ELSEVIER Computational Statistics & Data Analysis 42 (2003) 665–684

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Power comparisons for disease clustering tests

Martin Kulldorff[a,b,*], Toshiro Tango[c], Peter J. Park[d,1]

[a] *Department of Statistics, University of Connecticut, Storrs, CT 06269-4120, USA*
[b] *Department of Community Medicine and Health Care, University of Connecticut, 263 Farmington Avenue, Farmington, CT 06030-6325, USA*
[c] *Division of Theoretical Epidemiology, Department of Epidemiology, The Institute of Public Health, 6-1 Shirokanedai 4 chome, Minato-ku, Tokyo 108, Japan*
[d] *Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston, MA 02115, USA*

Received 1 November 2001; received in revised form 1 May 2002

Abstract

Many different methods have been proposed to test for geographical disease clustering, and more generally, for spatial clustering of any type of observations while adjusting for an inhomogeneous background population generating the observations. Despite the many proposed test statistics, there has been few formal comparisons conducted. We present a collection of 1,220,000 simulated benchmark data sets generated under 51 different cluster models and the null hypothesis, to be used for power evaluations. We then use these data sets to compare the power of the spatial scan statistic, the maximized excess events test and the nonparametric M statistic. All have good power, the first having an advantage for localized hot-spot type clusters and the second for global clustering where randomly located cases generate other cases close by. By making the simulated data sets publicly available, new tests can easily be compared with previously evaluated tests by analyzing the same benchmark data.

Keywords: Spatial statistics; Power; Geography; Spatial epidemiology; Hypothesis testing; Cluster detection

* Corresponding author. Department of Community Medicine and Health Care, University of Connecticut School of Medicine, 263 Farmington Avenue, Farmington, CT 06030-6325, USA. Tel.: +860-679-5473.

E-mail addresses: martink@neuron.uchc.edu (M. Kulldorff), tango@iph.go.jp (T. Tango), peter_park@harvard.edu (P.J. Park).

[1] Current address: Children's Hospital Informatics Program, 320 Longwood Ave, Boston, MA 02115, USA.

0167-9473/03/$ - see front matter
PII: S0167-9473(02)00160-3

1. Introduction

A large number of different tests for spatial randomness that adjust for an uneven background population have been proposed. Such test statistics are used to test, among other things, whether the geographical distribution of disease is random or not, adjusting for the geographical distribution of the population at large. They are also used in areas such as archaeology, botany, criminology, demography, ecology, economics, engineering, forestry, genetics, geography, history, neurology, sociology and zoology. Several review articles have been written (Biggeri and Marchi, 1993; Elliott et al., 1995; Heywood, 1991; Kulldorff, 1998; Lawson et al., 1999; Marshall, 1991; Moore and Carpenter, 1999; Orton, 1982; Sokal and Oden, 1978; Waller and Jacquez, 1995), the most extensive having identified over 100 different test statistics (Kulldorff).

There have been few systematic comparative evaluations of tests for spatial randomness. Different tests have sometimes been applied to the same data sets (Alexander and Boyle, 1996; Draper, 1991; Glaser, 1990; Shaw et al., 1988; Turnbull et al., 1990; Zoellner and Schmidtmann, 1999), but for a formal comparison of test statistics it is important to evaluate their power (Wartenberg, 1990), and only a small fraction of the proposed tests has undergone such evaluations. Three major considerations when designing a power comparison study are (i) the reproducibility of the clustering process, (ii) the clustering models considered as the alternative hypotheses, and (iii) minimization of the bias and variance when estimating the difference in power for different tests.

1.1. Reproducibility

While very important, simulated power comparisons are tedious, time consuming and unglamorous to perform. Each of the methods to be evaluated must be programmed, the simulated data must be generated, and each test statistic must be calculated for each simulated data set. If there are previously published power evaluations, it is sometimes possible to avoid redoing the calculations for already evaluated test statistics, but that requires that the earlier simulation models are described in complete detail, which is seldom the case. The ideal is to go one step further though, and build on previous power evaluations using not only the same alternative models but also the exact same simulated data. That minimizes the random variation when the methods are compared.

In this paper, we present and provide access to a set of benchmark simulated data sets. Using this benchmark, we evaluate the power of three test statistics for disease clustering. Other researchers can then easily compare tests of their interest to previously evaluated test statistics by simply reanalyzing the benchmark data sets. This is the most economical way to conduct power comparisons of many different test statistics. Past evaluations of tests for spatial randomness have for natural reasons been done mostly as pairwise power comparisons or more rarely in groups of three or four (Kulldorff and Nagarwalla, 1995; Oden, 1995; Rogerson, 1999; Swartz, 1998; Tango, 1995; Tango, 1999a; Tango, 2000; Vach, 1994). By establishing the benchmark data sets,

any new test evaluated will automatically be compared with all previously evaluated test statistics.

1.2. Clustering models

With one exception, earlier power comparisons all considered first-order clustering models where cases are located independently of each other, but where the relative risk is different in different geographical areas. Most of these evaluated the power for a clustering model with one (Kulldorff and Nagarwalla, 1995; Rogerson, 1999; Swartz, 1998; Tango, 1995; Tango, 1999a; Tango, 2000), two (Swartz, 1998; Tango, 1995; Tango, 2000), three (Tango, 2000) or four (Swartz, 1998) hot-spot clusters. As an important alternative, Oden (1995) used a clustering model with a different relative risk in each census area. Vach (1994) is the only one who has considered a second-order clustering model. In his model, the location of one case is dependent on the location of other previously generated cases, with the risk varying geographically at the same time. There has not been any power comparison using a pure second-order model, where each particular case is randomly located, so that the relative risk is constant throughout the map, but where the location of cases are dependent on each other. It is important to realize that while first- and second-order models are very different in how the points are generated, the resulting point patterns may be exactly the same, and hence indistinguishable. Bailey and Gatrell (1995, Chapter 3) provide an excellent discussion of this.

In this study we use 61 different clustering models, 15 with a single hot-spot cluster, 20 with multiple hot-spot clusters, and 26 with purely second-order clustering models where the risk is constant throughout so that any one particular case is spatially randomly located, but where the location of different cases are dependent on each other. For each model, the power is calculated conditioned on two different levels of the total number of cases. The number of alternative clustering models considered have in past studies been in the range of 3–8, with the exception of Vach (1994) and Rogerson (1999), who considered 12 and 20 different clustering models respectively. Tango (1995) is the only one who has evaluated the power for the same models but with different number of cases.

Another important aspect of a clustering model is the background population used. We use a real data set consisting of all women in 245 counties in Northeastern United States during 1988–1992. This is a fairly typical epidemiological data set, with data aggregated into a mix of rural and urban census areas.

1.3. Minimization of bias and variability

For some tests it is possible to evaluate the power using an asymptotic approximation of the test statistic distribution (Oden, 1995; Rogerson, 1999; Tango, 1995). Unfortunately, asymptotic approximations do not exist for most test statistics. When they do exist, the asymptotics may be defined in terms of the geographical area, the population size or the number of cases going to infinity, with the other two parameters held at a specific constant or rate, and the approximations must be interpreted considering these

asymptotic concepts. Unless the approximations for all test statistics are very good, it is necessary for comparison purposes to obtain the critical values through a large number of simulated data sets randomized under the null hypothesis. In this paper we present two groups of 100,000 simulated data sets to estimate the critical values, with 600 and 6000 cases, respectively.

