Environ Ecol Stat (2010) 17:55–71
DOI 10.1007/s10651-008-0101-0

Local multiplicity adjustments for spatial cluster detection

Ronald E. Gangnon

Received: 19 October 2007 / Revised: 2 May 2008 / Published online: 13 February 2009

Abstract The spatial scan statistic is a widely applied tool for cluster detection. The spatial scan statistic evaluates the significance of a series of potential circular clusters using Monte Carlo simulation to account for the multiplicity of comparisons. In most settings, the extent of the multiplicity problem varies across the study region. For example, urban areas typically have many overlapping clusters, while rural areas have few. The spatial scan statistic does not account for these local variations in the multiplicity problem. We propose two new spatially-varying multiplicity adjustments for spatial cluster detection, one based on a nested Bonferroni adjustment and one based on local averaging. Geographic variations in power for the spatial scan statistic and the two new statistics are explored through simulation studies, and the methods are applied to both the well-known New York leukemia data and data from a case–control study of breast cancer in Wisconsin.

Keywords Bonferroni · Case–control data · Case-count data · Likelihood ratio test · Score test · Spatial scan statistic

1 Introduction

Cluster detection, i.e., the identification of small regions of elevated disease risk, is a major problem in spatial epidemiology. Cluster detection methods can be distinguished from general clustering methods by the evaluation of specific locations for clusters and

R. E. Gangnon (✉)
Departments of Biostatistics and Medical Informatics and Population Health Sciences,
603 WARF Office Building, University of Wisconsin–Madison,
610 Walnut Street, Madison, WI 53726, USA
e-mail: ronald@biostat.wisc.edu

from focused clustering methods by the lack of a pre-specified source of possible elevated risk (Besag and Newell 1991). Cluster detection methods are typically based on a hypothesis testing paradigm, although some modeling approaches are now available (Gangnon and Clayton 2000, 2003; Lawson 2006; Hossain and Lawson 2006).

The spatial scan statistic (Kulldorff and Nagarwalla 1995; Kulldorff 1997) is one of the most widely used cluster detection methods. The spatial scan 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, typically circular clusters centered at the observed locations.

The operating characteristics of the spatial scan statistic vary across the study region (Gangnon and Clayton 2001, 2004). Under the null hypothesis of constant disease risk, the spatial scan statistic is more likely to select clusters in (urban) areas with fine geographic resolution (i.e. geographically smaller regions or more densely packed points) than clusters in (rural) areas with coarse geographic resolution, due to the larger number of potential clusters for a fixed radius size in such regions. We refer to this variation in the numbers of potential clusters across different portions of the study region as the local multiplicity problem. As a consequence of the local multiplicity problem, in many typical situations, there may be an overstatement of the evidence for clustering in urban areas and an understatement of the evidence for clustering in rural areas. Although one might be able to address this by restricting attention to circular clusters with a fixed grid of centers and radii, we instead propose the locally adjusted spatial scan (LASS) statistic, which incorporates a local multiplicity adjustment to provide a more balanced assessment of the evidence for clustering.

The weighted average likelihood ratio (WALR) statistic (Gangnon and Clayton 2001) and the weighted average likelihood ratio scan (WALRS) statistic (Gangnon and Clayton 2004) are alternatives to the spatial scan statistic. The WALR statistic is the weighted average of the likelihood ratios for all clusters; the WALRS statistic is the maximum, over all locations, of the WALR statistics for all clusters containing the location. These statistics require the specification of weights for each cluster. The weights could be used to incorporate prior information about cluster locations. With no prior information about cluster locations, weights based on an approximately uniform selection of a cluster have been recommended (Gangnon and Clayton 2001, 2004). However, it is unclear how to generalize these weights to collections of non-circular clusters or point location data. In this paper, we propose the local average likelihood ratio scan (LALRS) statistic, an unweighted version of the WALRS statistic, which is applicable in any setting.

Power comparisons of the spatial scan statistic, the WALR statistic and the WALRS statistic have been reported (Gangnon and Clayton 2001, 2004). These studies demonstrate that the spatial scan statistic has high power in areas with fine geographic resolution and low power in areas with coarse geographic resolution. However, these comparisons were limited to a small number of scenarios in which variations in cluster locations were combined with variations in other cluster properties, e.g., population and relative risk. In this paper, we present more comprehensive simulation studies which isolate the impact of cluster location on power by moving the center of a cluster of fixed population and risk across the study regions.

The paper is organized as follows. In Sect. 2, we discuss the methods for assessing the evidence for a single cluster using either the Poisson model for regional count data or the Bernoulli model for point location (case–control) data. In Sect. 3, we present the spatial scan statistic and propose the LASS and LALRS statistics. In Sect. 4, we evaluate the geography of power for these cluster detection methods using the spatial structures of the well-known New York leukemia data set and the Dane County (WI) breast cancer data. In Sect. 5, we present analyses of these data sets using the LASS, LALRS and spatial scan statistics. In Sect. 6, we provide some concluding remarks.

2 Available data and models

2.1 Point location (case–control) data

Consider the situation in which we observe the point locations of a sample of N cases and controls. The available data consist of $(y_i, \mathbf{x}_i, p_{i0})_{i=1}^N$, where y_i is the case–control indicator for subject i, $\mathbf{x}_i = (x_{1i}, x_{2i})$ is the location of subject i and p_{i0} is the baseline probability that subject i is a case. The baseline probability of being a case, p_{i0}, may incorporate covariate effects (e.g., age) or may simply reflect the overall proportion of cases. We model the y_i as independent Bernoulli random variables with mean p_i.

For any subset $\mathbf{Z}$ of the study region, we consider the following model: $\text{logit}(p_i) = \text{logit}(p_{i0}) + \alpha_{\mathbf{Z}} + \theta_{\mathbf{Z}}\delta_{\mathbf{Z}}(\mathbf{x_i})$, where $\delta_{\mathbf{Z}}(\mathbf{x_i}) = 1$ if $\mathbf{x}_i \in \mathbf{Z}$ and $\delta_{\mathbf{Z}}(\mathbf{x_i}) = 0$ otherwise, $\alpha_{\mathbf{Z}}$ is the disease risk for locations outside $\mathbf{Z}$ and $\theta_{\mathbf{Z}}$ is the log odds ratio for locations inside $\mathbf{Z}$. If $\mathbf{Z}$ is not a cluster, $\theta_{\mathbf{Z}} = 0$; if $\mathbf{Z}$ is a cluster, $\theta_{\mathbf{Z}} \neq 0$.

