17

Detection of Disease Clustering

Toshiro Tango

Department of Technology Assessment and Biostatistics, National Institute of Public Health, Wako-shi, Japan

Abstract: In epidemiological studies, it is often of interest to evaluate whether a disease is randomly distributed over time and/or space after being adjusted for a known heterogeneity, which may provide clues to the etiology of disease. To do this, we can apply tests for spatial randomness, or disease clustering. In this paper, I review the existing tests for disease clustering and discuss the advantages and disadvantages of these test statistics. These tests are illustrated and compared with several real temporal and spatial data sets.

Keywords and phrases: Cluster detection test, epidemiology, global clustering test, likelihood ratio, relative risk, spatial statistics

17.1 Introduction

There has been great public concern about the clustering of health events such as the occurrence of childhood leukemia, birth defects, and cancer. To investigate whether clustering is real and significant, many different tests have been proposed for different purposes. Besag and Newell (1991) classified these tests into two families: *focused tests* and *general tests*. The former family of tests assesses the clustering around a pre-fixed point like a nuclear installation. The latter is aimed at investigating the question of whether clustering occurs over the study region. *General tests* were further classified by Kulldorff (1998) into two groups: the first group, *global clustering tests* (GCTs), is designed for evaluating whether cases tend to come in groups or whether cases are located close to each other no matter when and where they occur, and the second group, *cluster detection tests* (CDTs), is designed to both detect local clusters and evaluate their significance. Recently, Kulldorff (2006) discussed the general framework into which most of the many different proposed test statistics for spatial randomness can be placed.

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_17,
© Birkhäuser Boston, a part of Springer Science+Business Media, LLC 2009

This paper is concerned with *general tests* and is organized as follows. Section 17.2 reviews tests for detecting temporal clustering, and Section 17.3 reviews tests for spatial clustering. This paper concludes with a discussion in Section 17.4.

17.2 Temporal Clustering

17.2.1 Disjoint tests

Ederer, Myers, and Mantel (1964) developed a GCT for temporal clustering using a cell-occupancy approach. They divided the time period into m disjoint subintervals. Under the null hypothesis of no clustering, the n cases are randomly distributed among the subintervals (i.e., are multinomially distributed). The test statistic M is the maximum number of cases occurring in a subinterval, i.e., $M = \max(n_1, ..., n_m)$. If the health event is rare and of unknown etiology, M is summed over several locations and time periods. The sum is tested by using a single degree of freedom chi-square test. Ederer, Myers, and Mantel (1964) and Mantel, Kryscio, and Myers (1976) provide tables of the exact null distribution of M for selected values of m and n.

17.2.2 Scan statistics for individual time points data

Naus (1965) proposed a CDT for temporal clustering that is known as the *scan statistic* and is applicable when individual time points data $(t_1, \ldots, t_n)$ are available during the study period. The test statistic S_d, the maximum number of cases observed in an interval of length d, is found by "scanning" all intervals of length d, known as the scanning window of fixed size d, in the time period. In certain cases, this approach is intuitively more appealing than the disjoint interval approach of Ederer, Myers, and Mantel (1964), but it is more complicated mathematically. A major challenge with the scan statistic has been to find analytical results concerning its statistical significance. Unfortunately, the computations necessary to obtain exact p-values for the scan statistic are complex and often not feasible. For selected interval lengths, time lengths, and sample sizes, the tables of p-values provided by Naus (1966) and Wallenstein (1980) can be used. Knox and Lancashire (1982) found a pragmatic approximation to the p-value but it was not so good. In 1987, Wallenstein and Neff proposed a simple but excellent approximation for small p-values such as $p < 0.10$. Let T denote the length of the entire study period and $w = d/T$. Then we have

$$\Pr\{S_d \geq k \mid n, T\} \approx (\frac{k}{w} - n - 1)b(k \mid n, w) + 2 \sum_{i=k+1}^{n} b(i \mid n, w), \qquad (17.1)$$

where

$$b(i \mid n, w) = \binom{n}{i} w^i (1-w)^{n-i}.$$

Although this formula often gives a poor approximation for larger p-values, it does not matter in terms of statistical significance. For example, when $n = 62$, $k = 7,\ d = 1,\ T = 24$ in examples of trisomy data, we have $p \approx 1.09 > 1$, indicating that the test result is not significant anyway.

Naus (1996) compared the power of the scan test with that of the Ederer, Myers, and Mantel (1964) test and concluded that if the scanning interval is small and the data are continuous over the interval, the scan test is the more powerful of the two. Weinstock (1981) proposed a generalization of the scan test that adjusts for changes in the population at risk. Later, Nagarwalla (1996) extended the scan statistic to one with a variable window, whose size does not need to be chosen *a priori*. Let $(t_1, ..., t_n)$ denote a random sample of n points from the density $f(t)$ in an interval $[0,\ T]$. For the hypothesis testing problem $H_0 : f(x) = 1/T,\ H_1 : f(x) = 1/T + \delta$ for $a \le x \le a + d$, the test is the maximized likelihood ratio test statistic λ, which allows for clusters of variable width d:

$$\lambda = \sup_{d,\ k \ge n_0} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \left(\frac{T}{d}\right)^k \left(\frac{T}{T-d}\right)^{n-k}, \qquad (17.2)$$

where $k = k(a, d)$ is the number of points in the window $(a,\ a + d]$. Nagarwalla gave a simple algorithm for the implementation of the method, but Monte Carlo hypothesis testing is used to obtain the p-value since it is not possible to obtain the null distribution of λ analytically.

