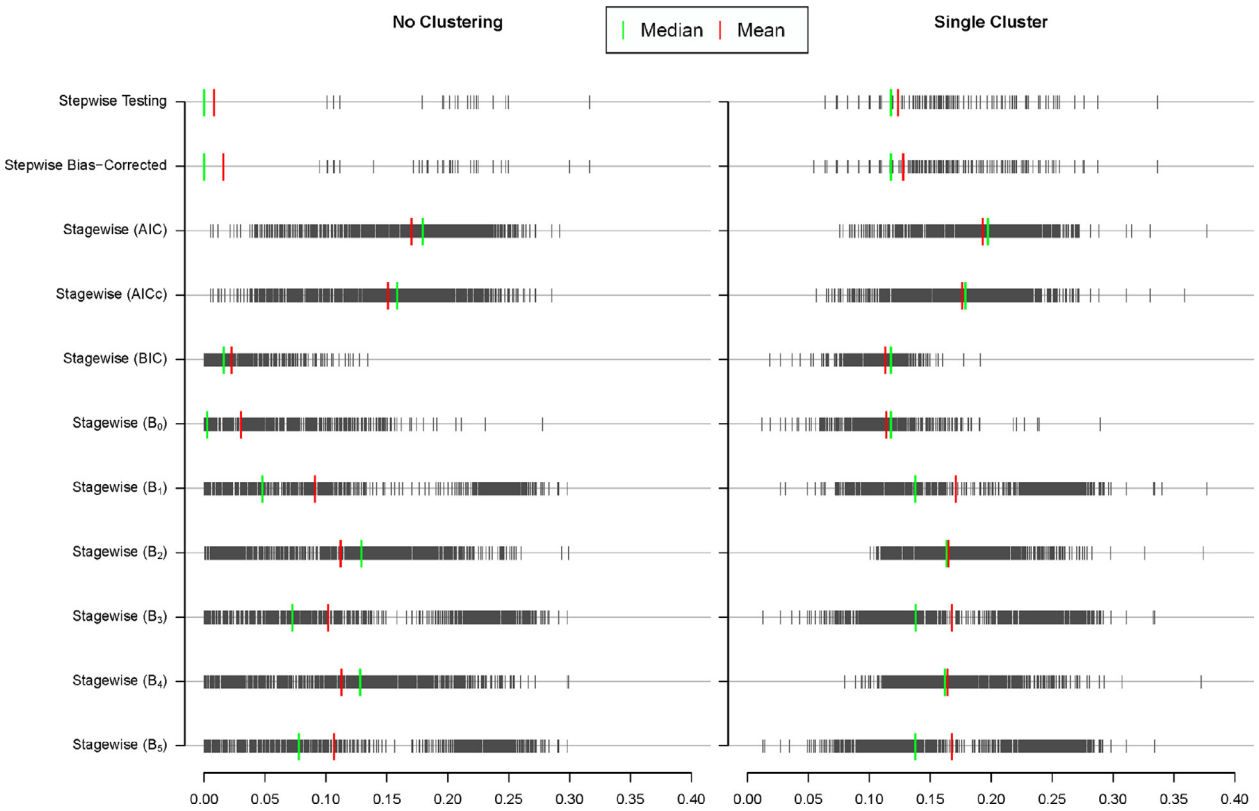

Fig. 1. Simulated distribution of the RAMSE of the estimated SIRs under the No Clustering and the Single Cluster scenario for the square grid. Mean and median RMSE are indicated by longer red and green bars, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

values. Among the remaining methods, stagewise methods by B_0, B_1, B_3 and B_5 have concentrations near minimum and median RAMSE values. For the single cluster scenario, we observe that 89.4% and 83.3% of the simulations have RAMSE values below 0.15 for stepwise testing method and stepwise bias-corrected method, respectively. Among the stagewise methods using information criteria, only 15.4% and 29.8% simulations have RAMSE values below 0.15 for AIC and AICc methods, and 99.5% simulations have RAMSE values below 0.15 for BIC method. The stagewise methods by B_0, B_3 and B_5 are preferable to the remaining methods. From the performance of these methods in terms of estimation accuracy, stepwise testing is as good as stepwise bias-corrected method and both methods outperform the stepwise information criteria methods, and stagewise by BIC produces smaller RAMSE values than any other stagewise methods. So stepwise testing, stagewise by BIC, stagewise by B_0 and stagewise by B_3 are considered best in each category of proposed methods.

We illustrate the cluster detection power of the above 4 methods in Fig. 2. An idealized result, labeled oracle, is presented in the first column of maps, where the pure red and light blue are observed. The figure shows that stagewise methods have generally higher detection percentages at the clustered regions than the stepwise testing method. Intuitively, the stagewise methods which involve a series of tiny iterative steps are more likely to detect a cluster than stepwise methods. The stagewise method based on bootstrap estimate B_3 tends to mistakenly select more background cells than the other two stagewise methods. In general, the stagewise method based on BIC has higher detection percentage in the clustered regions than stepwise method and lower false detection percentage in the background regions than the other stagewise methods.

4.2. Indiana

To further explore the differences between the proposed methods for binomial models, we perform a simulation study using the underlying geography and population structure from the 92 Indiana counties in the 2000 Census. The set of potential clusters consists of 1028 circular windows with radii ranging from 0 to 100 km. The number of cases in each county is assumed to follow a binomial distribution with the identical baseline probabilities. We simulate a total of 559, 484 cases in the 92 counties for each of the 4 cluster models: the null model and 3 models with a single cluster. We assume that $p_{i0} = 0.095, i = 1, 2, \ldots, N$. For the null model, the true probabilities of events are equal to the baseline probabilities. The single cluster models have a risk ratio of 1.05 and the cluster locations are shown in Fig. 3. We choose these cluster locations because they represent large, medium and small population size, respectively. The sensitivity of the proposed methods to