The evidence in favor of $\mathbf{Z}$ as a cluster is given by the likelihood ratio test statistic for $H_0 : \theta_{\mathbf{Z}} = 0$ versus $H_A : \theta_{\mathbf{Z}} \neq 0$. There is no closed-form solution for this statistic, but it can be found using iterative optimization algorithms. This is impractical for large-scale cluster detection. So, we use a 1-step Taylor series quadratic approximation to the log-likelihood to obtain the following approximation to the likelihood ratio test statistic

$$\text{LR}_{\mathbf{Z}} = \exp\left\{\frac{1}{2}[y(\mathbf{Z}) - E(\mathbf{Z})]^2\left[\frac{1}{V(\mathbf{Z})} + \frac{1}{V_{tot} - V(\mathbf{Z})}\right]\right\},$$

where $y(\mathbf{Z}) = \sum_{i=1}^N y_i\delta_{\mathbf{Z}}(\mathbf{x_i})$ is the number of cases inside $\mathbf{Z}$, $E(\mathbf{Z}) = E[y(\mathbf{Z})|H_0] = \sum_{i=1}^N p_{i0}\delta_{\mathbf{Z}}(\mathbf{x_i})$, $V(\mathbf{Z}) = \text{Var}[y(\mathbf{Z})|H_0] = \sum_{i=1}^N p_{i0}(1-p_{i0})\delta_{\mathbf{Z}}(\mathbf{x_i})$, and $V_{tot} = \sum_{i=1}^N p_{i0}(1-p_{i0})$. Using standard asymptotic results, we can also obtain the nominal p-value $p_{\mathbf{Z}} = P(X^2 > 2\log \text{LR}_{\mathbf{Z}})$ where X^2 is a χ_1^2 random variate.

2.2 Regional (case-count) data

In many situations, we only observe aggregated data for N administrative regions (e.g., counties, census tracts, ZIP codes) within the study area. The available data now consist of $(y_i, E_i, \mathbf{x}_i)_{i=1}^N$, where y_i is the number of cases of disease in region i,

E_i is the expected number of cases of disease in region i and $\mathbf{x}_i = (x_{1i}, x_{2i})$ is the geographic centroid of region i. The expected number of cases, E_i, may incorporate covariate effects (e.g., age) or may simply reflect the overall disease rate applied to the regional population. In any case, without loss of generality, we assume that the E_i have been internally standardized so that $\sum_{i=1}^{N} E_i = \sum_{i=1}^{N} y_i$. We assume that y_i are independent Poisson random variables with mean $\rho_i E_i$.

For any subset $\mathbf{Z}$ of the study region, we consider the following model for ρ_i: $\log(\rho_i) = \alpha_{\mathbf{Z}} + \theta_{\mathbf{Z}} \delta_{\mathbf{Z}}(\mathbf{x_i})$, where $\delta_{\mathbf{Z}}(\mathbf{x_i}) = 1$ if $\mathbf{x}_i \in \mathbf{Z}$ and $\delta_{\mathbf{Z}}(\mathbf{x_i}) = 0$ otherwise, $\alpha_{\mathbf{Z}}$ is the disease risk for locations outside $\mathbf{Z}$ and $\theta_{\mathbf{Z}}$ is the relative risk for locations inside $\mathbf{Z}$. For rare diseases, this model is an aggregated version of the Bernoulli model. If $\mathbf{Z}$ is not a cluster, $\theta_{\mathbf{Z}} = 0$; if $\mathbf{Z}$ is a cluster, $\theta_{\mathbf{Z}} \neq 0$.

The evidence in favor of $\mathbf{Z}$ as a cluster is given by the likelihood ratio test statistic for $H_0 : \theta_{\mathbf{Z}} = 0$ versus $H_A : \theta_{\mathbf{Z}} \neq 0$,

$$\mathrm{LR}_{\mathbf{Z}} = \left\{ \frac{y(\mathbf{Z})}{E(\mathbf{Z})} \right\}^{y(\mathbf{Z})} \left\{ \frac{y_{tot} - y(\mathbf{Z})}{E_{tot} - E(\mathbf{Z})} \right\}^{y_{tot} - y(\mathbf{Z})}$$

where $y(\mathbf{Z}) = \sum_{i=1}^{N} y_i \delta_{\mathbf{Z}}(\mathbf{x_i})$ is the number of cases inside $\mathbf{Z}$, $E(\mathbf{Z}) = \sum_{i=1}^{N} E_i \delta_{\mathbf{Z}}(\mathbf{x_i})$ is the expected number of cases inside $\mathbf{Z}$, $y_{tot} = \sum_{i=1}^{N} y_i$, and $E_{tot} = \sum_{i=1}^{N} E_i$. Using standard asymptotic results, we can also obtain the nominal p-value $p_{\mathbf{Z}} = P(X^2 > 2 \log \mathrm{LR}_{\mathbf{Z}})$ where X^2 is a χ^2_1 random variate.

2.3 Potential clusters

In cluster detection, we consider a large collection of subsets of the study region as potential clusters. Here, we select circular regions centered at the observed locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, with radii ranging from 0 up to a fixed maximum radius, $r_{\max}$. To identify the m_s unique clusters centered at $\mathbf{x}_s$ for $s = 1, 2, \ldots, N$, we let $0 = r_{s,1} < r_{s,2} < \ldots < r_{s,m_s} \leq r_{max}$ be the (unique) ordered distances from $\mathbf{x}_s$ to all locations, truncated at r_{max}. We denote the circular cluster centered at $\mathbf{x}_s$ with radius r_{st} by $\mathbf{Z}_{st}$, the associated likelihood ratio test statistic by LR_{st}, and the associated nominal p-value by p_{st} for $t = 1, 2, \ldots, m_s$; $s = 1, 2, \ldots, N$. Although we use a specific set of potential clusters, we note that the methods described here are quite general and could be applied to any discrete set of potential clusters.