17.2.3 Clustering index

Tango (1984) developed a GCT for temporal clustering based on the distribution of counts in m disjoint subintervals. However, it can provide a statistic to estimate the clustering periods which made large contributions to significant clustering. The test is useful when the data are grouped. The test statistic, known as a clustering index, is a quadratic form involving the relative frequencies in each subinterval and a measure of closeness between subintervals,

$$C = \boldsymbol{r}^t \mathbf{A} \boldsymbol{r} = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{n_i n_j}{n^2} a_{ij}, \quad (0 < C \le 1), \qquad (17.3)$$

where $\boldsymbol{r}^t = (n_1, ..., n_m)/n$ and the entries a_{ij} of the $m \times m$ symmetric matrix $\boldsymbol{A}$ are arbitrary known measures of closeness between the ith and jth subintervals with the property $a_{ii} = 1$ and where a_{ij} is a monotonically nonincreasing

function of d_{ij}, the time between the ith and jth subintervals. Tango used the following form as a natural choice:

$$a_{ij} = \exp(-d_{ij}) = \exp(- \mid i - j \mid).$$

The clustering index obtains a maximum value of 1 when all cases occur in the same subinterval. Although the statistic is easy to calculate, the proposed asymptotic null distribution was rather complex for simple use. Whittemore and Keller (1986) showed that the distribution of Tango's index is asymptotically normal with mean and variance that are simple to compute. However, later on, Tango (1990) showed that their normal approximation was very poor for moderately large sample sizes and suggested a central chi-square distribution with degrees of freedom ν adjusted by the skewness as a better approximation, i.e.,

$$\Pr\{C > c \mid H_0\} \approx \Pr\left\{\chi^2_\nu > \nu + \sqrt{2\nu}\left(\frac{c - E(C)}{\sqrt{Var(C)}}\right)\right\}, \tag{17.4}$$

where

$$\begin{aligned}
E(C) &= m^{-2}\{\mathbf{1}^t \boldsymbol{A}\mathbf{1} + n^{-1}\mathrm{tr}[\boldsymbol{AV}]\} \\
\mathrm{Var}(C) &= m^{-4}n^{-1}\{4\mathbf{1}^t \boldsymbol{AVA}\mathbf{1} + 2n^{-2}\mathrm{tr}[(\boldsymbol{AV})^2]\} \\
\nu &= 8/(\sqrt{\beta_1(C)})^2 \\
\sqrt{\beta_1(C)} &= \frac{8\{3\mathbf{1}^t(\boldsymbol{AV})^2\boldsymbol{A}\mathbf{1} + n^{-1}\mathrm{tr}[(\boldsymbol{AV})^3]\}}{\sqrt{n}\{4\mathbf{1}^t \boldsymbol{AVA}\mathbf{1} + 2n^{-1}\mathrm{tr}[(\boldsymbol{AV})^2]\}^{3/2}} \\
\mathbf{1} &= (1, ..., 1)^t \text{ (length } m) \\
\boldsymbol{V} &= \mathrm{diag}(m\mathbf{1}) - \mathbf{1}\mathbf{1}^t,
\end{aligned}$$

where $\mathrm{diag}(\boldsymbol{x})$ is the $m \times m$ diagonal matrix with the vector $\boldsymbol{x}$. If the null hypothesis of no clustering is rejected, we can apply the same idea adopted in the spatial clustering index [Tango (2000)], i.e., the most likely center of clustering period may be identified by the subinterval i with maximum of

$$U_i = \frac{1}{C}\sum_{j=1}^{m} \frac{n_i n_j}{n^2} a_{ij}, \quad (\sum_{i=1}^{m} U_i = 1), \tag{17.5}$$

which denotes the percentage of the ith subinterval's contribution to the significant clustering. Empirically, the subintervals with high outlying percentages will be likely periods of clusters.

17.2.4 Other methods

Bailar, Eisenberg, and Mantel (1970) suggested a GCT for detecting temporal clustering based on the number of pairs of cases in a given area that occur

within a specified length of time d of each other. The numbers of close pairs occurring in q areas are summed. The test statistic is assumed to be approximately normally distributed. Larsen, Holmes, and Heath (1973) developed a rank order GCT for detecting temporal clustering. The time period is divided into disjoint subintervals that are numbered sequentially (i.e., ranked). The test statistic K is the sum of absolute differences between the rank of the subinterval in which a case occurred and the median subinterval rank. Small values of K indicate unimodal clustering. Generally, the K statistics for multiple geographic areas are summed. The resulting statistic is asymptotically normal with simple mean and variance. This test is sensitive only to unimodal clustering; it cannot distinguish multiple clustering from randomness. Molinari, Bonaldi, and Daures (2001) proposed a CDT by applying a piecewise-constant regression model which allows for multiple cluster detection. They used the Akaike information criterion and the Bayesian information criterion to determine the optimal model including the number of clusters.