In order to minimize the variability of the estimated power difference between tests, it is important to condition the analysis on a particular population distribution, and on the total number of cases. Moreover, different tests should be evaluated using the same random data sets.

Another factor determining the variance of an estimated power difference is the number of random data sets generated under each alternative hypothesis. As part of the benchmark data, we present 10,000 random data sets for each alternative.

1.4. Test statistics compared

Tests for spatial randomness can be classified based on their purpose. Focused tests are designed to test whether a local cluster exist around a predetermined point source, while general tests looks for clusters without any preconceived assumptions about their location (Besag and Newell, 1991). Among general tests, cluster detection tests are used both to detect local clusters, without any preconceived idea of their location, and to determine their statistical significance. Global clustering tests, on the other hand, are used to determine whether there is clustering present throughout the study area, without determining statistical significance of individual clusters (Kulldorff, 1998; Tango, 1999a).

Discussions regarding the differences between the latter two types of general tests have been provided elsewhere (Kulldorff, 1998; Tango, 1999a), but their important difference is not always considered, and there has never been a formal study showing how they differ in terms of their power to detect different types of clustering. In fact, the power of global clustering tests has typically been evaluated using hot-spot cluster models. In this paper we evaluate the power of the spatial scan statistic (Kulldorff, 1997), the maximized excess events test (Tango, 2000) and Bonetti–Pagano's nonparametric M statistic (Bonetti and Pagano, 2001a,b). These were chosen so as to not only compare three different tests, but equally important, to illustrate the differences between the two types of tests. We show that the spatial scan statistic, a cluster detection test, has good power for hot-spot cluster alternatives, while the maximized excess events test, a global clustering test, has good power when clustering occurs throughout the geographical region of study. The M statistic, also a global clustering test, performs well for multiple hot-spot clusters.

Most tests for spatial clustering depend on a parameter that determines the scale of clustering. This includes the λ in Tango's excess events test (Tango, 1995), the k in Cuzick–Edward's k-nearest neighbor test (Cuzick and Edwards, 1990), and the radius of the circle in Turnbull's CEPP (Turnbull et al., 1990). The three tests compared in this paper do not depend on such a pre-specified parameter. This was a main reason for evaluating these particular test statistics. We expect that more such tests will be proposed as extensions of earlier methods, and it will then be of special interest to compare them with the tests evaluated here.

2. Benchmark data sets

For the benchmark data sets we use the female population in the 245 counties and county equivalents in the Northeastern United States, consisting of the states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware and Maryland, as well as the District of Columbia. Each county is geographically represented by a centroid coordinate. As the population for each county we used the number of women living there according to the 1990 United States census. This data has previously been used to evaluate the existence of geographical clusters of breast cancer mortality (Kulldorff et al., 1997). Both the population data and the geographical coordinates are available at 'http://www.commed.uchc.edu/biostat/datasets/'.

2.1. Hot-spot clusters

For the hot-spot alternatives, we constructed three different sets of local clusters in a rural, urban and mixed area, respectively. Within each of these three sets, we constructed five different size clusters using 1, 2, 4, 8 and 16 counties. The center of the rural cluster was Grand Isle County in northern Vermont, on the Canadian border. Among the 245 counties, Grand Isle has the smallest population. The center of the mixed cluster was Pittsburgh (Allegheny County) in western Pennsylvania. Pittsburgh is a large city, surrounded by rural areas. The center of the urban cluster was Manhattan (New York County) in New York City, closely surrounded by other very urban counties. Additional counties were added to the central county by order of geographic distance between county centroids. The clusters with 16 counties are shown on the map in Fig. 1. The New York City cluster is close to the population center of the region, while the Pittsburgh cluster contains the urban area furthest away from the population center.

The counties within each cluster were given a higher risk than the remaining counties. For each of the 15 clusters, the relative risks and the expected number of cases under both the null and the alternative hypotheses are given in Table 1. The relative risks were set so that the null hypothesis would be rejected with probability 0.999 when using a standard binomial test, had we known the 'cluster counties' a priori, not taking the multiple testing into account. Let n be the combined population in the cluster counties, and let N be the total population in all counties. Conditioned on the total number of cases C, the observed number of cases in the 'cluster counties' is under the null hypothesis binomially distributed with mean $m_0 = Cn/N$ and variance $v_0 = Cn/N(N-n)/N$. Using the normal approximation for the binomial distribution, the critical number of cases k needed in order for a one-sided test to reject the null hypothesis at the 0.05 level is then the k such that $(k-m_0)/\sqrt{v_0} = 1.645$. Under the alternative hypothesis with a relative risk of r for the 'cluster counties', the number of cases in those counties is binomially distributed with mean $m_A = Cnr/N - n + nr$ and variance $v_A = Cnr/(N-n+nr)\ (N-n)/(N-n+nr)$. Using the normal approximation again, we then selected the relative risk r such that $(k-m_A)/\sqrt{v_A} = 3.09$. This choice of relative risks provides an upper limit of 0.999 for the power attainable by any test

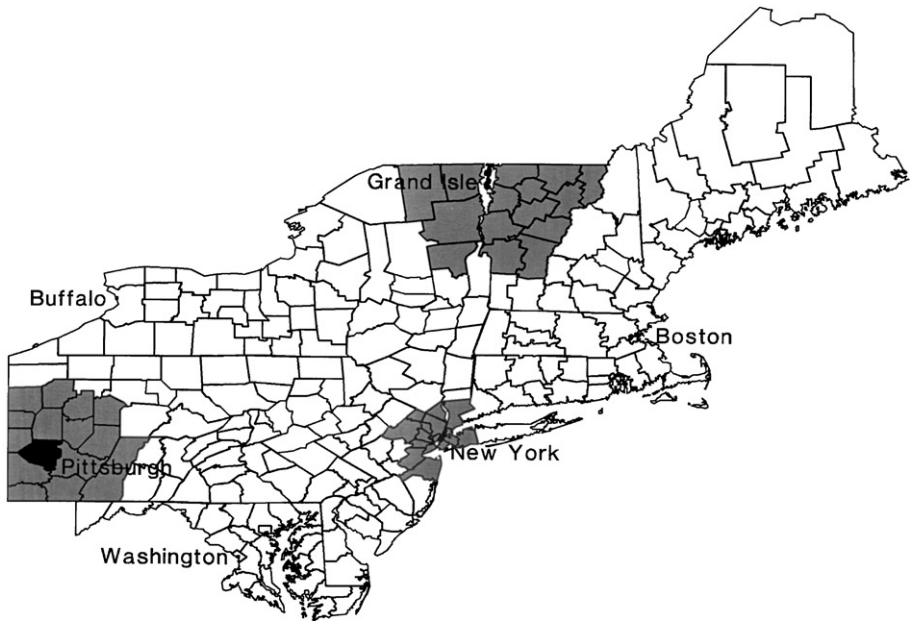

Fig. 1. Map showing the rural cluster centered around Grand Isle in the north, the mixed cluster centered around Pittsburgh in the west and the urban cluster centered around Manhattan, New York, in the center.