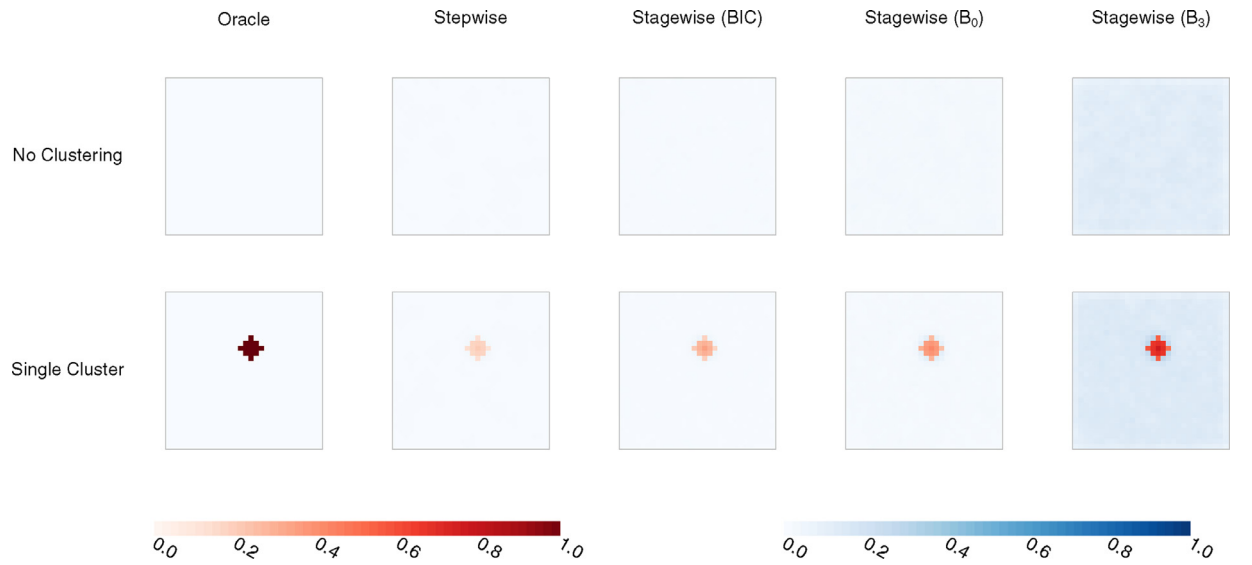

Fig. 2. Probability, under the null hypothesis of constant risk and single cluster scenario, of belonging to the estimated cluster. Cells in the background are shaded in blue, and cells in the cluster are shaded in red. The intensity of the color represents the probability of the cell belonging to the detected cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

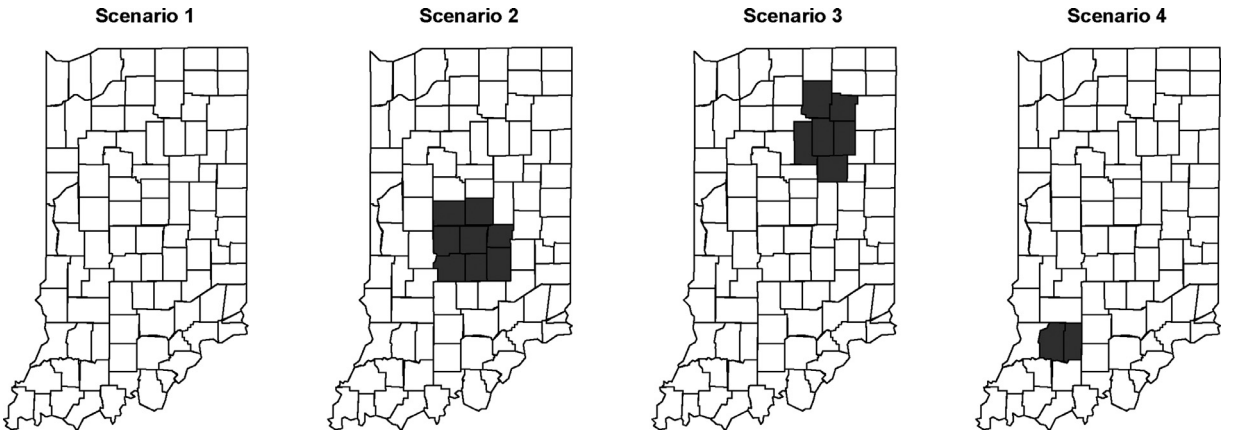

Fig. 3. Scenario 1 is the null model with no clusters, scenario 2–4 are single cluster models centered at Marion county, Wabash county and Martin county, with a population of 1,442,990, 275,633 and 39,460, respectively.

the population size within the clusters can therefore be analyzed.

The RAMSE of estimated SIR for 1000 simulations under the four scenarios are presented in Fig. 4. The stepwise information criterion method greatly overestimates the clustering effects and yields much larger RAMSE than other methods, so it will not be considered in the following studies. For scenario 1 where there is a constant risk in the whole study area, stepwise testing method and stepwise bias-corrected method both produce quite accurate estimates. About 95% simulations using the former and 90% simulations using latter yield estimates with 0 RAMSE. Stepwise methods using information criteria have much larger RAMSE values than all the other methods. Stagewise methods using information criteria and B_0 bias estimate have generally better performance than stagewise methods using bootstrap estimate of bias. Stagewise methods do not give as many 0 RAMSE values as the two stepwise methods, which is not surprising because stagewise methods are more inclined to detect a cluster than stepwise methods. However, we find that the relative risks of most detected clusters when using stagewise (BIC) method are very close to 1. Less than 1% of RAMSEs for stagewise (BIC) are larger than 0.003, whereas 5% RAMSEs for stepwise testing method and 8% for stepwise bias-corrected method are greater than 0.005. In the results for scenario 2 and 3, the stepwise testing and bias-corrected methods have better accuracy than all the stagewise methods. Stepwise information criteria methods are worse than any other methods. Among the stagewise methods, stagewise (B_0) is slightly favored than others. For scenario 4, 98.4% of the 1000 simulations have RAMSE values below 0.005 when stagewise (BIC) is applied, while only 92.4% and 85.3% RAMSE values are below 0.005 when stepwise