3 Global test statistics

The evidence for clustering is quantified by a global p-value, the probability of observing a collection of likelihood ratio test statistics at least as extreme as observed. A major difficulty is the lack of a single natural ordering of the sample space and hence of a uniformly most powerful test statistic. In this paper, we consider several different global test statistics. For each statistic, we calculate the p-value using its distribution under the null hypothesis of constant disease risk. For Bernoulli data without covariates or Poisson data, the null distribution, conditional on y_{tot} and the observed locations, is

free of unknown parameters and easily simulated (hypergeometric and multinomial, respectively). For Bernoulli data with covariates, the null distribution, conditional on y_{tot} and the observed locations of the cases and controls, is unfamiliar and difficult to simulate, so data $y_1^*, y_2^*, \ldots, y_N^*$ at each location are simulated as independent Bernoulli random variables with success probabilities $p_{10}, p_{20}, \ldots, p_{N0}$.

3.1 The spatial scan statistic

Kulldorff and Nagarwalla (1995) and Kulldorff (1997) proposed the spatial scan statistic, i.e., the maximum likelihood ratio test statistic over all potential clusters $\mathrm{LR}_{\max} = \max_{s,t} \mathrm{LR}_{st}$, as a global cluster detection test. In addition to a global p-value, one can also obtain adjusted p-values for each potential cluster by comparing LR_{st} with the simulated null distribution of $\mathrm{LR}_{\max}$.

3.2 The locally adjusted spatial scan (LASS) statistic

Under the null hypothesis of constant disease risk, the most likely cluster is more likely to be a cluster from an area with fine geographic resolution, e.g. geographically smaller regions or more densely packed points, than clusters in areas with coarse geographic resolution, due to the larger number of potential clusters for a fixed radius size in such regions. As a consequence, in many typical situations, the spatial scan statistic may overstate the evidence for clustering in urban areas and understate the evidence for clustering in rural areas.

To account for these local differences in the numbers of overlapping clusters, e.g. the local multiplicity problem, we propose a two-stage procedure. First, clusters are divided into mutually exclusive and exhaustive groups with variable numbers of clusters in each group. To capture local variations in the multiplicity problem, clusters within groups must have significant overlap. A Bonferroni adjustment is performed within each group. A simulation-based multiplicity adjustment is then applied across groups.

To accomplish this two-stage adjustment, the potential clusters are first partitioned into N groups, one for each observed location. Clusters are assigned to a group by selection of a random location inside each cluster. Since, by construction, all clusters within a group contain the same location, there will be substantial overlap within each group. Groups based on locations in dense urban areas which belong to many clusters will be large while groups based on locations in sparse rural areas which belong to few clusters will generally be small. We denote the assigned group for cluster $\mathbf{Z}_{st}$ by $g(\mathbf{Z}_{st})$ and the number of clusters assigned to group g by $m[g]$. By multiplying the nominal p-value for cluster $\mathbf{Z}_{st}$ (p_{st}) by both the number of groups (N) and the number of clusters within the group ($m[g(\mathbf{Z}_{st})]$), we obtain the two-stage Bonferroni adjusted p-value for cluster $\mathbf{Z}_{st}$, $Nm[g(\mathbf{Z}_{st})]p_{st}$.

These adjusted p-values are dependent on the random assignment of the clusters to groups. To avoid ambiguity, we use the average adjusted p-value over all possible random assignments of clusters to groups. That is, we replace $m[g(\mathbf{Z}_{st})]$ with its

expected value

$$M_{st} = \mathrm{E}\{m[g(\mathbf{Z}_{st})]\} = \sum_{g=1}^{N} M_g \delta_{\mathbf{Z}_{st}}(\mathbf{x}_g)/|\mathbf{Z}_{st}| + (1 - 1/|\mathbf{Z}_{st}|),$$

where $M_g = \sum_{s,t} \delta_{\mathbf{Z}_{st}}(\mathbf{x}_g)/|\mathbf{Z}_{st}|$ is the expected number of clusters in group g for $g = 1, 2, \ldots, N$ and $|\mathbf{Z}_{st}|$ is the number of locations inside cluster $\mathbf{Z}_{st}$. The adjusted p-value for cluster $\mathbf{Z}_{st}$ is $NM_{st}\,p_{st}$.

We call the minimum adjusted p-value across all potential clusters, $p_{min}^{adj} = \min_{s,t} NM_{st}\,p_{st}$, the locally adjusted spatial scan (LASS) statistic. We use the LASS statistic as a global test statistic. The adjustment factors M_{st} account for local variations in the numbers and/or overlap of potential clusters. In addition to a global p-value, as with the spatial scan statistic, one can also obtain cluster-specific adjusted p-values by comparing $NM_{st}\,p_{st}$ with the simulated null distribution of p_{min}^{adj}.

3.3 The local average likelihood ratio scan (LALRS) statistic

As an alternative method for accounting for local multiplicity, we propose the local average likelihood ratio scan (LALRS) statistic. The LALRS statistic is also based on a two-stage procedure. First, for each location $\mathbf{x}_k$, we calculate the average of the likelihood ratios associated with clusters containing $\mathbf{x}_k$, i.e., the local average likelihood ratio $\mathrm{LALR}(\mathbf{x}_k) = \sum_{s,t} \mathrm{LR}_{st}\delta_{\mathbf{Z}_{st}}(\mathbf{x}_k)/\sum_{s,t}\delta_{\mathbf{Z}_{st}}(\mathbf{x}_k)$. Local averages for locations in dense (urban) areas will be based on a large number of clusters, while local averages for locations in sparse (rural) areas will be based on a small number of clusters. The global LALRS statistic is the maximum, over all locations, of these LALR statistics, $\mathrm{LALRS} = \max_k \mathrm{LALR}(\mathbf{x}_k)$.

The LALRS statistic is related to two other cluster detection statistics, the weighted average likelihood ratio (WALR) statistic (Gangnon and Clayton 2001) and the (local) weighted average likelihood ratio scan (WALRS) statistic (Gangnon and Clayton 2004). The WALR statistic is the weighted average of the likelihood ratios across all clusters, $\mathrm{WALR} = \sum_{s=1}^{N}\sum_{t=1}^{m_s} w_{st}\mathrm{LR}_{st}$, where $w_{st} \geq 0$ is a known weight associated with cluster $\mathbf{Z}_{st}$ and $\sum_{s,t} w_{st} = 1$. It can be motivated as an approximation to the marginal likelihood ratio (or Bayes factor) for the composite one cluster model relative to the model with no clusters (viewing the weights as a prior distribution).