Table 1
Hot-spot clusters $E[c|H_0]$ is the expected number of cases under the null hypothesis, and $E[c|H_A]$ is the same under the alternative. RR is the relative risk

		With 600 simulated cases			With 6000 simulated cases		
Counties		$E[c\|H_0]$	$E[c\|H_A]$	RR	$E[c\|H_0]$	$E[c\|H_A]$	RR
Rural	1	0.05	10	192.89	0.5	13	23.73
(On edge)	2	0.46	12	27.03	4.6	23	4.96
	4	2.69	18	7.05	26.9	59	2.21
	8	4.16	22	5.35	41.6	80	1.92
	16	7.32	28	3.90	73.2	121	1.66
Mixed	1	14.43	39	2.85	144.3	208	1.45
(Corner)	2	16.41	42	2.70	164.1	231	1.42
	4	22.52	51	2.40	225.2	302	1.36
	8	27.47	58	2.24	275.7	358	1.32
	16	34.22	67	2.10	342.2	434	1.29
Urban	1	15.97	42	2.73	159.7	226	1.43
(Central)	2	21.78	50	2.43	217.8	293	1.36
	4	59.99	100	1.81	599.9	716	1.22
	8	101.96	150	1.63	1019.6	1162	1.17
	16	154.94	209	1.53	1549.4	1713	1.15

for spatial clustering, and a yard stick by which to compare the performance of a test statistic on different hot-spot clusters.

In order to evaluate how the disease clustering tests perform when there are multiple hot-spot clusters, we constructed fifteen alternative models that included two clusters, urban/rural, urban/mixed or mixed/rural, with the same number of counties in both clusters. For another five alternative models, we included three clusters, one of each type, with an identical number of counties in each. For each cluster we set the relative risks as before, according to Table 1. Note that while the number of counties is the same in each cluster, the relative risks and the population sizes are different.

In total we constructed 35 different hot-spot cluster models, with varying characteristics.

2.2. Global chain clustering

For the global clustering alternative, we want cases to be clustered wherever they occur in the region. Moreover, for all counties we want the expected number of cases to be the same under the null and alternative hypotheses. These requirements pose a special challenge in constructing a clustering model.

For the global clustering model a certain number of cases are first located randomly on the map, according to the null hypothesis. These original cases then generate other new cases close by. If each original case generates one additional case, we call them twins, and if two additional cases are generated, we call them triplets.

Let n_i be the population of county i, and let $N = \sum_i n_i$. Let d_{ij} be the Euclidean distance between counties i and j. If the original case is in county i, a natural way to assign its twin is to put it in county j, where j is chosen so that $\sum_k I(d_{ik} < d_{ij})n_k < rN \leqslant \sum_k I(d_{ik} \leqslant d_{ij})n_k$, for some constant $0 \leqslant r < 0.5$. This means that the twin is selected as the rN-nearest neighbor of the original case. In other words, a randomly selected case has probability r of being closer than the twin to the original case. A problem with this type of approach is that the additional cases will not be spatially randomly distributed, but have a higher chance to occur in the central areas of the map as compared to outlying areas. This is because someone in the center of the map is a closer neighbor to more other individuals as compared to someone that lives close to the border. Hence, the requirement that every county has the same expected number of cases under the null and alternative hypotheses is not met.

To overcome the above problem, we used what we call a global chain clustering model. The counties are tied together sequentially on a chain that passes through each county exactly once, after which it reconnects with the first county on the chain, forming a Hamiltonian cycle. The randomization of twins and triplets is then embedded within this chain, so that an additional case is assigned to county j if $\sum_k I(d'_{ik} < d'_{ij})n_k < rN \leqslant \sum_k I(d'_{ik} \leqslant d'_{ij})n_k$, where d'_{ij} is now the distance in one particular direction along the chain connecting the counties. Hence, the twins are assigned as the rN nearest neighbor along one direction of the chain. For twins, the probability model is the same independent of the direction used. For triplets, the two new cases were assigned in opposite directions. Note that the chain does not imply

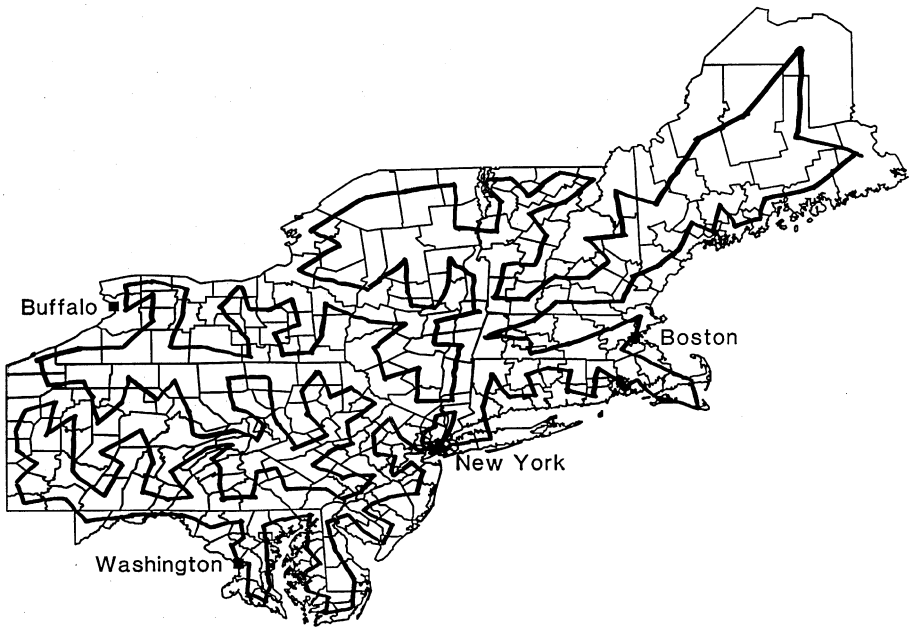

Fig. 2. Map showing the circular chain of counties used for the global chain clustering model.

that the 'disease' spreads itself around the chain, just that twin and triplet cases are located in either of only two directions, as defined by the chain.

The chain used is shown in Fig. 2. The chain was constructed so that two counties next to each other on the chain always border each other geographically. Moreover, it was constructed so that all counties in a state occur consecutively along the chain, except for New York and New Hampshire where that was not possible as these two states stretch from the Canadian border to the Atlantic coast. Within these parameters, the exact construction of the chain is arbitrary, but an attempt was made to have counties that are geographically close to be close along the chain as well, to the largest extent possible.

We constructed different clustering models by using different constants r for the population based distances between the twins. In the most clustered scenario, the distance was zero, so that twins are always assigned to the same county. The chain was not needed for this scenario. In a second set of clustering models, r was deterministically set to be 0.005, 0.01, 0.02, 0.04, 0.08 and 0.16, respectively. In a third set of clustering models, r was exponentially distributed with mean 0.005, 0.01, 0.02, 0.04, 0.08 and 0.16, respectively.

With the chain model, to use $r = 1$ is the same as using $r = 0$. To use $r = 0.5$ would assign the twins at the opposite ends of the chain, typically putting them further away from each other than they would be by chance, leading to the opposite of clustering. If the chain were a perfect circle with even population distribution along the circle, then $r = 0.22$ would correspond to a situation where the expected distance between twins is

the same as the expected distance between any two cases under the null model (see the appendix). If the chain was circular, with an uneven population distribution, that equality would hold for a smaller r. In our case, we do not have a circular chain so it is not clear what value of r represents a situation of no clustering, but the above reasoning for a perfectly circular chain means that we should not necessarily expect to see clustering for distances greater than $r = 0.22$, and the largest distance used at $r = 0.16$ represents at most a very weak amount of clustering.