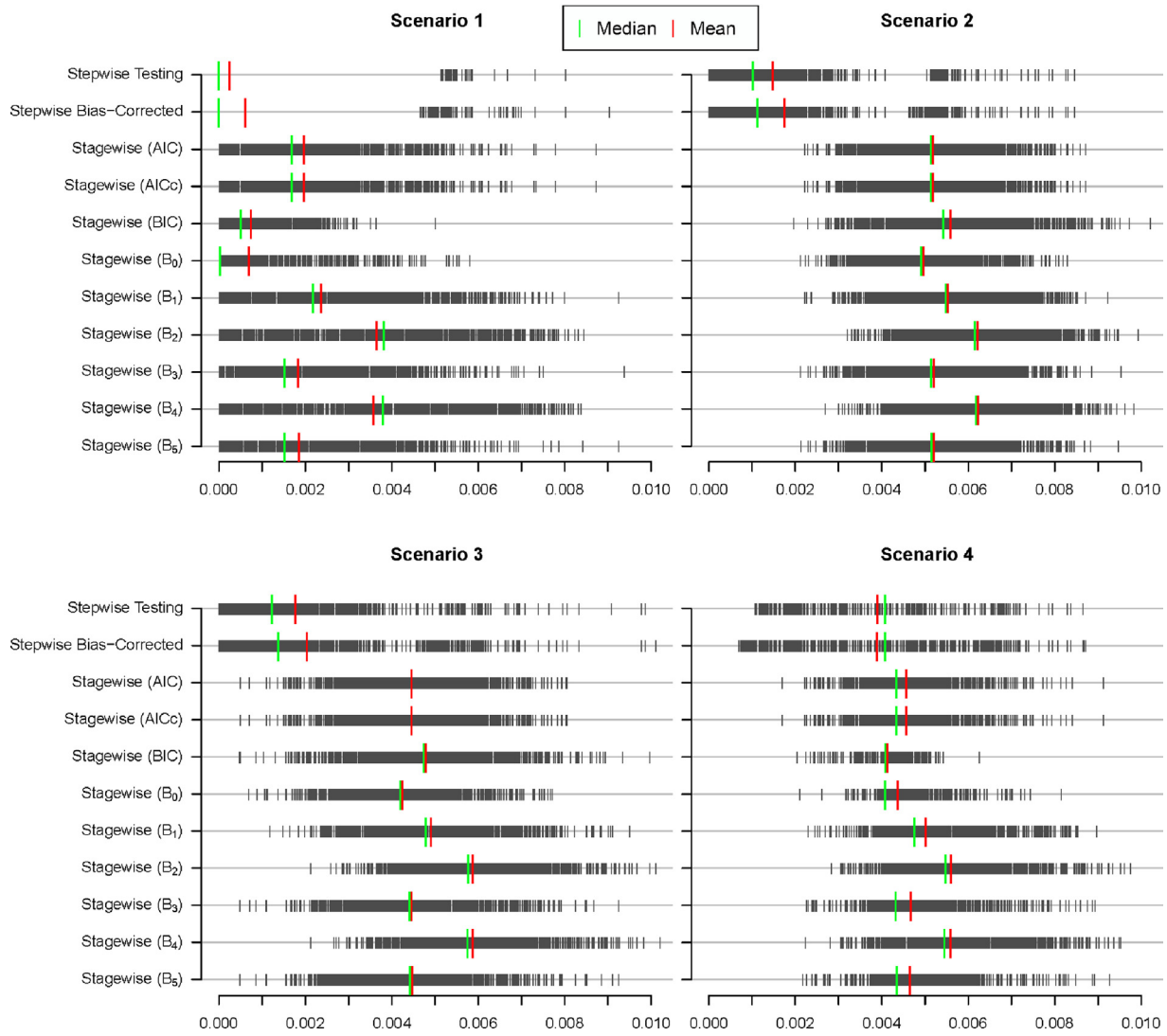

Fig. 4. Simulated distribution of the RMSE of the estimated SIRs under the No Clustering and three Single Cluster scenarios for the Indiana data. Mean and median RMSE are indicated by longer red and green bars, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

testing and stepwise bias-corrected methods are used respectively. In summary, stepwise testing method and stepwise bias-corrected method perform almost equally well and stagewise (BIC) and stagewise (B_0) are more accurate than the other stagewise methods.

Fig. 4 demonstrates that the stepwise testing method and the stagewise (BIC) method are the most favored stepwise and stagewise methods respectively. Now we also present maps of the estimated SIR for some selected simulations when applying these two methods. We choose the simulations with minimum, lower quartile, median, upper quartile and maximum RAMSE values from the 1000 simulations under scenario 1 and scenarios 2 respectively and display the observed and estimated SIR's of the selected simulations in Figs. 5 and 6. For scenario 1, we observe that both methods produce quite accurate estimates for most of the simulations except the ones with maximum RAMSE value. The worst situation for stagewise (BIC) method seems better than the worst situation for stepwise testing method since there are fewer false detections when stagewise (BIC) method is applied to the simulation with maximum RAMSE. For scenario 2, both methods produce generally good estimates for those simulations. Both methods correctly identify the elevated cluster centered at Marion county, however there are more wrong detections at background area for stepwise testing method than stagewise (BIC) method.