17.2.5 Illustration with congenital oesophageal atresia data

The data we use to illustrate several tests here consists of individual dates of birth of $n = 35$ cases of the birth defects oesophageal atresia and tracheo-oesophageal fistula observed in a hospital in Birmingham, U.K., from 1950 through 1955. The study was first published by Knox (1959) and subsequently analyzed by Weinstock (1981) using a scan statistic with a fixed window and by Nagarwalla (1996) using a scan statistic with a variable window. The data is shown in Table 17.1. The second column is the number of days past 1 January 1950 on which each case was observed. The third, fourth, and fifth columns of the table denote the frequency of cases per 100 days, 200 days, and 365 days (one year), respectively. Visual inspection of the data suggests that there occurs a clustering during three close subintervals [1200, 1299], [1300, 1399], [1400, 1499] and another less striking concentration occurs in the last three subintervals [1900, 1999], [2000, 2099], [2100, 2199]. We shall show the results of application of the scan statistic with a fixed window, the scan statistics with a variable window, and the clustering index.

1. Scan statistic with a fixed window $d = 100$

   $S_d = 7$ for the cluster of 7 cases from the day 1233 (17 May 1953) to the day 1305 (28 July 1953). Using the approximation (17.1) we obtain $p = 0.088$.

2. Scan statistic with a fixed window $d = 200$

   $S_d = 10$ for the cluster of 10 cases from the day 1233 (17 May 1953) to the day 1390 (21 October 1953). Using (17.1) we obtain $p = 0.0499$.

Table 17.1. $n = 35$ cases of oesophageal atresia and tracheo-oesophageal fistula over 2191 days from 1950 to 1955. Day 1 was set as *1 January 1950*. (Data from Knox, 1959)

Interval	Day number	Frequency per d days		
		$d = 100$	200	365
0–99		0		
100–199	170	1	1	
200–299		0		
300–399	316	1	1	2
400–499	445, 468	2		
500–599		0	2	
600–699		0		
700–799		0	0	2
800–899		0		
900–999	938	1	1	
1000–1099	1034	1		2
1100–1199	1128	1	2	
1200–1299	1233, 1248, 1249, 1252, 1259, 1267	6		
1300–1399	1305, 1385, 1388, 1390	4	10	
1400–1499	1446, 1454, 1458, 1461, 1491	5		14
1500–1599	1583	1	6	
1600–1699	1699	1		
1700–1799	1702, 1787	2	3	
1800–1899		0		6
1900–1999	1924, 1974	2	2	
2000–2099	2049, 2051, 2067, 2075	4		
2100–2199	2108, 2151, 2174	3	7	9
Total		35		

3. Scan statistic with a fixed window $d = 300$

   $S_d = 15$ for the cluster of 15 cases from the day 1233 (17 May 1953) to the day 1491 (30 January 1954). Using (17.1) we obtain $p = 0.0014$.

4. Scan statistic with a fixed window $d = 365$ [Weinstock (1981)]

   $S_d = 16$ for the cluster of 16 cases from the day 1233 (17 May 1953) to the day 1583 (2 May 1954). Using (17.1) we obtain $p = 0.0027$.

5. Scan statistic with a variable window [Nagarwalla (1996)]

   Results of four different scan statistics with fixed windows $d = 100, 200, 300$, and 365 suggest the optimal window could exist between 200

and 365. With $n_0 = 5$, the maximum likelihood ratio (17.2) is $\lambda^* = 43,968$, and the most likely cluster is the set of 15 cases from the day 1233 (17 May 1953) to the day 1491 (30 January 1954), which is the same as that of the scan statistic with fixed window $d = 300$. The optimal and minimum window is $1491 - 1233 + 1 = 259$. Using Monte Carlo testing with 9999 replicates, the observed rank of λ^* due to Nagarwalla's computation is 58, i.e., $p = 0.0058$.

6. Clustering index for the frequency data per 100 days

   Observed standardized clustering index is $c = 5.015$ and using the approximation (17.4) we obtain $p = 0.00027$. By examining the percent contribution U_i to C, we can see that three successive subintervals [1200, 1299], [1300, 1399], [1400, 1499] (15 cases from the day 1233 to the day 1491) have quite large values compared with those of other subintervals, and their contribution is 61.7%, indicating strong clustering period in these three successive subintervals. Furthermore, we can indicate another possible clustering period in two successive subintervals [2000, 2099], [2100, 2199] (7 cases from the day 2049 to the last day 2174) which contributed about 18.7%.