The WALRS statistic substitutes a weighted average for the unweighted LALR used in the LALRS statistic, i.e., $\mathrm{WALR}(\mathbf{x}_k) = \sum_{s,t} w_{st}\mathrm{LR}_{st}\delta_{\mathbf{Z}_{st}}(\mathbf{x}_k)/\sum_{s,t} w_{st}\delta_{\mathbf{Z}_{st}}(\mathbf{x}_k)$. The WALRS statistic is the maximum of these local WALR statistics, $\mathrm{WALRS} = \max_k \mathrm{WALR}(\mathbf{x}_k)$. The LALRS statistic is a special case of the WALRS statistic with all weights set to 1 (alternate representations of the same collection of cells/points are only counted once).

Location-specific WALR (or LALR) statistics can be motivated as an approximate Bayes factor comparing a composite one cluster model in which the cluster contains $\mathbf{x}_k$ to the no cluster model. As such, the location-specific WALR statistics are similar to the local Bayes factors for clustering proposed by Gangnon and Clayton (2007). Gangnon (2006) demonstrated that local Bayes factors are robust to the prior distribution for cluster locations. Thus, we expect little difference between location-specific WALR and LALR statistics and hence little difference between the global WALRS and LALRS statistics, since the only difference between the statistics is the choice of weights (priors).

The WALR and WALRS statistics can be used with any weights. Gangnon and Clayton (2001, 2004) recommended the following weights for circular clusters with regional data:

$$w_{st} = \frac{a_s}{A} \frac{r_{s,t+1} - r_{s,t}}{r_{\max}},$$

where a_s is the area of the region s, $A = \sum_{s=1}^{N} a_s$, and $r_{s,m_s+1} = r_{max}$. These weights were motivated by a spatially uniform selection of a circle adjusted to avoid empty circles. The generalization of these weights to point location data or other sets of clusters is not straightforward and, to some extent, motivated the development of the LALRS statistic. The LALRS statistic depends only on the cluster memberships and is easily applied to both regional and point location data and any enumerated set of clusters, including the flexible spatial scan statistic (Tango and Takahashi 2005) and the elliptic spatial scan statistic (Kulldorff et al. 2006).

4 Inference for cluster locations: maps of location-specific p-values

For the spatial scan statistic, inferences about cluster locations are typically based on displays of the most likely cluster and any non-overlapping, statistically significant secondary clusters. The identified cluster(s) are properly viewed as approximate cluster locations, since there will typically be a large number of clusters that overlap the identified cluster(s) with similar evidence for clustering (e.g. similar likelihood ratio test statistics). Here, we propose the use of location-specific adjusted p-values as a method for identifying the locations of the most likely cluster and any secondary clusters and conveying some of the uncertainty about the exact composition of the clusters.

For the spatial scan statistic, location-specific adjusted p-values can be obtained by comparing location-specific scan ($\mathrm{LR}_{max}(\mathbf{x}) = \max\{\mathrm{LR}_{st} : \mathbf{x} \in \mathbf{Z}_{st}\}$) with the simulated null distribution of the spatial scan statistic (LR_{max}). Unlike the complete set of cluster-specific p-values, location-specific adjusted p-values can be easily displayed on a grayscale map.

The comparison of the location-specific scan statistics with the null distribution of the global spatial scan statistic serves two purposes. First, the adjusted p-value for any location in the most likely cluster will be the same as the global p-value. Thus, a grayscale map of the adjusted p-values will convey both the location and significance

of the most likely cluster. It similarly conveys the location and significance of any non-overlapping secondary clusters. Second, the representation of the location-specific scan statistics in terms of location-specific adjusted p-values provides a simple, intuitive calibration of the scan statistics.

For example, adjusted p-values of 0.002 and 0.004 (or 0.02 and 0.04) for locations in the most likely cluster and an adjacent location would indicate roughly similar evidence for a secondary cluster which includes the second location and the most likely cluster. Here, we might consider it plausible that the second location belongs to the true cluster On the other hand, adjusted p-values of 0.002 and 0.02 (or 0.02 and 0.24) would indicate much less evidence for a secondary cluster which includes the second location. Here, we might consider it unlikely that the second location belongs to the true cluster. In this manner, the location-specific p-values provide an informal assessment of the plausible extent of each cluster. Similar objectives could be achieved using displays of the location-specific scan statistics. However, without the p-value calibration, it would be more difficult to assign an interpretation to observed differences in the location-specific values.

Location-specific adjusted p-values can also be obtained for the LASS statistic (comparing $p_{min}^{adj}(\mathbf{x}) = \min\{NM_{st}p_{st} : \mathbf{x} \in \mathbf{Z}_{st}\}$ with the simulated null distribution of p_{min}^{adj}), the LALRS statistic (comparing $\mathrm{LALR}(\mathbf{x}_k)$ with the simulated null distribution of LALRS) and the WALRS statistic (comparing $\mathrm{WALR}(\mathbf{x}_k)$ with the simulated null distribution of WALRS).

5 Simulation results: geographic variations in power

In this section we evaluate the performance of these tests in terms of power to detect specified clusters via simulation. Simulations were conducted using the structures of two different data sets: the New York leukemia data and the Dane County (WI) breast cancer data. For these simulations we introduced a single circular cluster centered at each observed location. The cluster incorporated a fixed number of subjects and elevation in risk. Thus, variations in power will reflect the impact of geographic resolution on test performance.

Waller et al. (2006) performed a similar evaluation of the geography of power for the spatial scan statistic and Tango's index of clustering (Tango 1995). Their simulations used a fixed geographic cluster size rather than a fixed population size. Hence, their maps of power emphasize the impact of local sample size, whereas our maps emphasize the impact of local variations in geographic resolution (and consequent local variations in numbers of overlapping potential clusters).

In this simulation study, we determined power as the proportion of simulations in which the global null hypothesis is rejected at the 5% level. We did not assess the accuracy of the location of the identified cluster for two reasons. First, it is not obvious how to define a correct identification of a cluster. For the spatial scan statistic alone, several definitions of correct cluster identifications have been used, e.g., rejections in which the center of the most likely cluster is the true cluster center, rejections in which the most likely cluster contains the true cluster center (Waller et al. 2006) and rejections in which the most likely cluster overlaps the true cluster (Gangnon and Clayton 2004).

Second, and more importantly, incorporating correct cluster identifications has little, if any, impact on comparisons of performance, either for different clusters using the same test statistic (Gangnon and Clayton 2004; Waller et al. 2006) or for the same cluster using different test statistics (Gangnon and Clayton 2004).