The probability of belonging to the estimated cluster at each cell for oracle, stepwise testing and three stagewise methods are displayed in Fig. 7. For scenario 1, stepwise and stagewise (BIC) have very low wrong detection percentage. The stagewise methods have a few wrong detections around Marion county, which has the largest population in Indiana. For scenario 2–4, every method

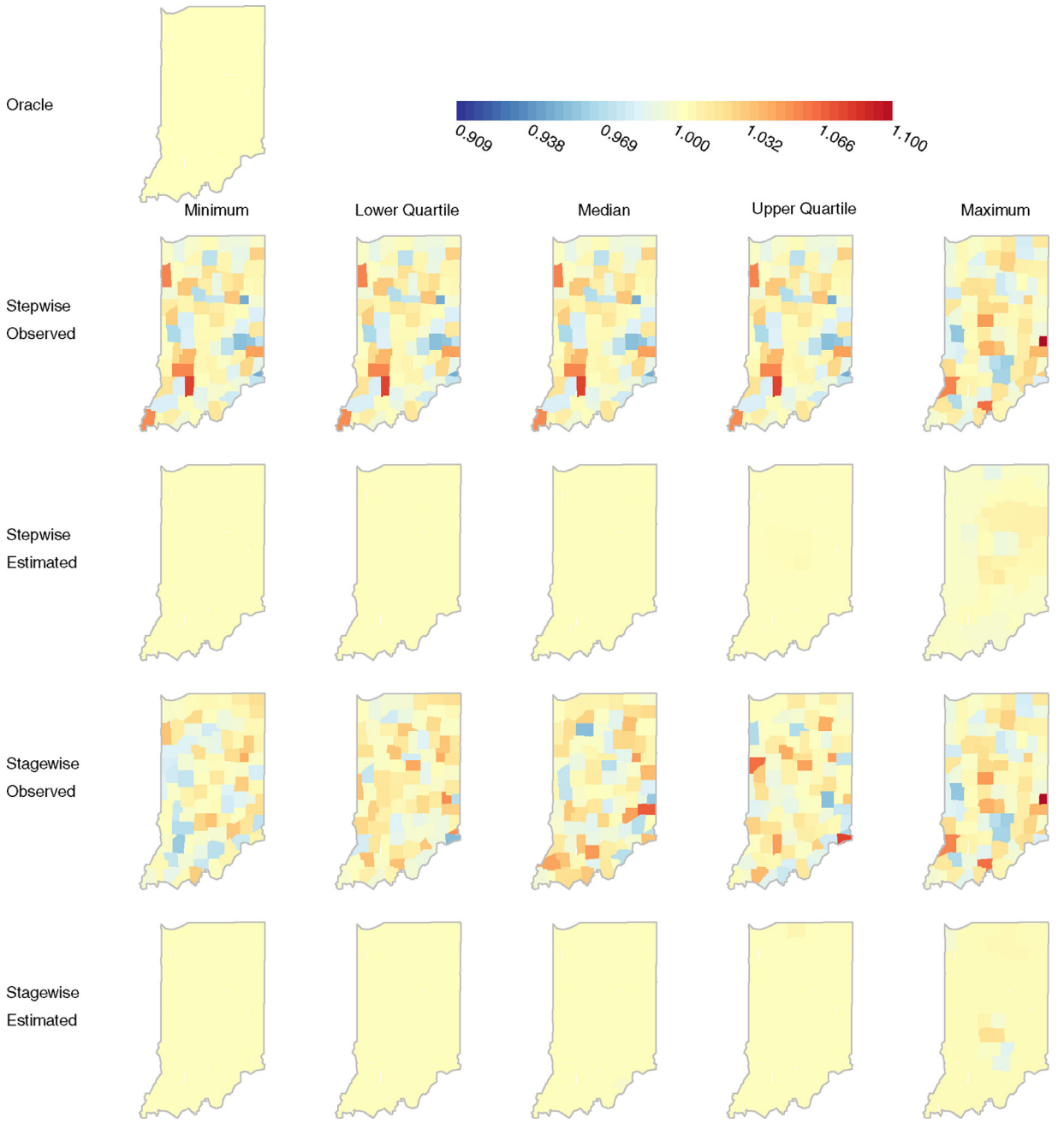

Fig. 5. Observed and estimated relative risks using stepwise testing algorithm and stagewise (BIC) algorithm for five simulations, which have minimum, lower quartile, median, upper quartile and maximum RAMSE out of 1000 replicates under the no clustering scenario (scenario 1). The true relative risks are presented in the top left corner.

successfully finds the elevated cluster. Still, the percentages at some background regions are relatively higher for the stagewise methods than stepwise methods. Given that stagewise methods are based on a series of tiny steps starting from the null model of constant risk, the result is acceptable if the percentages of detections at background regions are slightly higher than 0 and the estimated SIR's at these cells are close to 1.

5. Examples

5.1. Example: New York leukemia data

In year 1986, the New York State Department of Health released a data set on leukemia incidence for a five-year period (1978–1982) in an eight-county region of upstate New York. In alphabetical order, the eight counties are