7. Clustering index for the frequency data per 200 days

   Observed standardized clustering index is $c = 5.222$ and using (17.4) we obtain $p = 0.0004$. By examining the percent contribution U_i to C, we can see a cluster in the two successive subintervals [1200, 1399], [1400, 1599] (16 cases from the day 1233 to the day 1583) which has 61.8% contribution. Furthermore, we can indicate another possible clustering period in the last subinterval [2000, 2199] (7 cases from the day 2049 to the last day 2174) which contributed about 18.0%.

8. Clustering index for the frequency data per one year (365 days)

   Observed standardized clustering index is $c = 4.745$ and using (17.4) we obtain $p = 0.0014$. By examining the percent contribution U_i to C, we can see a cluster in the subinterval [1095, 1459] (14 cases from the day 1128 to the day 1458) which has about 51.3% contribution. Furthermore, we can indicate another possible clustering period in the last subinterval [1825, 2190] (9 cases from the day 1924 to the last day 2174) which contributed about 23.6%.

17.2.6 Illustration with trisomy data

In this section, we shall consider a grouped data of $N = 62$ cases of trisomy among karyotyped spontaneous abortions of pregnancies, by calendar month of

the last menstrual period, July 1975 to June 1977, in three New York hospitals. This study was first analyzed by Wallenstein (1980) and subsequently by Tango (1984, 1990). The data is shown in Table 17.2. The trisomy data was tabulated in two ways: (1) monthly data over 24 months, (ii) bimonthly data over 24 months. Visual inspection of the data suggests that a cluster seems to occur during the period November 1976 to January 1977. The results are as follows.

1. Scan statistic [Wallenstein (1980)]

   Wallenstein (1980) applied the scan statistic with a fixed window to individual trisomy data (not shown in his paper). In his illustration, he set $d = 60$ days and found $S_d = 14$, $p = 0.038$ based on his unpub-

Table 17.2. Frequency of trisomy among karyotyped spontaneous abortions of pregnancies, by calendar month of the last menstrual period, July 1975 to June 1977, in three New York hospitals. (Data from Wallenstein, 1980; Tango, 1984)

Year	Month	Frequency	
		per month	per two months
1975	7	0	
	8	4	4
	9	1	
	10	2	3
	11	1	
	12	3	4
1976	1	1	
	2	3	4
	3	2	
	4	2	4
	5	3	
	6	4	7
	7	1	
	8	1	2
	9	1	
	10	2	3
	11	4	
	12	7	11
1977	1	7	
	2	2	9
	3	2	
	4	6	8
	5	1	
	6	2	3
Total		62	

lished extensive table. Linear interpolation based on his Table 17.1 yields $p = 0.040$. Using the approximation (17.1) we obtain $p = 0.037$. In this example, the maximum number of trisomies in two consecutive months was also 14. In general, inspection of *all* 60-day intervals may yield a higher value than the maximum number of two consecutive months.

2. Clustering index [Tango (1984, 1990)]

   All the following three results are significant at the 5% level: (i) for monthly data over 24 months, $C = 0.1139,\ p = 0.023$, (ii) for bimonthly data over 24 months, $C = 0.1975,\ p = 0.035$, and (iii) for monthly data over the last 12 months, $C = 0.2354,\ p = 0.0046$. Using U_i, we can find a likely cluster in the period from November 1976 to January 1977 which has 18 cases and 45.5% contribution.

3. Use of SaTScan

   SaTScan is a free software developed by Kulldorff *et al.* (2007) implementing several types of spatial, temporal, and space-time scan statistics. Purely temporal, analysis is essentially the same idea as Nagarwalla's scan statistic with a variable window for individual data. The details will be described in the next section. We shall show the results only for monthly data over 24 months. The most likely cluster is the set of 28 cases from November 1976 to April 1977. Using Monte Carlo testing with 999 replicates, the observed rank of the log-likelihood ratio statistic is 22, i.e., $p = 0.022$.

17.3 Spatial Clustering

For spatial analysis, it was/is sometimes practically impossible to obtain individual point location data in space due to confidentiality restrictions on individual privacy. Therefore, most tests for spatial clustering developed so far have been designed for regional count data. Although there are some important tests using individual point data or a sample of case-control location data, e.g., Cuzick and Edwards's test (1990) based on k-nearest neighbors and its generalized version by Tango (2007), in what follows, I shall confine myself to considering the situation where an entire study area is divided into m administrative regions (for example, county, census tract, block group) and the region $i(= 1, ..., m)$ has the observed number of cases n_i and the expected number of cases e_i under the null hypothesis of no clustering such that