5.1 Regional data: New York leukemia data

The New York leukemia data set (Waller et al. 1994) describes leukemia incidence between 1978 and 1982 in eight counties in upstate New York. The two largest cities in the study region are Syracuse in the north-central portion of the study region and Binghamton in the south-central portion of the study region. The eight-county region is divided into 790 cells, either census blocks or census tracts, for which the population at risk, count of incident leukemia cases, and geographic centroid are available. The reported centroids of two cells are identical, so those two cells are merged for analysis. Cell areas are obtained from the Dirichlet tessellation of the centroids (Gangnon and Clayton 2000). The study region is approximately 136 km from north to south and 115 km from east to west. Figure 1 displays the observed leukemia incidence rates.

The null distributions of the test statistics were obtained by simulating 10,000 data sets from the multinomial distribution, conditional on the total of 592 cases. For each alternative, we simulated 1,000 data sets. We considered circles centered at cell centroids with radii no greater than 20 km as potential clusters. The choice of maximum

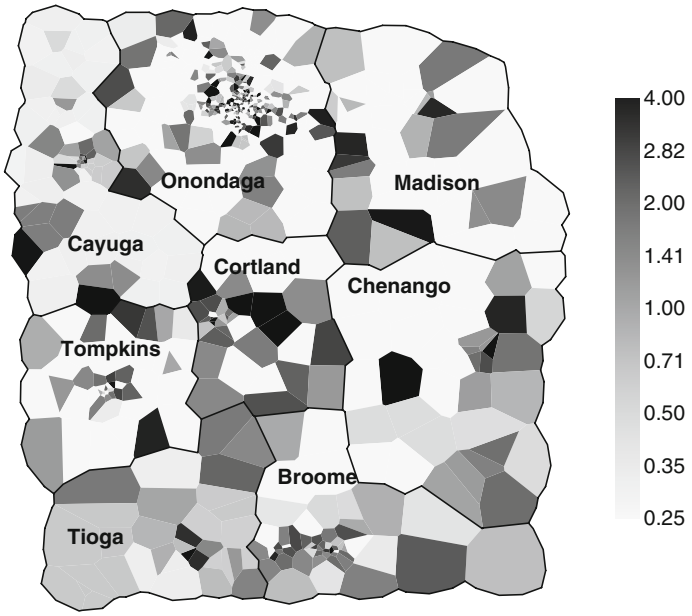

Fig. 1 Map of the observed 5-year leukemia incidence rates (relative to the overall rate of 5.5 per 10,000 persons) for the New York data. Relative rates above 4 and below 0.25 are displayed in black and light gray, respectively. Regions are based on the Dirichlet tessellation of the cell centroids. County borders and names are given in black

radius is arbitrary, but seems reasonable given the size of the region. We considered 789 single cluster alternatives, which consisted of a circular cluster of 30,000 subjects (roughly 3% of the total population) with a relative risk of 2 centered at one of the cell centroids. The relative risk of 2 was chosen based on a crude power calculation (approximately 80% power for a nominal 1% level test). Power calculations were based on 5% level tests.

Maps of power to detect the standard cluster are provided in Fig. 2. The maps for the WALR, WALRS and LALRS statistics are quite similar, with low power in the two large cities, Binghamton and Syracuse, and high power elsewhere. The mean power for the 457 clusters centered in Onondaga county (Syracuse) is 25% (WALR), 23% (WALRS) and 26% (LALRS). The mean power for the 55 clusters centered in Broome county (Binghamton) is 32% (WALR), 31% (WALRS) and 36% (LALRS). The mean power for the 477 clusters centered elsewhere is 48% (WALR), 47% (WALRS) and 49% (LALRS). The similar performance of the WALRS and LALRS statistics agrees with the robustness of local Bayes factors to prior specification observed by Gangnon (2006). We recommend the use of the LALRS statistic in place of either the WALR or WALRS statistic, since it does not require weights.

The maps of power for the LALRS, LASS and (spatial) scan statistics are quite different. The scan statistic has high power in Syracuse, while the LALRS statistic has the low power. The LALRS statistic has the high power in rural areas (outside Broome and Onondaga counties), while the scan statistic has the low power.

The mean power for the clusters centered within each county for these three statistics is provided in Table 1. The LALRS statistic has the highest power in 6 of the 8

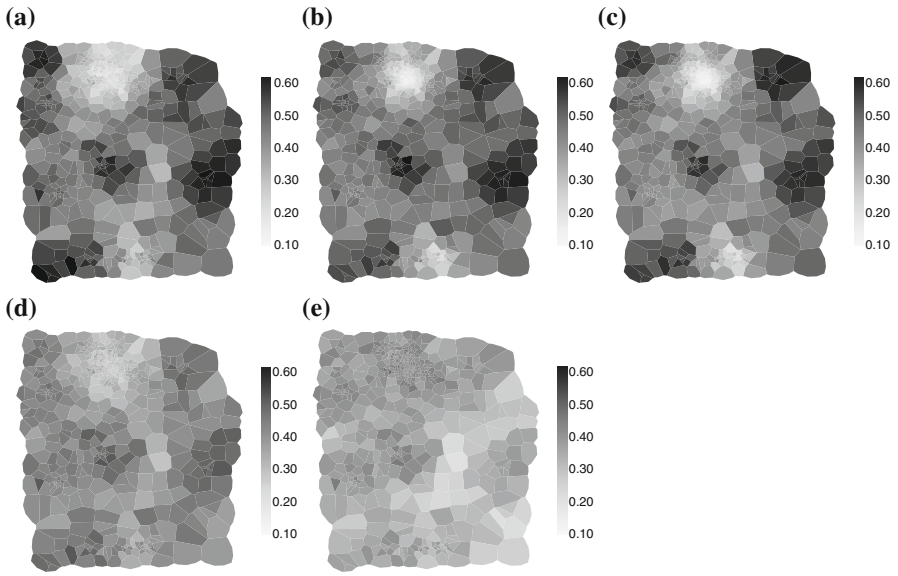

Fig. 2 Estimated power of the **a** LALRS, **b** WALR, **c** WALRS, **d** LASS and **e** spatial scan statistics to detect a circular cluster of 30,000 persons with a relative risk of 2 centered at each of the 789 cell centroids in the New York data using a 5% level test. Estimated power is based on 1,000 simulations

Table 1 Mean power for clusters centered within each county for the scan, LASS and LALRS statistics with $r_{max} = 20$ and $r_{max} = 40$