The equivalent 13 models were also used for triplets, resulting in a total of 26 global chain clustering models.

2.3. Simulated data

In order to perform power comparisons, we constructed a number of random data sets. These are in two groups, with 600 and 6000 simulated cases respectively. These numbers were chosen because we wanted the total number of cases to be divisible by both 2 and 3, to fit with both the twin and triplet models. The same data sets were used to evaluate power both at the $\alpha = 0.05$ and 0.01 significance levels.

The same null hypothesis is used throughout, where the relative risk is set to one for each county, and case locations are independent of each other. This means that a particular case is assigned to county i with probability n_i/N. We generated 100,000 random data sets with 600 cases and the same number of data sets with 6,000 cases. These are used to estimate the cut-off point for significance. For each alternative hypothesis, we generated 10,000 random data sets. Using the null cut-off points, these were used to estimate the power. A Lehmer random number generator was employed, with modulus $2,147,483,647$ and multiplier 48,271 (Park and Miller, 1988).

The same data sets were used for the three tests, so as to eliminate any power differential due to some data sets being by chance more clustered than others. All 1,220,000 data sets can be downloaded from the world wide web at 'http://www.commed.uchc.edu/biostat/datasets/'.

3. A cluster detection test and two global clustering tests

The clustering models described above can be used for power analysis of any number of disease clustering tests. In this paper we estimate the power of one cluster detection test, the spatial scan statistic (Kulldorff, 1997), and two global clustering tests, the maximized excess events test (Tango, 2000) and Bonetti–Pagano's M statistic (Bonetti and Pagano, 2001a,b). As others use the same data sets to evaluate other disease clustering tests, they only need to do the power calculations for the new tests, enabling an automatic comparison with these three test statistics.

3.1. The spatial scan statistic

The spatial scan statistic (Kulldorff, 1997) imposes a circular window on a map and lets the circle centroid move across the study region. For any given position of the

centroid, the radius of the window is changed continuously to take any value between zero and some upper limit. In total, the method uses a set $\mathscr{Z}$ containing an infinite number of distinct circles, each with a different location and size, and each being a potential cluster. We set the upper limit so that the circle may contain at most 50 percent of the total population.

Under the alternative hypothesis, there is at least one circle for which the underlying risk is higher inside the circle as compared to outside. For each circle, it is possible to calculate the likelihood to observe the observed number of cases within and outside the circle respectively. The circle with the maximum likelihood is defined as the most likely cluster. This is the cluster that is least likely to have occurred by chance.

The likelihood can be calculated assuming either a Poisson or Bernoulli model, depending on how the cases are generated. We use the former.

Let $c(Z)$ be the observed number of cases in circle Z. Let $n(Z)$ be the expected number of cases in circle Z under the null hypothesis, so that $n(A) = c(A) = C$, where A is the total region under study. Let $L(Z)$ be the likelihood under the alternative hypothesis that there is a cluster in circle Z, and let L_0 be the likelihood under the null hypothesis. It can then be shown that

$$\frac{L(Z)}{L_0} = \left(\frac{c(Z)}{n(Z)}\right)^{c(Z)} \left(\frac{C - c(Z)}{C - n(Z)}\right)^{C - c(Z)} \tag{1}$$

if $c(Z) > n(Z)$ and one otherwise. Details, including derivations as a likelihood ratio test, have been given elsewhere (Kulldorff, 1997). As this likelihood ratio is maximized over all the circles, it identifies the one that constitutes the most likely cluster. The test statistic is

$$\max_Z \frac{L(Z)}{L_0}.$$

When derived as a likelihood ratio test, it is based on a set of alternative hypotheses, each with a single circular cluster of different size, location and relative risk. This does not mean that the test statistic can only detect circular clusters, but should expect higher power for more compact clusters if everything else is equal. Its p-value is obtained through Monte Carlo hypothesis testing (Dwass, 1957). Calculations were done using SaTScan (Kulldorff et al., 1998). The method has been applied in a wide variety of epidemiological studies (e.g. Cousens et al., 2001; Fèvre et al., 2001; Imai, 1998; Kojima, 1999; Kulldorff et al., 1997; Nakatani, 1999; Sankoh et al., 2001; Viel et al., 2000; Walsh and Fenster, 1997).

3.2. The maximized excess events test

Let c_i be the observed number of cases in county i, and let $C = \sum_i c_i$ be the total number of cases. Let n_i be the expected number of cases in county i under the null hypothesis, so that $\sum_i n_i = C$. For a given constant λ, the excess events test statistic (Tango, 1995) is defined as

$$EET(\lambda) = \sum_i \sum_j a_{ij}(d_{ij}, \lambda)\ (c_i - n_i)(c_j - n_j),$$

where

$$a_{ij}(d_{ij},\lambda) = \mathrm{e}^{-4d_{ij}^2/\lambda^2}$$

and d_{ij} is the Euclidean distance between counties i and j. Other choices of $a_{ij}(d_{ij},\lambda)$ are also possible (Oden, 1995; Tango, 1995; Rogerson, 1999). The choice of λ relates to the geographical scale of clustering, and is to some extent arbitrary. A large λ will give a test sensitive to geographically large clusters, while a small λ will make the test more sensitive to small ones.

To be able to detect clustering irrespectively of its geographical scale, Tango (2000) proposed the maximized excess events test (*MEET*):

$$MEET = \min_{0\leqslant\lambda\leqslant U} P[EET(\lambda) > eet(\lambda)|H_0,\lambda],$$

where $eet(\lambda)$ is the observed value of the excess events test statistic conditioning on λ, and U is an upper limit on λ. Practical implementation of the test uses 'line search' by discretization of λ, and the *MEET* statistic is evaluated using Monte Carlo hypothesis testing (Dwass, 1957).

Calculations were done using a specially written S-Plus code (Tango, 1998). The method has been applied to various epidemiological data sets (Imai, 1998; Kojima, 1999; Nakatani, 1999; Tango, 1999b).

3.3. Bonetti–Pagano's M statistic

The M statistic proposed by Bonetti and Pagano (2001a,b) uses the interpoint distance distribution function to describe the spatial pattern of a set of points. Let $F(d) = P(D \leqslant d)$ be the cumulative distribution function of the random variable $D = dist(X_1,X_2)$ that represents the distance between the coordinates X_1 and X_2 of two individuals chosen at random from a spatial distribution $\mu_X(\cdot)$. F can be estimated consistently from a random sample of n individuals $X_1,X_2,\ldots,X_n$ by the quantity

$$\hat{F}_n(d) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I(dist(X_i,X_j) \leqslant d).$$

Moreover, $\sqrt{n}(\hat{F}_n(d) - F(d))$ converges to a Gaussian process (Bonetti and Pagano, 2001b). A nonparametric test for deviations from the null distribution $\mu_X(\cdot)$ can thus be constructed by defining a statistic that measures the distance between the function $\hat{F}_n(\cdot)$ based on the observed case locations and the function $F(\cdot)$ based on the underlying population at risk. One such test statistic can be constructed by splitting the distance axis $[0, d_{\max}]$ (with $d_{\max}$ being the largest observed interpoint distance) into the K intervals $[(t-1)d_{\max}/K, td_{\max}/K]$, for $t = 1,\ldots,K$. From the vectors $\mathbf{F} = [F(d_{\max}/K),\ldots,F(d_{\max})]$ and $\hat{\mathbf{F}}_\mathbf{n} = [\hat{F}_n(d_{\max}/K),\ldots,\hat{F}_n(d_{\max})]$ one can then construct the quadratic form $M = (\hat{\mathbf{F}}_\mathbf{n} - \mathbf{F})'\mathbf{S}^-(\hat{\mathbf{F}}_\mathbf{n} - \mathbf{F})$, where $\mathbf{S}^-$ is the sample (generalized) inverse of the variance–covariance matrix of $\hat{\mathbf{F}}_\mathbf{n}$. Note that if one creates the vector of first differences from $\hat{\mathbf{F}}_\mathbf{n}$ and $\mathbf{F}$, then M can be written equivalently in terms of such first differences.