$$n = \sum_{i=1}^{m} n_i = \sum_{i}^{m} e_i. \tag{17.6}$$

17.3.1 Tests based on adjacencies

Geary (1954) developed a test of spatial clustering that assesses whether rates for adjacent areas are more similar than would be expected if they were randomly distributed among the geographic areas. The test statistic is the ratio of the sum of mean squared differences between rates for pairs of adjacent areas to the weighted sum of mean-squared differences between rates for all pairs of areas. If the rates are geographically distributed at random, the test statistic is close to one; otherwise, it is less than one. Geary derived an expression for the approximate variance of the ratio. If the number of areas is not too small, the ratio is asymptotically normally distributed. Ohno, Aoki, and Aoki (1979) and Ohno and Aoki (1981) developed a simple test for spatial clustering that uses rates for geographic areas (e.g., census tracts, counties, or states) rather than data for individual cases. The test assesses whether the rates in adjacent areas are more similar than would be expected under the null hypothesis of no clustering. For this test, the rate for each area is classified into one of several categories, and each pair of adjacent areas is identified. The test statistic is the number of adjacent concordant pairs; i.e., the number of pairs of areas that are adjacent and have rates in the same category. An overall clustering measure summed across all categories can be obtained as well as category-specific clustering measures. The observed number of adjacent concordant pairs is compared with the expected number by using a chi-square test. Ohno, Aoki, and Aoki (1979) provide a simple formula for calculating the expected number of pairs. Grimson, Wang, and Johnson (1981) proposed a test of spatial clustering for use in detecting clusters of geographic areas designated as high risk. The null hypothesis is that high-risk areas are randomly distributed within a larger area and do not cluster. Given the number of high-risk areas, the test statistic is the number of pairs of high-risk areas that are adjacent to each other. This statistic is equivalent to the category-specific statistic from Ohno, Aoki, and Aoki (1979).

Note that, although these tests based on adjacencies are easy to use, they do not properly take the sampling variability of rates into account, and so they are not recommendable in the sense that they may produce spurious results in practice.

17.3.2 Tests based on scanning regions

As the first method using scanning local regional rates, Openshaw *et al.* (1988) developed a geographical analysis machine (GAM) that is an exploratory tool

for searching for potential clusters. GAM constructs overlapping circles of different radii centered at each grid point defined *a priori*, counts the number of cases and the number of people at risk within the circle, and displays those circles with local incidence proportions exceeding some predefined threshold. However, GAM has attracted much criticism since it produces large numbers of highly correlated ovelapped circles. Turnbull *et al.* (1990), on the other hand, proposed a more statistically sound cluster evaluation permutation procedure (CEPP), where, for each region, a window is constructed by absorbing the nearest neighboring regions such that each window contains just a pre-fixed population size R. These windows vary in geographic shape and size but maintain a constant population size at risk so that observed counts are identically distributed. However, these windows of cases and populations overlap, and the counts are not independently distributed. The test statistic of the CEPP is given by the maximum number of cases in the window, which is not necessarily integer due to the adjustment of each population size to R. Monte Carlo testing is needed to obtain the p-value for the test statistic.

Besag and Newell (1991) considered windows with a pre-fixed number of cases k rather than a pre-fixed population size. It was originally designed for quite rare diseases, and thus a typical value of k might be small such as $k = 2, 4, ...$ Each region with nonzero cases is considered in turn as the center of a possible cluster. When considering a particular region, we label it as region 0 and order the remaining regions by their distance to the region 0. We label these regions $j = 1, 2, ..., m - 1$ and define

$$D_i = \sum_{j=0}^{i} n_{(j)}, \ u_i = \sum_{j=0}^{i} \xi_{(j)},$$

where $n_{(j)}$ and $\xi_{(j)}$ denote the number of cases and population in the region labelled j, respectively. Then, the test statistic for detecting individual clusters is

$$S = \min\{i : D_i \geq k\}. \tag{17.7}$$

Namely, the nearest S regions contain the closest k cases. A small observed value of S indicates a cluster centered at region 0. The significance level for each potential cluster is

$$\Pr\{S \leq s\} = 1 - \sum_{t=0}^{k-1} \exp(-u_s Q)(u_s Q)^t / t!, \quad Q = n_+ / \xi_+. \tag{17.8}$$

As the test statistic of overall clustering within the entire study area, Besag and Newell (1991) suggested the total number T_{BN} of significant ($p < 0.05$, say) individual clusters. The significance of the observed T_{BN} may be determined by Monte Carlo simulation.

17.3.3 Spatial scan statistics

Kulldorff and Nagarwalla (1995) and Kulldorff (1997) proposed the spatial scan statistic, which is a spatial version of the scan statistic with a variable window size and is a generalization of CEPP. The spatial scan statistic imposes a circular window $\mathbf{Z}$ on each centroid of a region. For any of those centroids, the radius of the circle varies from zero to some preset upper limit. If the window contains the centroid of a region, then that whole region is included in the window. In total, a very large number of different but overlapping circular windows are created, each with a different location and size, and each being a potential cluster. Let $\mathbf{Z}_{ik},\ k = 1, \ldots, K_i,$ denote the window composed by the $(k-1)$-nearest neighbors to region i. Then, all the windows to be scanned by the spatial scan statistic are included in the set

$$\mathcal{Z}_1 = \{\mathbf{Z}_{ik} \mid 1 \leq i \leq m,\ 1 \leq k \leq K_i\}.$$