	$r_{max} = 20$			$r_{max} = 40$		
	Scan (%)	LASS (%)	LALRS (%)	Scan (%)	LASS (%)	LALRS (%)
Cayuga	38.4	42.9	49.1	38.5	40.1	35.2
Onondaga	41.8	31.0	26.1	40.0	33.8	26.8
Madison	34.6	41.0	48.9	35.9	40.3	37.9
Tompkins	36.9	42.1	46.3	36.0	41.4	40.0
Cortland	37.5	44.4	50.6	38.7	42.7	40.8
Chenango	30.6	43.5	50.6	36.3	46.8	52.7
Tioga	34.2	43.3	48.3	34.8	44.2	43.0
Broome	32.8	38.3	36.3	32.0	39.5	34.1

counties, but very low power in Onondaga county. The scan statistic has the highest power in Onondaga county, but the lowest power in the remaining 7 counties. Although the LASS statistic has the highest power for just one county (Broome), it has higher power than the scan statistic for all counties except Onondaga and more balanced power across the entire study region than the LALRS statistic.

To evaluate the sensitivity of these findings to the choice of maximum radius for potential clusters, the simulation study was repeated using a maximum radius of 40 km. The mean power for clusters centered within each county for the scan, LASS and LALRS statistics from these simulations is provided in Table 1. The LASS statistic now has the highest power in 6 of the 8 counties, with higher power for the LALRS statistic only in Chenango county and higher power for the scan statistic only in Onondaga county.

5.2 Point location data: Dane county (WI) breast cancer data

The Dane county breast cancer data set is a subset of a larger population-based case–control study of breast cancer in Wisconsin (Newcomb et al. 1994, 1999). Dane county is the second largest county, by population, in Wisconsin and consists of an urban/suburban center, the city of Madison and its suburbs, and a rural/small town periphery. The Dane county data set consists of 471 incident breast cancer cases and 451 control subjects. Cases and controls were geocoded to their residence (McElroy et al. 2003). For each subject, the available data are the case/control status indicator, the geographic location and fitted values from a logistic regression model including known breast cancer risk factors (age, parity, age at menarche, education, family history of breast cancer, alcohol use and body mass index) (Dumitrescu and Cotarla 2005). The study region is approximately 50 km from north to south and 76 km from east to west. Figure 3 displays the point locations of cases and controls.

The null distributions of the test statistics were obtained by simulating 10,000 data sets from the Bernoulli distribution using the fitted values from the logistic regression.

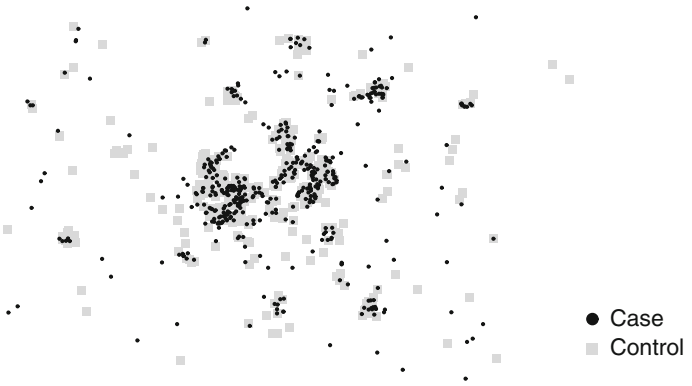

Fig. 3 Map of geocoded locations of breast cancer cases and controls for the Dane county data

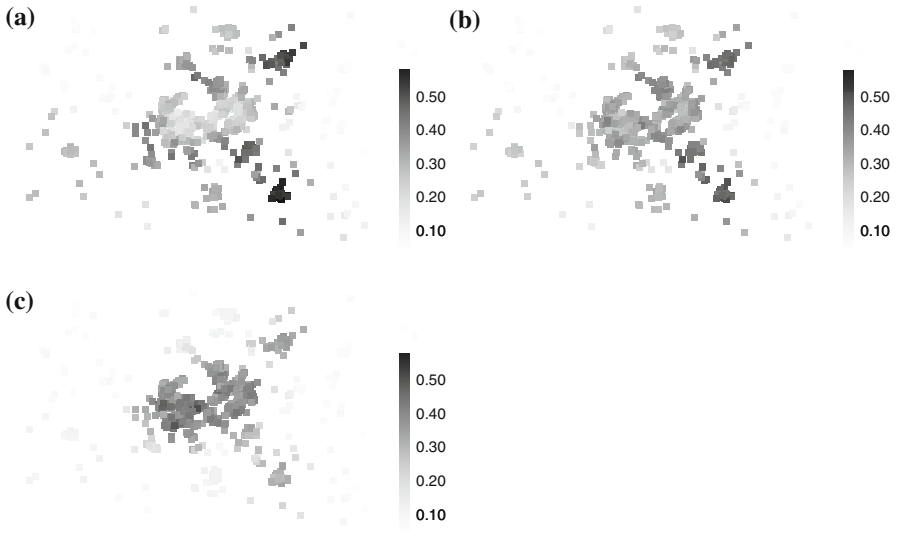

Fig. 4 Estimated power of the **a** LALRS, **b** LASS and **c** spatial scan statistics to detect a circular cluster of 30 persons with an odds ratio of 4 centered at each of the 908 unique locations in the Dane county data using a 5% level test. Estimated power is based on 1,000 simulations

For each alternative, we simulated 1,000 data sets. We considered circles centered at the observed case/control locations with radii no greater than 5 km as potential clusters. As noted earlier, the choice of maximum radius is arbitrary. We considered 908 single cluster alternatives, which consisted of a circular cluster of 30 subjects (roughly 3% of the study population) with an odds ratio of 4 centered at one of the observed locations. The odds ratio of 4 was chosen based on a crude power calculation (approximately 80% power for a nominal 1% level test). Power calculations were based on 5% level tests.

Maps of power to detect the standard cluster are provided in Fig. 4. Overall, the patterns are similar to those observed with the New York data. The scan statistic has

high power in the city of Madison, while the LALRS statistic has the low power. The LALRS statistic has the high power in the suburban and rural areas, while the scan statistic has low power. The power of the LASS statistic generally falls between the power for the other two statistics; it is also more consistent across the region. In sparsely populated areas in the east and northwest, none of the tests has much power. The mean power for clusters centered at the 457 locations in the city of Madison is 39% (scan), 33% (LASS) and 24% (LALRS). The mean power for clusters centered at the 317 locations in other cities or villages in Dane county is 25% (scan), 31% (LASS) and 31% (LALRS). The mean power for clusters centered at the 148 rural locations is 18% (scan), 25% (LASS) and 27% (LALRS).