In the discrete setting such as the one considered here, the population distribution μ and the sample are summarized by probabilities p_j and by counts n_j at some T fixed locations (indexed by $j = 1, \ldots, T$). $\hat{F}_n$ can then be written as

$$\hat{F}_n(d) = \frac{1}{n^2} \sum_{i=1}^{T} \sum_{j=1}^{T} n_i n_j I(d_{ij} \leqslant d),$$

where d_{ij} is the distance between the two locations indexed by i and j, respectively. While the asymptotic distribution of the M statistic can be derived, convergence to this distribution appears to be slow. So we obtain an estimate of the distribution of M by sampling from the (known) probabilities p_j at the T locations.

This statistic incorporates the geographical information by taking into account the behavior of the interpoint distance distribution over its whole range and capturing the dependence among the interpoint distances through the covariance matrix. Because of its definition, the interpoint distance distribution (and thus the M statistic, or other measures of the distance between $\hat{F}_n$ and F) is a summary of the spatial distribution that can be defined over continuous settings, and in very high dimensional spaces.

Note that in the settings described in this paper only an individual's county of residence is reported. Thus only crude approximations of all the interpoint distances are available, and thus the interpoint distance distribution is less informative than it could be.

4. Power comparison results

4.1. Hot-spot clusters

The results of the power analyses for the hot-spot clusters are shown in Table 2. For the rural clusters, the spatial scan statistic has very high power while the power of the other two tests is low. For the mixed clusters, all tests have high power with a slight advantage for the spatial scan statistic. For the urban clusters, it is instead the maximized excess events test that has a slight advantage.

All tests have higher power when there are two or three different hot-spot clusters in the same model. This is expected since more clustering is introduced when additional clusters are added. With some exceptions, the spatial scan statistic has slightly higher power, but the differences are in most cases small.

Comparing the power between different cluster models, we find that the spatial scan statistic has highest power for the rural cluster models, followed by the mixed, and then the urban. The maximized excess events test on the other hand, has highest power for the urban cluster models, followed by the mixed and the rural. As a contrast, the M test reaches its highest power for the mixed cluster models. Does this mean that the power is a function of a cluster's population size? Not necessarily. Within the mixed cluster models, the opposite relationship occurs. The power of the spatial scan statistic increases with increasing population size, while for the maximized excess events test, the power sometimes decreases with increasing cluster population size. The power is not a simple function of a cluster's population size, but is also a function of the

Table 2
Estimated power of the spatial scan statistic, the M test and the maximized excess events test ($MEET$) for 35 different alternative models with different hot-spot clusters, for 600 and 6000 simulated cases, respectively, and for significance levels 0.05 and 0.01

		With 600 simulated cases						With 6000 simulated cases					
		$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.01$		
Counties		Scan	M	MEET	Scan	M	MEET	Scan	M	MEET	Scan	M	MEET
Rural	1	0.998	0.355	0.196	0.992	0.127	0.057	0.991	0.102	0.058	0.974	0.024	0.009
(Edge)	2	0.991	0.406	0.221	0.986	0.154	0.072	0.955	0.155	0.072	0.901	0.042	0.016
	4	0.973	0.292	0.229	0.946	0.082	0.064	0.920	0.189	0.088	0.844	0.056	0.019
	8	0.971	0.241	0.213	0.937	0.058	0.055	0.929	0.182	0.097	0.846	0.055	0.019
	16	0.969	0.197	0.229	0.936	0.041	0.062	0.936	0.191	0.112	0.849	0.059	0.027
Mixed	1	0.936	0.909	0.925	0.871	0.757	0.833	0.885	0.791	0.831	0.783	0.585	0.643
(Corner)	2	0.939	0.883	0.896	0.871	0.703	0.771	0.890	0.751	0.773	0.784	0.522	0.556
	4	0.937	0.815	0.838	0.873	0.590	0.654	0.891	0.645	0.694	0.784	0.398	0.416
	8	0.941	0.794	0.817	0.876	0.567	0.599	0.905	0.649	0.687	0.810	0.408	0.390
	16	0.949	0.745	0.832	0.886	0.484	0.602	0.923	0.607	0.705	0.830	0.364	0.407
Urban	1	0.922	0.342	0.941	0.818	0.115	0.870	0.841	0.198	0.859	0.733	0.061	0.697
(Central)	2	0.903	0.397	0.920	0.823	0.154	0.830	0.848	0.253	0.840	0.728	0.089	0.669
	4	0.892	0.711	0.961	0.794	0.428	0.902	0.862	0.568	0.945	0.730	0.320	0.861
	8	0.913	0.844	0.983	0.824	0.619	0.951	0.896	0.740	0.978	0.781	0.508	0.939
	16	0.926	0.777	0.986	0.836	0.504	0.950	0.918	0.721	0.982	0.816	0.477	0.945
Rural	1	1.000	0.980	0.964	0.999	0.916	0.910	0.998	0.838	0.834	0.994	0.652	0.643
and	2	0.999	0.970	0.952	0.997	0.894	0.871	0.993	0.890	0.812	0.975	0.666	0.594
mixed	4	0.997	0.931	0.930	0.987	0.804	0.793	0.990	0.928	0.802	0.961	0.588	0.535
	8	0.996	0.915	0.931	0.986	0.741	0.772	0.990	0.962	0.802	0.958	0.564	0.523
	16	0.996	0.827	0.941	0.982	0.590	0.804	0.991	0.921	0.854	0.961	0.483	0.607
Rural	1	1.000	0.709	0.970	0.998	0.400	0.923	0.998	0.301	0.877	0.992	0.112	0.727
and	2	0.999	0.644	0.962	0.996	0.334	0.895	0.991	0.310	0.864	0.970	0.116	0.706
urban	4	0.992	0.811	0.971	0.974	0.538	0.912	0.981	0.664	0.945	0.938	0.410	0.857
	8	0.991	0.884	0.977	0.968	0.667	0.936	0.985	0.817	0.973	0.941	0.596	0.920
	16	0.987	0.776	0.975	0.947	0.481	0.915	0.978	0.768	0.975	0.917	0.518	0.917
Mixed	1	0.987	0.964	0.998	0.950	0.868	0.995	0.968	0.864	0.994	0.907	0.691	0.974
and	2	0.984	0.950	0.995	0.950	0.829	0.984	0.966	0.843	0.984	0.897	0.647	0.947
urban	4	0.966	0.954	0.991	0.901	0.830	0.969	0.958	0.903	0.987	0.873	0.746	0.948
	8	0.954	0.970	0.990	0.871	0.873	0.960	0.944	0.936	0.989	0.841	0.810	0.946
	16	0.935	0.929	0.984	0.811	0.742	0.935	0.934	0.911	0.987	0.816	0.757	0.936
Rural,	1	1.000	0.991	0.999	0.999	0.958	0.997	0.999	0.918	0.994	0.996	0.783	0.975
mixed	2	1.000	0.981	0.998	0.999	0.920	0.992	0.998	0.890	0.989	0.990	0.721	0.960
and	4	0.996	0.979	0.994	0.981	0.895	0.973	0.994	0.928	0.988	0.965	0.792	0.949
urban	8	0.992	0.980	0.989	0.964	0.901	0.952	0.991	0.962	0.988	0.950	0.870	0.941
	16	0.977	0.929	0.983	0.916	0.744	0.918	0.980	0.921	0.982	0.910	0.770	0.924