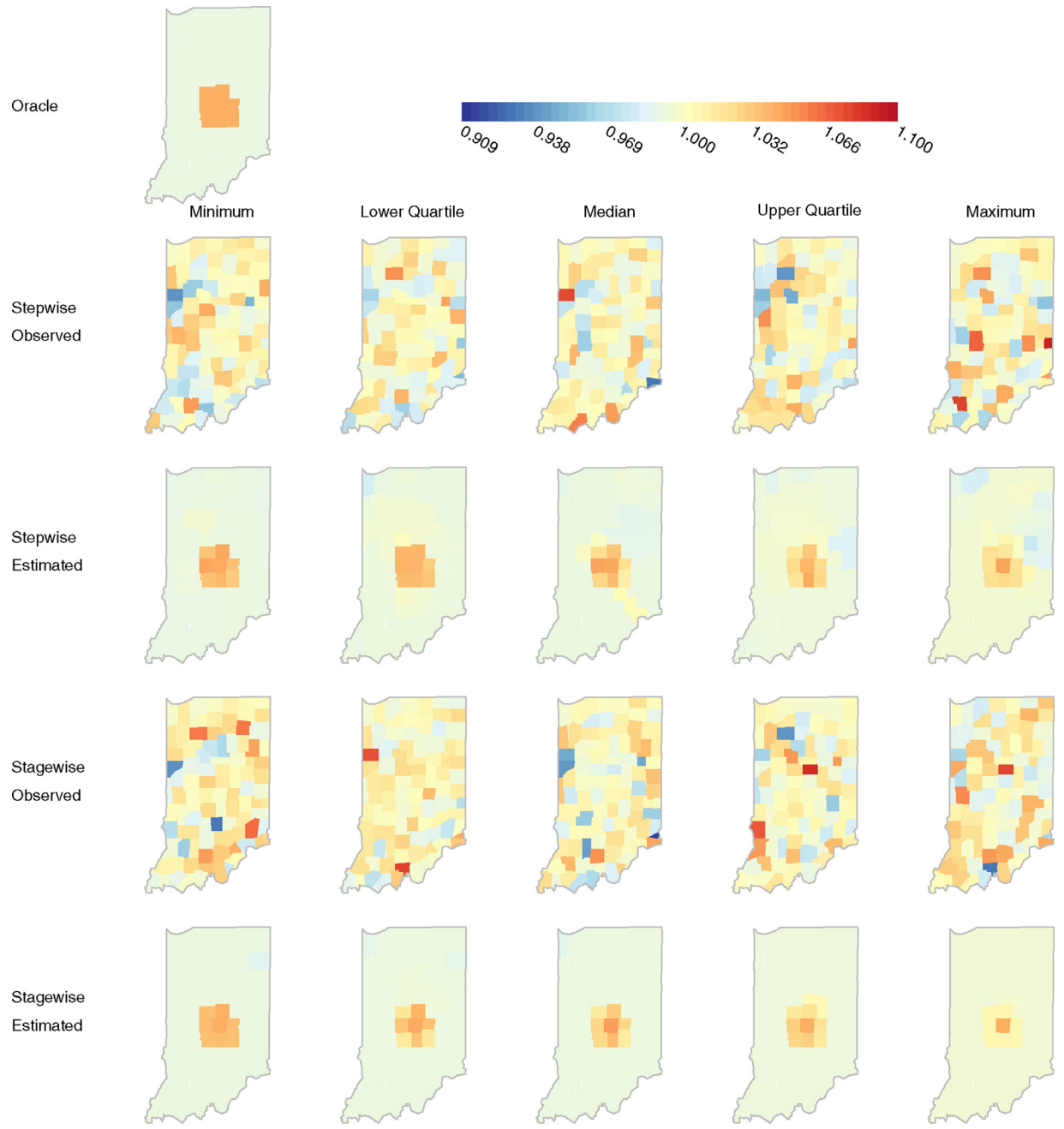

Fig. 6. Observed and estimated relative risks using stepwise testing algorithm and stagewise (BIC) algorithm for five simulations, which have minimum, lower quartile, median, upper quartile and maximum RAMSE out of 1000 replicates under the hypothesis of a single cluster centered at Marion county (scenario 2). The true relative risks are presented in the top left corner.

Broome, Cayuga, Chenango, Cortland, Madison, Onondaga, Tioga and Tompkins. The two largest cities in the study region are Syracuse in Onondaga County and Binghamton in Broome County. As displayed in Fig. 8, the eight-county region is divided into 790 cells (census blocks or census tracts). For each cell, the population at risk is the population from the 1980 U.S. census. Following previous work, we used a maximum radius of 20 km for which the largest potential cluster is roughly 10% of the total study area. There are 191, 129 potential clusters given this upper bound of cluster radius. We choose a step size $\epsilon = 0.0001$ and run the stagewise procedure for 5000 steps. Many previous analyses of the New York leukemia data have been based on hypothesis testing methods or focused on detecting a single cluster with an elevated or lowered risk. These methods showed evidence of clustering in either