6 Data analyses

6.1 New York leukemia data

We assessed the evidence for clustering in the New York leukemia data using the LASS, LALRS and (spatial) scan statistics using circular clusters centered at the cell centroids with radii less than or equal to 20 km. Cell-specific adjusted p-values for each statistic based on 10,000 Monte Carlo simulations are displayed in Fig. 5.

All three statistics identify very strong evidence for a cluster of elevated risk in the city of Binghamton in Broome county (LASS: $p = 0.0006$, LALRS: $p = 0.0004$, scan: $p = 0.0017$). All three statistics also identify evidence for a secondary cluster of lowered risk north of Syracuse in Onondaga county (LASS: $p = 0.048$, LALRS: $p = 0.0097$, scan: $p = 0.017$) and a secondary cluster of elevated risk in Cortland county (LASS: $p = 0.022$, LALRS: $p = 0.011$, scan: $p = 0.048$).

Although all three statistics find evidence for clusters in the same areas, differences in the p-values tend to reflect interactions with the geography observed previously. For example, the spatial scan statistic identifies less evidence (higher p-values) for the clusters in Binghamton and in Cortland county than the other two statistics. Both the spatial scan statistic and the LALRS statistic identify more evidence for clusters north of Syracuse than the LASS statistic, albeit in two distinct areas. The spatial scan statistic identifies the strongest evidence for clustering in urban/suburban cells near Syracuse, while the LALRS statistic identifies the strongest evidence for clustering in rural cells near the border with Cayuga county.

6.2 Dane county breast cancer data

We assessed the evidence for clustering in the Dane county breast cancer data using the LASS, LALRS and (spatial) scan statistics using circular clusters centered at the cell centroids with radii less than or equal to 5 km. Location-specific adjusted p-values for each statistic based on 10,000 Monte Carlo simulations are displayed in Fig. 6.

There is little evidence for clustering in these data. The LASS and LALRS statistics identify very weak evidence for a cluster associated with lowered risk west of Madison (LASS: $p = 0.072$, LALRS: $p = 0.11$); the scan statistic finds no evidence

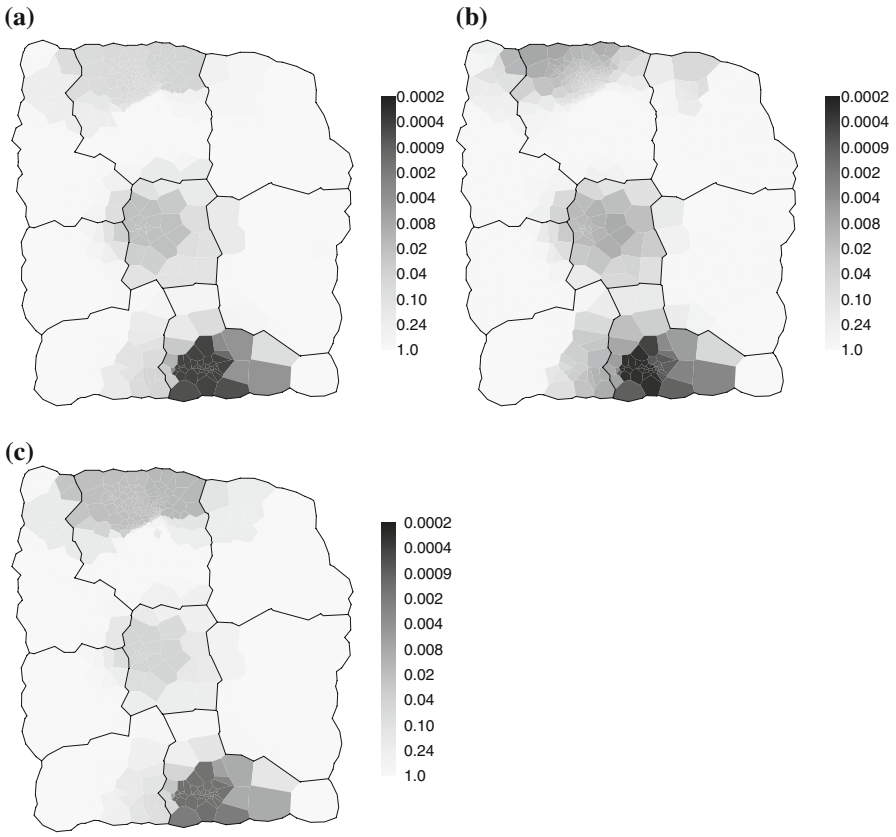

Fig. 5 Cell-specific multiplicity-adjusted p-values for the New York leukemia data using the **a** LASS (global test $p = 0.0006$), **b** LALRS (global test $p = 0.0004$) and **c** spatial scan (global test $p = 0.0017$) statistics

for clustering ($p = 0.55$). The differences in p-values again reflect the interaction of the underlying geography with each statistic.

7 Conclusions

We have proposed two new cluster detection methods, the locally adjusted spatial scan (LASS) statistic and the local average likelihood ratio scan (LALRS) statistic. Both statistics can be used anytime the spatial scan statistic is applicable. The LASS statistic introduces a local multiplicity adjustment into the spatial scan statistic. The LALRS statistic is a generalization of the weighted average likelihood ratio (WALR) and weighted average likelihood ratio scan (WALRS) statistics.