Table 3
Estimated power of the spatial scan statistic, the M test and the maximized excess events test (*MEET*) for 26 different global chain clustering models, for 600 and 6000 simulated cases, respectively, and for significance levels 0.05 and 0.01

		With 600 simulated cases						With 6000 simulated cases					
		$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.01$		
Distance(r)		Scan	M	MEET	Scan	M	MEET	Scan	M	MEET	Scan	M	MEET
						Twins							
No distance	0	0.791	0.860	0.990	0.513	0.616	0.945	0.826	0.911	0.988	0.549	0.773	0.942
Fixed	0.005	0.392	0.346	0.624	0.197	0.130	0.376	0.389	0.397	0.621	0.190	0.178	0.362
distance	0.01	0.285	0.163	0.406	0.131	0.044	0.201	0.277	0.181	0.398	0.124	0.061	0.186
	0.02	0.194	0.087	0.264	0.084	0.019	0.110	0.188	0.100	0.259	0.077	0.025	0.106
	0.04	0.124	0.060	0.174	0.049	0.014	0.068	0.119	0.067	0.169	0.042	0.015	0.065
	0.08	0.080	0.051	0.109	0.024	0.009	0.038	0.082	0.056	0.108	0.024	0.010	0.037
	0.16	0.055	0.050	0.059	0.014	0.009	0.014	0.051	0.053	0.061	0.012	0.012	0.013
Exponential	0.005	0.452	0.449	0.738	0.229	0.189	0.486	0.457	0.527	0.735	0.229	0.285	0.481
distance	0.01	0.351	0.304	0.556	0.165	0.106	0.299	0.348	0.358	0.548	0.163	0.154	0.297
	0.02	0.262	0.184	0.378	0.110	0.051	0.171	0.258	0.224	0.379	0.111	0.075	0.169
	0.04	0.185	0.114	0.250	0.073	0.027	0.096	0.180	0.138	0.252	0.071	0.036	0.099
	0.08	0.124	0.083	0.166	0.042	0.018	0.056	0.120	0.094	0.163	0.037	0.021	0.053
	0.16	0.080	0.059	0.107	0.023	0.010	0.029	0.077	0.070	0.100	0.020	0.015	0.029
						Triplets							
No distance	0	0.995	0.996	1.000	0.949	0.969	1.000	0.966	0.998	1.000	0.962	0.991	1.000
Fixed	0.005	0.674	0.569	0.884	0.460	0.291	0.728	0.680	0.631	0.883	0.453	0.402	0.717
distance	0.01	0.491	0.253	0.646	0.309	0.087	0.415	0.485	0.293	0.643	0.297	0.123	0.405
	0.02	0.318	0.117	0.430	0.178	0.032	0.237	0.313	0.134	0.423	0.171	0.044	0.231
	0.04	0.189	0.070	0.265	0.094	0.018	0.135	0.177	0.080	0.255	0.088	0.021	0.128
	0.08	0.102	0.053	0.141	0.038	0.010	0.057	0.098	0.059	0.142	0.035	0.013	0.061
	0.16	0.046	0.049	0.050	0.010	0.011	0.015	0.041	0.053	0.049	0.008	0.012	0.011
Exponential	0.005	0.762	0.734	0.960	0.538	0.457	0.862	0.767	0.804	0.958	0.549	0.604	0.860
distance	0.01	0.610	0.497	0.826	0.388	0.232	0.615	0.608	0.580	0.821	0.379	0.336	0.608
	0.02	0.436	0.294	0.599	0.253	0.099	0.363	0.435	0.350	0.590	0.244	0.153	0.354
	0.04	0.289	0.162	0.390	0.144	0.043	0.202	0.282	0.188	0.384	0.134	0.063	0.197
	0.08	0.171	0.096	0.226	0.068	0.021	0.096	0.166	0.104	0.223	0.064	0.026	0.090
	0.16	0.091	0.062	0.115	0.027	0.013	0.036	0.091	0.066	0.115	0.025	0.013	0.035

geographical cluster size and location, as well as of the level of spatial aggregation (i.e. the average county population size) in and around the cluster.

4.2. Global chain clustering

The power estimates for global chain clustering are shown in Table 3. When the distance is zero, all tests have good power but with a clear advantage for the maximized

excess events test followed by the M statistic. As expected, the power goes down with increasing distance between twins and between triplets, and approaches the nominal significance level at the greatest distance of $r = 0.16$. Note that the power is consistently higher for the clustering models in which the distances between twins/triplets are random according to the exponential distribution, as compared to the fixed distance model, even though the expected distances are the same.

5. Discussion

The results of this study clearly show how different disease clustering tests are good for different types of alternative hypotheses. If one is interested in detecting and evaluating localized clusters, it is better to use the spatial scan statistic, while the maximized excess events test is better at detecting global type clustering that is present throughout the study region. This is to some extent intuitive.

The maximized excess events test is based on the evidence of clustering found throughout the map, as the test statistic is a summation over all the counties. When a cluster is large in population size, it also performs well for hot-spot clusters, since there is then a large proportion of the population that is affected by the cluster. The test statistic uses geographical distance to define closeness of cases. This may explain why it has higher power for hot-spot clusters in urban as opposed to rural areas, as the latter clusters are more dispersed. It also means that it may perform better for a global chain clustering model where distance between twins are defined in terms of geographical distance rather than the nearest neighbors. Although not shown here, a feature of the maximized excess events test that is of great additional value is that it is possible to determine how much each county contributes towards the total amount of clustering observed, as represented by the magnitude of the test statistic corresponding to that particular county (Tango, 2000).

The spatial scan statistic ignores all the information about the location of cases except whether the case is inside or outside the currently evaluated circular zone. The disadvantage of this is that the power is lower when clustering occurs throughout the study region. An advantage is that the rejection can be wholly attributed to a particular cluster, since any rearrangement of the cases outside the cluster cannot reduce the value of the test statistic, no matter how the rearrangement is done. This is discussed elsewhere in formal mathematical language (Kulldorff, 1997). By use of circles, the power depends on the compactness of the cluster shape. The true cluster need not be circular to obtain good power, but the test should not be expected to have good power for a long and narrow cluster, such as along the Hudson river.

Hence, these two tests put weight on different aspects of clustering, and we can classify the spatial scan statistic primarily as a cluster detection test, and the maximized excess events test as primarily a global clustering test. The differences between these two types of tests have been discussed earlier (Kulldorff, 1998; Tango, 1999a), but this is the first time that the difference has been made clear through power comparisons using different types of clustering models.