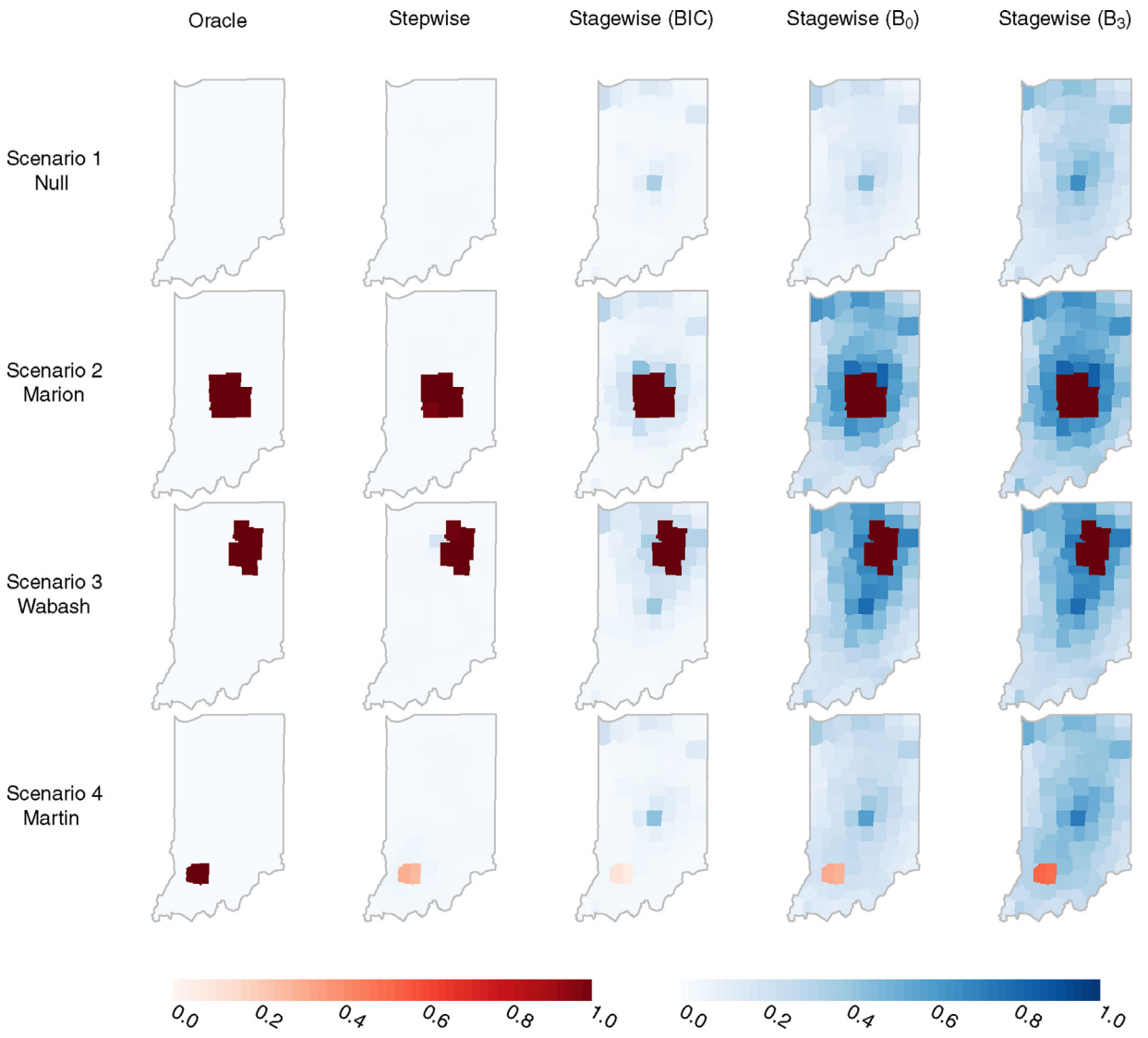

Fig. 7. Probability of belonging to the estimated cluster by oracle, stepwise testing, stagewise (BIC), stagewise(B_0) and stagewise(B_3) methods under scenario 1–4.

Broome county or Cortland county (Waller et al., 1994). Some methods (Gangnon and Clayton, 2001; Kulldorff and Nagarwalla, 1995) detected clusters of elevated risk in both locations. Gangnon and Clayton (2000, 2003) found evidence for three clusters: areas of clustering in Broome and Cortland counties with an increased risk of leukemia and an area of clustering in Onondaga county, north of Syracuse, with a decreased incidence of leukemia.

The observed SIR and the estimated SIR from five selected methods are displayed in Fig. 9. The stagewise methods by AIC, AICc and B_0 identify two areas of clustering in Broome and Cortland counties with an elevated risk of leukemia and an area of clustering in Onondaga county, which is associated with a lowered risk of leukemia. The term 'area of clustering' is used instead of 'clusters' to indicate that many different clusters are detected in a particular area. Besides these three areas of clustering, some additional clusters are also found in the maps of these three methods, which may be a sign of overfitting. The stepwise testing method shows evidence of clustering in three areas of clustering in Broome, Cortland and Onondaga counties, all of which are associated with an increased risk of leukemia. The stagewise method by BIC only identify an elevated risk of leukemia in Broome county. Both stepwise testing and stagewise (BIC) methods have very clean background and are considered to be very useful in identifying the most obvious clusters. The stagewise (BIC) method is more conservative. The remaining stagewise methods tend to produce a noisy background and are less attractive than the stepwise testing and stagewise (BIC) methods.

5.2. Example: Indiana poverty data

The Indiana poverty data set covers 92 counties in Indiana. For each county, the counts of individual poverty cases

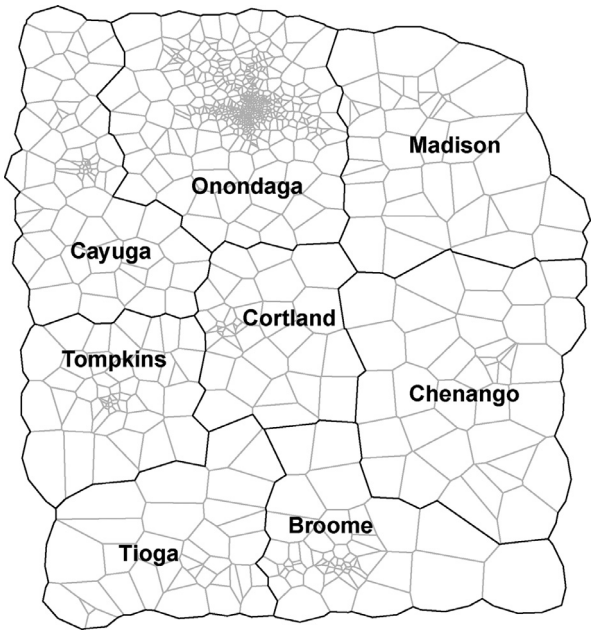

Fig. 8. Map of Dirichlet tessellation of 789 cell centroids for the New York data. Cell boundaries are in light gray and county borders are in black.

and the population are available in U.S. census year 2000. The latitudes and longitudes of geographical centroid of each county are included in the data set. The set of potential clusters consists of 1028 circular clusters centered at the 92 distinct cell centroids with radii ranging from 0 up to 100 km. The maximum cluster radius was chosen such that the largest potential cluster is roughly 1/3 of the total study area. The data set also contains racial composition and labor force composition of the county. We fit a logistic regression model incorporating these predictors and use the fitted value as the baseline probability in model (2).

In Fig. 10, we present estimated relative risk for stepwise testing and stagewise (AIC and BIC) methods using the Indiana poverty data in year 2000. The estimates are all very close to the observed poverty rates. All these methods identify much more clusters than expected, indicating that the data support a saturated fit. From the analysis of Indiana poverty data, we find no evidence for a parsimonious set of clusters. It indicates overfitting occurs in our clustering model where spatial heterogeneity is overlooked.