Under the alternative hypothesis, there is an elevated risk within some window $\mathbf{Z}$ as compared to outside:

$$\begin{aligned} H_0 &: E(N(\mathbf{Z})) = e(\mathbf{Z}), \quad \text{for all } \mathbf{Z}, \\ H_1 &: E(N(\mathbf{Z})) > e(\mathbf{Z}), \quad \text{for some } \mathbf{Z}, \end{aligned}$$

where $N()$ and $e()$ denote the random number of cases and the null expected number of cases within the specified window, respectively. For each window, it is possible to compute the likelihood to observe the observed number of cases within and outside the window, respectively. Under the Poisson assumption, which is a typical distribution for rare diseases, the test statistic is the likelihood ratio maximized for $\mathbf{Z}$:

$$\sup_{\mathbf{Z} \in \mathcal{Z}_1} \left(\frac{n(\mathbf{Z})}{e(\mathbf{Z})}\right)^{n(\mathbf{Z})} \left(\frac{n(\mathbf{Z}^c)}{e(\mathbf{Z}^c)}\right)^{n(\mathbf{Z}^c)} I\left(\frac{n(\mathbf{Z})}{e(\mathbf{Z})} > \frac{n(\mathbf{Z}^c)}{e(\mathbf{Z}^c)}\right), \tag{17.9}$$

where $\mathbf{Z}^c$ indicates all the regions outside the window $\mathbf{Z}$, and $n()$ denotes the observed number of cases within the specified window and $I()$ is the indicator function. The window $\mathbf{Z}^*$ that attains the maximum likelihood is defined as the *most likely cluster* (MLC). To find the distribution of the test statistic under the null hypothesis, Monte Carlo hypothesis testing is required. Kulldorff's spatial scan statistic has been applied to a wide variety of epidemiological studies and also to disease surveillance for the detection of disease clusters along with SaTScan Software (Kulldorff *et al.* 2007).

However, since it uses a circular window to scan the potential cluster areas, it has difficulty in correctly detecting actual noncircular clusters. To detect arbitrarily shaped clusters which cannot be detected by the circular spatial scan statistic, Patil and Taillie (2004), Duczmal and Assunção (2004), Tango

and Takahashi (2005), and Assunção *et al.* (2006) have proposed different spatial scan statistics. Patil and Taillie (2004) used the notion of "upper level set" to reduce the size of windows to be scanned and proposed the "upper level set scan statistic." However, they do not discuss how to select the level g which defines the upper level set. Duczmal and Assunção (2004), on the other hand, have applied a simulated annealing method, in which they try to examine only the most promising windows using a graph-based algorithm to obtain the local maxima of a certain likelihood function over a subset of the collection of all the connected regions. Their method seems to be very complicated, but they do not show any programmable procedure for it. Tango and Takahashi (2005) called their spatial scan statistic the *flexible spatial scan statistic* in contrast to Kulldorff's *circular spatial scan statistic* and provided FleXScan Software [Takahashi, Yokoyama, and Tango (2007)].

The *flexible spatial scan statistic* imposes an irregularly shaped window $\mathbf{Z}$ on each region by connecting its adjacent regions. For any given region i, we create the set of irregularly shaped windows with *length* k consisting of k connected regions including i and let k move from 1 to the preset maximum length of cluster K. To avoid detecting a cluster of *unlikely peculiar shape*, the connected regions are restricted as the subsets of the set of regions i and $(K-1)$-nearest neighbors to the region i. In total, as in the circular spatial scan statistic, a very large number of different but overlapping arbitrarily shaped windows are created. Let $\mathbf{Z}_{ik(j)},\ j = 1, \ldots, J_{ik}$ denote the jth window which is a set of k regions connected starting from the region i, where J_{ik} is the number of j satisfying $\mathbf{Z}_{ik(j)} \subseteq \mathbf{Z}_{ik}$ for $k = 1, \ldots, K_i = K$. Then, all the windows to be scanned are included in the set

$$\mathcal{Z}_2 = \{\mathbf{Z}_{ik(j)} \mid 1 \le i \le m,\ 1 \le k \le K,\ 1 \le j \le J_{ik}\}. \tag{17.10}$$

In other words, for any given region i, the circular spatial scan statistic considers K concentric circles, whereas the flexible scan statistic considers K concentric circles plus all the sets of connected regions (including the single region i) whose centroids are located within the Kth largest concentric circle. So, the size of $\mathcal{Z}_2$ is far larger than that of $\mathcal{Z}_1$, which is at most mK. Under the Poisson assumption, the test statistic is the same form as (17.9) where $\mathcal{Z}_1$ is replaced by $\mathcal{Z}_2$.