In the majority of cases, the M statistic had lower power than the other two. This is a reflection of the alternative models considered. From a practical perspective, what is important is to know how different tests perform for different types of alternative models. We believe that there are alternative models for which the M statistic would have the highest power among the three. For example, as a general nonparametric statistic, the M statistic is similar in spirit to Kolmogorov's statistic to test for a sample's being originated by a particular distribution. The M statistic is thus a global clustering test for deviations from the assumed null spatial distribution, and it can be expected to have reasonable performance over a large range of deviations from the null, including the presence of long and narrow clusters. As one such example, a clustering process that would produce several small clusters over the map would be harder to detect for the spatial scan statistic since it would be far from the assumed probabilistic model, while such a process might influence the interpoint distance distribution very strongly and thus be easier to detect by the M statistic. This is suggested by the fact that M statistic performs comparatively better for multiple as compared to single hot-spots, and it may be the top performer for alternatives with more than three hot-spots. It is important to keep in mind that any simulated power comparison is dependent on the particular data set and alternative models used. Most tests will have relative strengths and weaknesses for different clustering models and no single test can have optimal (or suboptimal) power for all alternative hypotheses.

For this study we used a data set typical of epidemiological applications, where both the population and cases are aggregated into census areas of different population size. While all three test statistics can be used for either aggregated or non-aggregated point data, the $MEET$ was designed for the former, the M statistic for the latter and the spatial scan statistic for both types of data in mind. The M statistic may hence have been especially disadvantaged in this setting, and may be suspected to perform better for nonaggregated data.

A limiting factor in this power evaluation is that only one set of spatially distributed populations numbers was used, and the strength of various test statistics depends not only on the alternative clustering models, but also on the spatial distribution of the aggregated areas as well as the relative population sizes in these areas. For example, the M statistic is designed to detect any deviation from the underlying population distribution of the interpoint distances. This seems to penalize it in the settings examined here, and may be a consequence of the fact that the population distribution in the map is driven by the population concentration in the east-coast corridor, so that the interpoint distance distribution may not be a very informative summary of this population. The presence of several peaks in that underlying population distribution have been found to work in favor of this statistic (Bonetti and Pagano, 2001b).

In terms of different number of cases, the comparative results were very similar for 600 and 6000, respectively, so the sensitivity to this model parameter is of lesser concern.

After rejecting the null hypothesis, concluding that there is some form of clustering, it is of course of interest to know the exact nature of the clustering process. For example, is it global type clustering or are there hot-spot clusters? If the former, do the cases consist of twins, or triplets, or do they consist of small groups with a variable

number of cases, or are all cases generated through one single process where each new case generates another one? If the latter, how many hot-spots are there and where are they located? It is important to note that the power estimates provided reflect the power to reject the null hypothesis for whatever reason and that the probability of both rejecting the null hypothesis and correctly determining the type of clustering process is a different matter.

Other scientists are encouraged to use the benchmark data sets presented in this paper to evaluate disease clustering tests that they consider using, or to create new tests that will perform better than those evaluated here. Existing tests of potential interest include the k-nearest neighbors test (Cliff and Ord, 1973; Cuzick and Edwards, 1990), Swartz' entropy test (Swartz, 1998), Besag–Newell's R (Besag and Newell, 1991), the isotonic spatial scan statistic (Kulldorff, 1999), Grimson's method (Grimson and Rose, 1991), Martuzzi–Hills' gamma method (Martuzzi and Hills, 1995), Oden's I_{pop} (Oden, 1995), Rogerson's R (Rogerson, 1999), Ord and Getis' max G_i (Ord and Getis, 1995), Diggle–Chetwynd's D (Diggle and Chetwynd, 1991) and Bithell's M (Bithell, 1999). It would also be worth investigating the maximized excess events test with other weight functions $a_{ij}(d_{ij}, \lambda)$. Comparisons are of great interest regardless of whether other tests turn out to have greater or lower power than those presented here, as it will spread light on the question of what types of tests are good for what types of clustering models.

This paper can be viewed as presenting only a first batch of simulated benchmark data sets for disease clustering test evaluations. Others investigators are encouraged to contribute simulated data generated from other alternative hypotheses of interest. Ideally, this will produce a collection of simulated benchmark data sets for the communal use of all researchers in this area. Other clustering models to consider may be (i) interior hot-spot clusters, (ii) hot-spot clusters with different levels of risk in the center and peripheral areas, (iii) a long and narrow hot-spot cluster, (iv) a very large number of geographically small hot-spot clusters, say about one or two dozen, (v) a global clustering model where each original case has a random number of 'siblings' rather than the fixed number that we used, and (vi) a global double-chain clustering model, with two separate disconnected chains covering two different parts of the map, such as the more rural and urban areas respectively, and with the strength of clustering being different within the two chains. One could also use a Cox process (Cressie, 1993; Lundberg, 1940), where cluster locations and relative risks are random rather than deterministic. The advantage of this is that the comparison of the test statistics would reflect the average performance for a large group of different hot-spot clusters. The disadvantage is that one will not learn for what specific types of hot-spot clusters a particular test statistic has high or low power. It would also be worthwhile to create benchmark data sets for nonaggregated data sets, where each case has unique coordinates.

There are many more tests for spatial clustering, and many more clustering models, for which it is worth while to estimate power. We hope that other researchers will build upon this work, and evaluate other tests using the clustering models used here and, equally important, that they will generate and share simulated data from other clustering models. Most importantly, with the existence of a set of benchmark data

sets, each new power comparison does not need to start from scratch, but can build upon previously calculated power estimates for already evaluated test statistics.

6. Appendix

Suppose we have a circle with radius one centered at $(0,0)$. The distance from $(1,0)$ to the point on the circle corresponding to x degrees is

$$\sqrt{(1-\cos x)^2+\sin^2 x}=\sqrt{2-2\cos x}.$$

The distance to a point 22 percent along the circle is $\sqrt{2-2\cos(2\pi 0.22)}=1.27$. The expected distance from $(1,0)$ to a random point on the circle is

$$\int_0^{2\pi}\sqrt{2-2\cos x}\,\mathrm{d}x=1.27.$$

Acknowledgements

The authors thank Marco Bonetti for advice concerning the implementation of the M test, and two anonymous reviewers for valuable comments that improved the quality of the paper.

References

Alexander, F.E., Boyle, P. (Eds.), 1996. Methods for investigating localized clustering of disease. IARC scientific publication no. 135, International Agency for Research on Cancer, Lyon.

Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Longman Scientific & Technical, Harlow Essex, England.

Besag, J., Newell, J., 1991. The detection of clusters in rare diseases. J. Roy. Statist. Soc. A 154, 143–155.

Biggeri, A., Marchi, M., 1993. Metodi di analisi spazio-temporale in campo epidemiologico: una rassegna. In: Zani (Ed.), Metodi statistici per le analisi territoriali, Franco Angeli, Milan.

Bithell, J.F., 1999. Disease mapping using the relative risk function estimated from areal data. In: Lawson et al. (Eds.), Disease Mapping and Risk Assessment for Public Health, Wiley, London.

Bonetti, M., Pagano, M., 2001a. On detecting clustering. Proceedings of the Biometrics Section, American Statistical Association, pp. 24–33.

Bonetti, M., Pagano, M., 2001b. The interpoint distance distribution as a descriptor of point patterns: an application to cluster detection. J. Amer. Statist. Assoc., under review.