6. Discussion

The standard spatial scan statistic method is useful in detecting a single cluster in the study area. A strong cluster may hide the existence of a secondary cluster in another region in the map. Several sequential approaches are

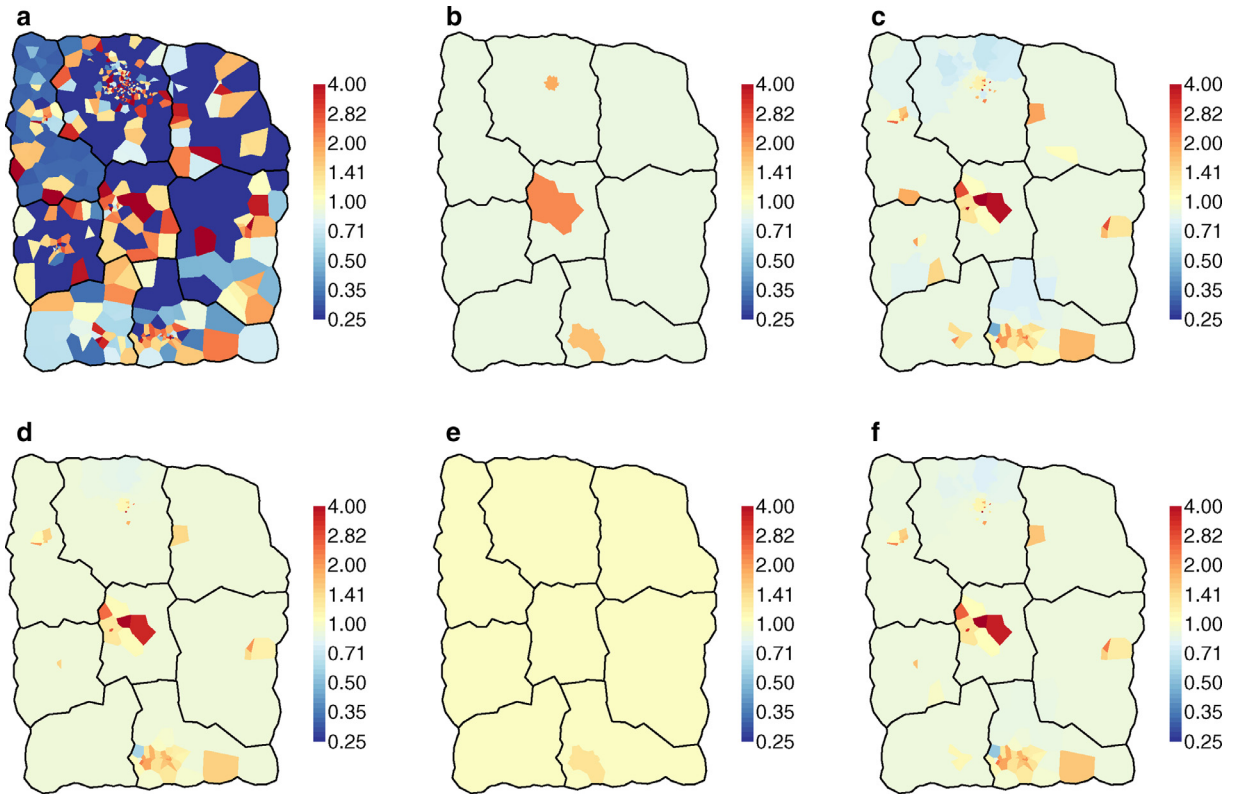

Fig. 9. Maps of (a) the observed leukemia incidence rates (relative to the overall rate of 5.6 per 10,000 persons), and the estimated SIR using (b) stepwise testing, (c) stagewise (AIC), (d) stagewise (AICc), (e) stagewise (BIC), (f) stagewise (B_0) methods for the New York data.

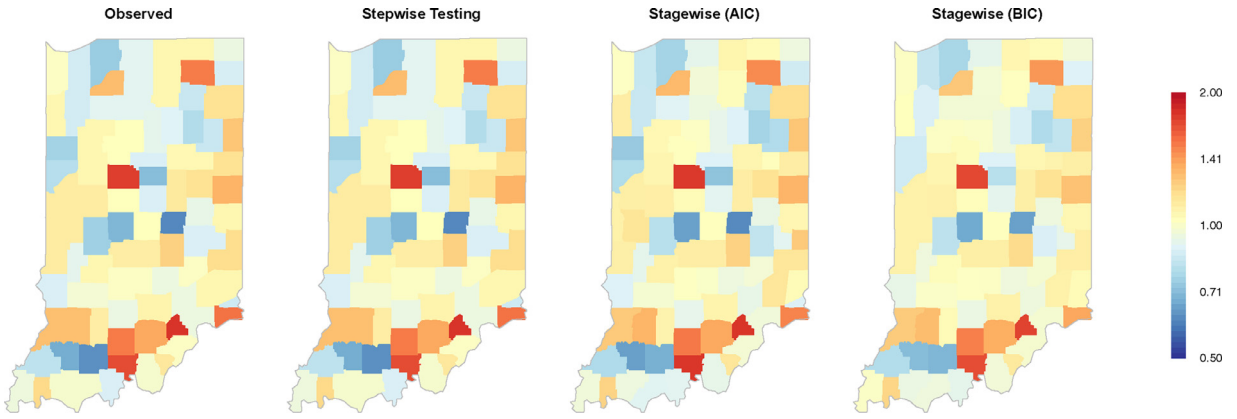

Fig. 10. Maps of the observed poverty rates (relative to fitted values from logistic regression model), and the estimated SIR using stepwise testing, stagewise (AIC) and stagewise (BIC) methods for the Indiana poverty data (year 2000).

proposed to recursively find the location of other clusters conditional on the presence of previously detected clusters. We have proposed three forward stepwise cluster detection methods to detect multiple clusters. Our methods differ from previous sequential methods (Li et al., 2011; Zhang et al., 2010) in two aspects. First, our methods allow for overlapping clusters, while previous methods only consider non-overlapping clusters. Second, our methods iteratively update the expected number of cases for Poisson model and baseline probability for binomial model, and use the standard spatial scan statistic in each iteration. The sequential methods, which are also based on a series of hypothesis testing, need to adjust the spatial scan statistic for multiple clusters in each iteration. Our proposed forward stepwise methods use maximum likelihood ratio test statistic to find the most likely cluster, and select a secondary cluster after updating the expected number of cases or baseline probability for each cell, and do this iteratively. The stepwise testing method stops whenever the most likely cluster in any iteration becomes nonsignificant. The stepwise bias-corrected method identifies the optimal solution by maximizing the bias-corrected log-likelihood. The stepwise information criterion approach is similar to the stepwise bias-corrected method, but tends to be more liberal, detecting more spurious clusters. The stepwise testing method and the stepwise bias-corrected method perform similarly in terms of the accuracy of estimates and the power of detecting the true clusters. Due to computational efficiency, the stepwise testing method is preferred to the stepwise bias-corrected method.