Our assessments of the geography of power for spatial scan, LASS and LALRS statistics provide insight into their relative performance. No statistic is uniformly better than another statistic. The spatial scan statistic performs best in urban areas, while

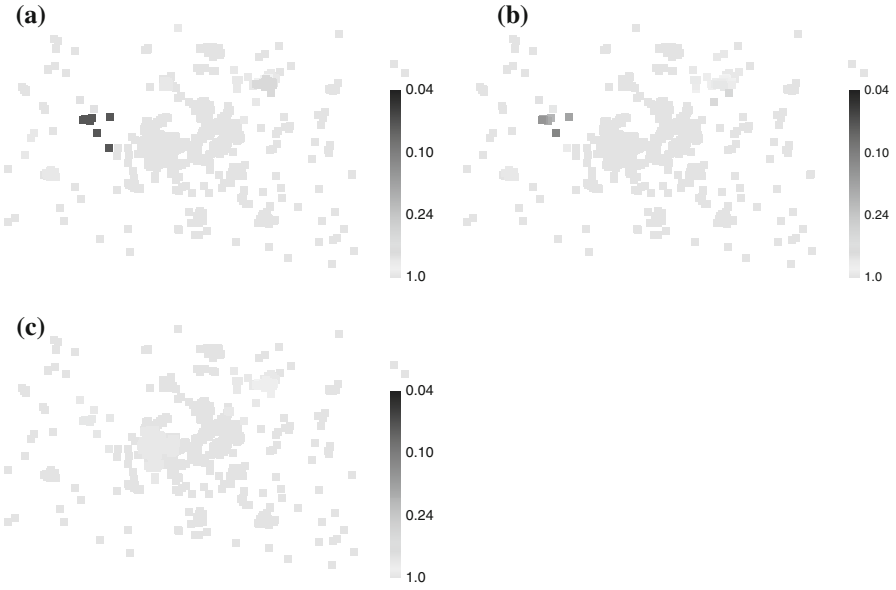

Fig. 6 Location-specific multiplicity-adjusted p-values for the Dane county breast cancer data using the **a** LASS (global test $p = 0.072$), **b** LALRS (global test $p = 0.11$) and **c** spatial scan (global test $p = 0.55$) statistics

the LASS and LALRS statistics perform better in rural areas. The performance for the LASS statistic is somewhat more consistent (although still not ideal) than the performance of the LALRS statistic across urban and rural areas. For these reason, the LASS statistic might be recommended for general use with the scan statistic reserved for situations in which clusters in urban areas are of special concern. Obviously, these studies can provide only a limited assessment of the relative performance of these methods, given the use of only two study regions. In future work, we will expand these assessments to other study regions.

Location-specific adjusted p-values provide a useful inferential summary for these cluster detection methods. Grayscale maps of these p-values effectively communicate (1) results of the global test, (2) evidence for clustering in different portions of the study region and (3) uncertainty about the membership of the true cluster(s). By conveying our uncertainty about the exact composition of the cluster(s), location-specific p-values provide a more informative summary than the current practice of reporting the "most likely cluster" and "non-overlapping secondary clusters" for the spatial scan statistic.

The LASS and LALRS statistics are only suitable for the detection of hot spot clusters, i.e., clusters of constant risk. Adaptations of these statistics to the detection of clinal clusters, i.e., clusters in which risk increases with greater proximity to the cluster center, are not straightforward. Any successful adaptation of these statistics to a more general setting will likely require a better understanding of the dependence of both statistics on the labeling of locations as falling inside or outside the cluster,

i.e., the particular coding of the 0-1 contrast associated with the cluster. Switching the coding of the 0-1 contrast does not change either the model or the likelihood ratio statistic, but it will produce a different LASS and LALRS statistic. Since a useful extension of these statistics should be invariant with respect to the coding of contrasts, we will attempt to identify principles that justify the use of the preferred coding of 0-1 contrasts with these statistics.

Acknowledgements I would like to thank Amy Trentham-Dietz and Polly Newcomb for providing access to the Dane County breast cancer dataset and Jane McElroy and John Hampton for their assistance with the dataset. I also thank Murray Clayton, Tom Cook and several anonymous reviewers for their helpful comments and suggestions on earlier drafts of this manuscript. Partially supported by grants CA47147 and CA82004 from the National Cancer Institute.

References

Besag J, Newell J (1991) The detection of clusters of rare diseases. J Roy Stat Soc Ser A 154:143–155

Dumitrescu RG, Cotarla I (2005) Understanding breast cancer risk—where do we stand in 2005. J Cell Mol Med 9:208–221

Gangnon RE (2006) Impact of prior choice on local Bayes factors for cluster detection. Stat Med 25:883–895

Gangnon RE, Clayton MK (2000) Bayesian detection and modeling of spatial disease clustering. Biometrics 56:922–935

Gangnon RE, Clayton MK (2001) A weighted average likelihood ratio test for spatial clustering of disease. Stat Med 20:2977–2987

Gangnon RE, Clayton MK (2003) A hierarchical model for spatially clustered disease rates. Stat Med 22:3213–3228

Gangnon RE, Clayton MK (2004) Likelihood-based tests for detecting spatial clustering of disease. Environmetrics 15:797–810

Gangnon RE, Clayton MK (2007) Cluster detection using Bayes factors from overparameterized cluster models. Environ Ecol Stat 14:69–82

Hossain MM, Lawson AB (2006) Cluster detection diagnostics for small area health data: with reference to evaluation of local likelihood models. Stat Med 25:771–786

Kulldorff M (1997) A spatial scan statistic. Commun Stat Part A 26:1481–1496

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

Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. Stat Med 25:3929–3943

Lawson AB (2006) Disease cluster detection: a critique and a Bayesian proposal. Stat Med 25:897–916

McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA (2003) Geocoding addresses from a large population-based study: lessons learned. Epidemiology 14:399–407

Newcomb PA, Storer BE, Longnecker MP, Mittendorf R, Greenberg ER, Clapp RW, Burke KP, Willett WC, MacMahon B (1994) Lactation and a reduced risk of premenopausal breast cancer. N Engl J Med 330:81–87

Newcomb PA, Egan KM, Titus-Ernstoff L, Trentham-Dietz A, Greenberg ER, Baron JA, Willett WC, Stampfer MJ (1999) Lactation in relation to postmenopausal breast cancer. Am J Epidemiol 150:174–182

Tango T (1995) A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Stat Med 14:2323–2334

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

Waller LA, Turnbull BW, Clark LC, Nasca P (1994) Spatial pattern analyses to detect rare disease clusters. In: Lange N, Ryan L, Billiard L (eds) Case studies in biometry. Wiley, New York, pp 3–22

Waller LA, Hill EG, Rudd RA (2006) The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. Stat Med 25:853–865

Author Biography

Dr. Ronald E. Gangnon is an Assistant Professor in the Departments of Biostatistics and Medical Informatics and Population Health Sciences at the University of Wisconsin-Madison. He received his bachelor's degree in Mathematics and Economics in 1992 from the University of Minnesota-Duluth and his M.S. and Ph.D. in Statistics from the University of Wisconsin-Madison. His research interests are in spatial epidemiology, with particular emphasis on cluster detection and, more generally, in the application of statistics to medical and epidemiologic data.