17.3.4 Clustering index

Tango (1995) proposed the following test statistic for spatial disease clustering:

$$\begin{aligned} C &= (\boldsymbol{r} - \boldsymbol{p})^t \boldsymbol{A} (\boldsymbol{r} - \boldsymbol{p}) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\frac{n_i - e_i}{n} \right) \left(\frac{n_j - e_j}{n} \right) a_{ij}, \end{aligned} \tag{17.11}$$

where $\boldsymbol{r}^t = (n_1, ..., n_m)/n$ denotes a vector of the observed relative frequencies, $\boldsymbol{p} = E_{H_0}(\boldsymbol{r})$, and $e_i = np_i,\ i = 1, ..., m$. This is a generalization of his temporal clustering index in that it allows for heterogeneous population size and confounding factors based on indirect standardization. Namely, let us partition the population into K categories and let n_{ik} and ξ_{ik} denote the observed number of cases and the population size, respectively, in the kth category of the confounding factor of the ith region. Then, we have

$$\boldsymbol{p} = \sum_{k=1}^{K} \frac{n_{+k}}{n} \boldsymbol{p}_k = \sum_{k=1}^{K} \frac{n_{+k}}{n} (p_{1k}, ..., p_{mk})^t, \tag{17.12}$$

where $p_{ik} = \xi_{ik} / \sum_{j=1}^{m} \xi_{jk}$. As a measure of closeness, $a_{ij}(\lambda)$, between the regions i and j, Tango (1995, 2000) recommended the double exponential form:

$$a_{ij} = \exp\left\{-4\left(\frac{d_{ij}}{\lambda}\right)^2\right\}, \tag{17.13}$$

where λ is a measure of *cluster size* and is essentially equal to the maximum distance between cases, such that any pair of cases far apart beyond the distance λ cannot be considered as a cluster. Large λ will give a test sensitive to a large cluster and small λ to a small cluster. In practical application, it is rare that we can predict the cluster size before examining data. Therefore, we usually repeat the procedure using different parameter settings and, consequently, face multiple testing problems. To take this problem into account, Tango (2000) propose, as an extended test statistic, *the minimum of the profile P-value of C for* λ where λ varies continuously from a small value near zero upwards until λ reaches about one-fourth the maximum distance d_{ij} in the study area. The proposed test statistic P_{min} is defined as

$$P_{min} = \min_{\lambda} \Pr\{C > c \mid H_0, \lambda\} = \Pr\{C > c \mid H_0, \lambda = \lambda^*\}, \tag{17.14}$$

where λ^* attains the minimum p-values of C. A practical implementation of this procedure is to use "line search" by discretization of λ. The null distribution of P_{min} can be obtained by using Monte Carlo simulation. This test is also called Tango's MEET (maximized excess event test) in the literature [e.g., Kulldorff *et al.* (2003, 2006); Song and Kulldorff (2003, 2005)].

Given λ and under the null hypothesis H_0, the test statistic C was shown to be asymptotically approximated by the same type of chi-square distribution as (17.4), where

$$
\begin{aligned}
E(C) &= n^{-1}tr(\boldsymbol{AV}) \\
Var(C) &= 2n^{-2}tr(\boldsymbol{AV})^2 \\
\nu &= 8/\{\sqrt{\beta_1(C)}\}^2 \\
\sqrt{\beta_1(C)} &= 2\sqrt{2}tr(\boldsymbol{AV})^3/\{tr(\boldsymbol{AV})^2\}^{3/2} \\
\boldsymbol{V} &= \sum_{k=1}^{K} \frac{n_{+k}}{n}\{\mathrm{diag}(\boldsymbol{p}_k) - \boldsymbol{p}_k\boldsymbol{p}_k^t\}.
\end{aligned}
$$

This chi-square approximation is generally quite accurate even for small n. If the null hypothesis of no clustering is rejected, we can use a statistic similar to (17.5) to indicate *the most likely center* i of clustering area with large values of

$$
U_i = \frac{1}{C}\sum_{j=1}^{m}\left(\frac{n_i - e_i}{n}\right)\left(\frac{n_j - e_j}{n}\right)a_{ij}, \tag{17.15}
$$

which denote the percentage of the ith region's contribution to the significant clustering. More specifically, we may use the following condition of standardized U_i to suggest the center of clustering areas:

$$
(U_i - \bar{U})/\mathrm{SD}_U \geq 2.0 \text{ or } 3.0.
$$

17.3.5 Other methods

Whittemore *et al.* (1987) developed a test statistic for spatial clustering,

$$
W = \boldsymbol{r}^t\boldsymbol{D}\boldsymbol{r},
$$

which is identical in form to Tango's clustering index C (17.3), but for which $\boldsymbol{D} = (d_{ij})$ is used as a measure of distance. They proved the asymptotic distribution of this index to be normal and insisted that the clustering index C (17.3) also has an asymptotic normal distribution. However, it does depend largely on the element $\boldsymbol{A}$ or $\boldsymbol{D}$ used. When the distance measure $\boldsymbol{D}$ is used, convergence to normality is very fast. On the contrary, when the closeness measure $\boldsymbol{A}$ is used, the speed is shown to be too slow, and thus normality is not valid even for fairly large sample sizes such as $n = 1000$ [Tango (1986, 1990)]. Furthermore, more substantially, it has been shown that (1) the quadratic form in $(\boldsymbol{r} - \boldsymbol{p})$ should be used to properly adjust for heterogeneous populations, and (2) the power of W often falls below the nominal α level depending on the clustering models due to the use of distance measure $\boldsymbol{D}$ [Tango (1995, 1999)]. Therefore, the test of Whittemore *et al.* cannot be recommended for practical use. Bonetti and Pagano (2005) proposed a test using the interpoint distance distribution for spatial clustering, but it generally does not perform quite as well as the spatial scan statistic and Tango's clustering index [Kulldorff *et al.* (2003), Song and Kulldorff (2003)].