Cliff, A.D., Ord, J.K., 1973. Spatial Autocorrelation. Pion, London.

Cousens, S., Smith, P.G., Ward, H., et al., 2001. Geographical distribution of variant Creutzfeldt–Jakob disease in Great Britain, 1994–2000. The Lancet 357, 1002–1007.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.

Cuzick, J., Edwards, R., 1990. Spatial clustering for inhomogeneous populations. J. Roy. Statist. Soc. B 52, 73–104.

Diggle, P.J., Chetwynd, A.D., 1991. Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics 47, 1155–1163.

Draper, G. (Ed.), 1991. The geographical epidemiology of childhood leukemia and non-Hodgkins lymphomas in Great Britain, 1966–83. Studies on Medical and Population Subjects No. 53. HMSO, London.

Dwass, M., 1957. Modified randomization tests for nonparametric hypotheses. Ann. Math. Statist. 28, 181–187.

Elliott, P., Martuzzi, M., Shaddick, G., 1995. Spatial statistical methods in environmental epidemiology: a critique. Statist. Meth. Med. Res. 4, 137–159.

Fèvre, E.M., Coleman, P.G., Odit, M., et al., 2001. The origins of a new Trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern Uganda. The Lancet 358, 625–628.

Glaser, S.L., 1990. Spatial clustering of Hodgkin's disease in San Francisco Bay area. Amer. J. Epidemiol. 132, S167–S177.

Grimson, R.C., Rose, R.D., 1991. A versatile test for clustering and a proximity analysis of neurons. Meth. Inform. Med. 30, 299–303.

Heywood, J.S., 1991. Spatial analysis of genetic variation in plant populations. Annu. Rev. Ecol. Systematics 22, 335–355.

Imai, J., 1998. Spatial disease clustering in Kochi prefecture in Japan—evaluation of disease indices and disease mapping. NIPH Epidemiology and Biostatistics Research 1998. National Institute of Public Health, Tokyo, pp. 57–96.

Kojima, M., 1999. Spatial clusters of diseases and nutritional factors in Shiga prefecture in Japan, 1987–1996'. NIPH Epidemiology and Biostatistics Research 1999, National Institute of Public Health, Tokyo, 73–120.

Kulldorff, M., 1997. A spatial scan statistic. Commun. Statist. Theory and Methods 26, 1481–1496.

Kulldorff, M., 1998. Statistical methods for spatial epidemiology: tests for randomness. In: Gatrell, Lö ytö nen (Eds.), GIS and Health. Taylor & Francis, London, pp. 49–62.

Kulldorff, M., 1999. An isotonic spatial scan statistic for geographical disease surveillance. J. Natl. Inst. Public Health 48, 94–101.

Kulldorff, M., Tests for spatial randomness adjusted for an underlying inhomogeneity: a general framework, under review.

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

Kulldorff, M., Feuer, E., Miller, B., Freedman, L., 1997. Breast cancer in northeast United States: a geographic analysis. Amer. J. Epidemiol. 146, 161–170.

Kulldorff, M., Rand, K., Gherman, G., Williams, G., DeFrancesco, D., 1998. SaTScan v 2.1: Software for the spatial and space-time scan statistics. National Cancer Institute, Bethesda (http://www.cancer.gov/prevention/BB/SaTScan.html)

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.F., Bertollini, R. (Eds.), 1999. Disease Mapping and Risk Assessment for Public Health. Wiley, London.

Lundberg, O., 1940. On Random Processes and their Application to Sickness and Accident Statistics. Almqvist & Wiksells, Uppsala, Sweden.

Marshall, R.J., 1991. A review of methods for the statistical analysis of spatial patterns of disease. J. Roy. Statist. Soc. A 154, 421–441.

Martuzzi, M., Hills, M., 1995. Estimating the degree of heterogeneity between event rates using likelihood. Amer. J. Epidemiol. 141, 369–374.

Moore, D.A., Carpenter, T.E., 1999. Spatial analytical methods and geographic information systems: use in health research and epidemiology. Epidemiol. Rev. 21, 143–161.

Nakatani, M., 1999. Spatial clusters of diseases and environmental factors in Aomori prefecture in Japan, 1988–1997. NIPH Epidemiology and Biostatistics Research 1999. National Institute of Public Health, Tokyo, pp. 14–45.

Oden, N., 1995. Adjusting Moran's I for population density. Statist. Med. 14, 17–26.

Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. Geograph. Anal. 27, 286–306.

Orton, C.R., 1982. Stochastic process and archaeological mechanism in spatial analysis. J. Archaeol. Sci. 9, 1–23.

Park, S.K., Miller, K.W., 1988. Random number generators: good ones are hard to find. Commun. ACM 31, 1192–1201.

Rogerson, P.A., 1999. The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic. Geogr. Anal. 31, 130–147.

Sankoh, O.A., Ye, Y., Sauerborn, R., Muller, O., Becher, H., 2001. Clustering of childhood mortality in rural Burkina Faso. Int. J. Epidemiol. 30, 485–492.

Shaw, G.M., Selvin, S., Swan, S.H., Merrill, D., Schulman, J., 1988. An examination of three spatial disease clustering methodologies. Int. J. Epidemiol. 17, 913–919.

Sokal, R.R., Oden, N.L., 1978. Spatial autocorrelation in biology 1. methodology. Biol. J. Linnean Soc. 10, 199–228.

Swartz, J.B., 1998. An entropy-based algorithm for detecting clusters of cases and controls and its comparison with a method using nearest neighbours. Health Place 4, 67–77.

Tango, T., 1995. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Statist. Med. 14, 2323–2334.

Tango, T., 1998. S-Plus Code for the Maximized Excess Events Test. National Institute of Public Health, Tokyo.

Tango, T., 1999a. Comparison of general tests for spatial clustering. In: Lawson et al. (Eds.), Disease Mapping and Risk Assessment for Public Health. Wiley, London, pp. 111–117.

Tango, T., 1999b. Disease mapping and spatial disease clustering: toward an appropriate interpretation and use of disease indices. J. Nat. Inst. Public Health 48, 84–93.

Tango, T., 2000. A test for spatial disease clustering adjusted for multiple testing. Statist. Med. 19, 191–204.

Turnbull, B., Iwano, E.J., Burnett, W.S., Howe, H.L., Clark, L.C., 1990. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. Amer. J. Epidemiol. 132, S136–S143.

Vach, W., 1994. Locally optimal tests on spatial clustering'. In: Diday et al. (Eds.), New Approaches in Classification and Data Analysis, Springer, Berlin, pp. 161–168.

Viel, J.F., Arveux, P., Baverel, J., Cahn, J.Y., 2000. Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels. Amer. J. Epidemiol. 151, 13–19.

Walsh, S.J., Fenster, J.R., 1997. Geographical clustering of mortality from systemic sclerosis in the Southeastern United States, 1981–1990. J. Rheumatol. 2348–2352.

Waller, L.A., Jacquez, G.M., 1995. Disease models implicit in statistical tests of disease clustering. Epidemiology 6, 584–590.

Wartenberg, D., 1990. Detecting disease clusters: the importance of statistical power. Amer. J. Epidemiol. 132, S156–S166.

Zoellner, I.K., Schmidtmann, I.M., 1999. Empirical studies of cluster detection—different cluster tests in application to German cancer maps. In: Lawson et al. (Eds.), Disease Mapping and Risk Assessment for Public Health. Wiley, London.