In addition, we developed a forward stagewise approach for spatial cluster detection. We considered a generalized linear regression model in which each column of design matrix corresponds to a potential cluster. A forward stagewise algorithm is applied to the full set of covariates and yields a solution path for either Poisson model or binomial model. We discussed several stopping criteria, including information criteria, the criterion based on bias estimates using Monte Carlo simulations under null model and criteria based on bootstrap estimation of Kullback–Leibler information.

Simulation studies indicate that the stagewise method using BIC as stopping rule usually yields more accurate estimated maps (smaller RAMSE values) and fewer false cluster detections in the background than other methods. It also has a substantially lower computational burden than the bootstrap-based bias correction approaches. In practical applications where one typically expects few true clusters, we recommend the stagewise (BIC) method as the best option for general use.

In the analysis of New York leukemia data, we note that the stagewise (BIC) method is more conservative than other approaches, identifying the primary cluster but missing some previously identified clusters, while stepwise testing and bias-corrected methods find evidence for one additional cluster; other stagewise methods find many additional clusters. In the Indiana poverty data analysis, almost all regions are identified as belonging to clusters using any of the proposed methods. This suggests more general spatial heterogeneity rather than clustering is present. We are working on extensions of these methods that include spatially unstructured random effects in addition to spatial clusters.

Although we use circular potential clusters for illustration, adaptations of clustering model to other shapes such as rectangles, ellipses, or even irregularly shaped clusters are feasible. In our simulation studies, we have focused on scenarios of no clustering and single cluster with increased risk. Results from simulations with multiple clusters were qualitatively similar and are not presented here. To further evaluate the proposed methods, we can extend the simulation studies to reflect more practical scenarios. Some possible extensions may include: (1) replacing circular clusters by rectangles, ellipses or even irregularly shaped clusters; (2) considering multiple clusters with unequal risks. Additional simulation studies under different situations can facilitate our understanding of the limitations of the proposed methods and help us improve them.

A broader extension of the stagewise method is from spatial cluster detection to spatio-temporal cluster detection. In a purely spatial model, a large collection of moving windows centered at the observed locations are considered

as potential clusters. Within the spatio-temporal model, we can consider cylindrical space-time potential clusters, e.g. circular windows during certain time intervals with different pattern of risk from the remainder of the study region or the other time intervals. The extended algorithm for spatio-temporal cluster detection is somewhat more complicated and will be part of our future work.

References

Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control 1974;19(6):716–23.

Cavanaugh J, shumway R. A bootstrap variant of aic for state-space model selection. Stat Sin 1997;7:473–96.

Clark A, Lawson AB. Spatio-temporal cluster modelling of small area health data. In: Lawson AB, Denison D, editors Spatial cluster modelling 2002;154:235–58.

Efron B, Hastie T, Johnstone I, Tibshiranii O. Least angle regression. Ann Stat 2004;32:407–99.

Gangnon RE. Impact of prior choice on local bayes factors for cluster detection. Stat Med 2006;25:883–95.

Gangnon RE, Clayton M. Bayesian detection and modeling of spatial disease clustering. Biometrics 2000;56:922–35.

Gangnon RE, Clayton M. A weighted average likelihood ratio test for spatial clustering of disease. Stat Med 2001;20:2977–87.

Gangnon RE, Clayton M. A hierarchical model for spatially clustered disease rates. Stat Med 2003;22:3213–28.

Gangnon RE, Clayton M. Cluster dection using bayes factors from overparameterized cluster models. Environ Ecol Stat 2007;14:69–82.

Hastie T, Taylor J, Tibshirani R, Wlather G. Forward stagewise regression and the monotone lasso. Electron J Stat 2007;1:1–29.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning; data mining, inference and prediction. New York: Springer Verlag; 2001.

Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika 1989;76:297–307.

Ishiguro M, Sakamoto Y. WIC: An estimation-free information criterion. Tokyo: Research Memorandum, Institute of Statistical Mathematics; 1991.

Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods 1997;26(6):1481–96.

Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. Stat Med 2006;25:3929–43.

Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med 1995;14:799–810.

Lawson AB. Cluster modelling of disease incidence via rjmcmc methods: a comparative evaluation. Stat Med 2000;19:2361–75.

Li XZ, Wang JF, Yang WZ, Li ZJ, Lai SJ. A spatial scan statistic for multiple clusters. Math Biosci 2011;233(2):135–42.

Schwarz GE. Estimating the dimension of a model. Ann Stat 1978;6:461–4.

Shibata R. Bootstrap estimate of kullback–leibler information for model selection. Stat Sin 1997;7:375–94.

Shu L, Jiang W, Tsui KL. A standardized scan statistic for detecting spatial clusters with estimated parameters. Naval Res Logist 2012;59(6):397–410.

Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 2005;4(1):11.

Wakefield J, Kim A. A bayesian model for cluster detection. Biostatistics 2013;14 (4):752–65.

Wakefield JC, Kelsall J, Morris S. Spatial epidemiology - methods and applications. Oxford: Oxford University Press; 2000.

Waller LA, Turnbull BW, Clark L, Nasca P. Spatial pattern analyses to detect rare disease clusters. Wiley, New York; 1994.

Yan P, Clayton MK. A cluster model for space-time disease counts. Stat Med 2006;25:867–81.

Zhang Z, Assunção R, Kulldorff M. Spatial scan statistic adjusted for multiple clusters. J Probab Stat 2010;2010. Article ID: 642379, 11 pages.