17.3.6 Illustration with gallbladder cancer mortality data

As an illustration, we shall apply three tests, 1) circular spatial scan statistic, 2) flexible spatial scan statistic, and 3) spatial clustering index, to the mortality data from gallbladder cancer (male, 1993–1997) in the areas of three adjacent prefectures (Niigata, Fukushima, and Yamagata) in Japan. The total observed number of deaths for five years was 665 in this area with $m = 246$ regions (cities and villages). Before applying these three tests for spatial clustering, we drew a disease map based on the standardized mortality ratio (SMR) in Figure 17.1, which shows the maximum likelihood estimates for the relative risks. No clear spatial pattern emerges from this map. SMRs are commonly used in disease mapping, but they are very unstable in the sense that they can yield large changes in estimate with relatively small changes in expected number of cases. So, to overcome the drawbacks of the SMRs in disease mapping, Bayesian approaches have been used to obtain more smoothed estimates [for example, see Lawson, Browne, and Vidal Rodeiro (2003)]. In this paper we shall omit Bayes estimates of for disease mapping.

The results of Kulldorff's circular spatial scan statistic and Tango–Takahashi's flexible spatial scan statistic are shown in Figure 17.2 and Figure 17.3, respectively, where $K = 20$. The most likely cluster and the secondary cluster detected by the flexible spatial scan statistic are very similar to, but have a slightly different shape than, those of the circular spatial scan statistic. Regarding the application of Tango's clustering index, we took a sequence of

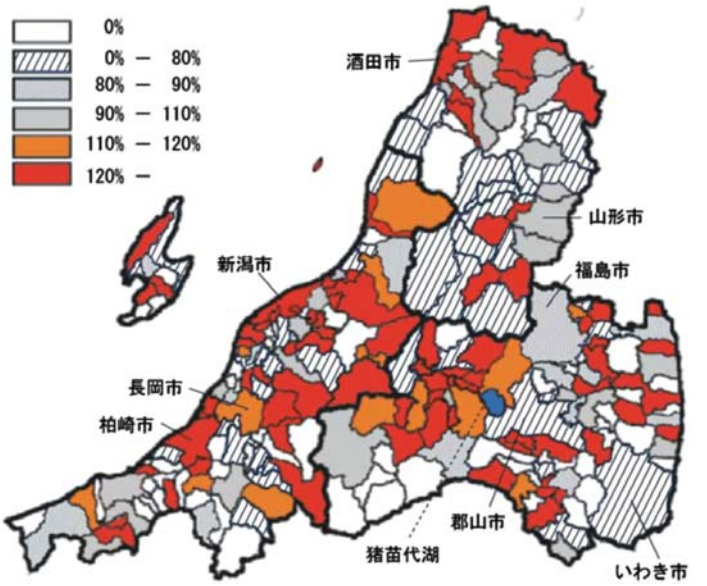

Figure 17.1. The SMRs of gallbladder cancer (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan (1996–2000).

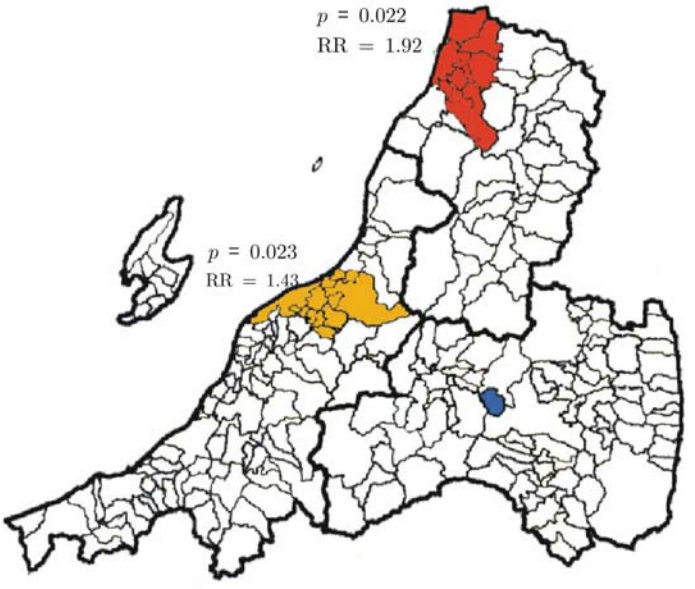

Figure 17.2. The most likely cluster (shaded area) and the secondary cluster (a lighter shaded area) detected by SaTScan for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan.

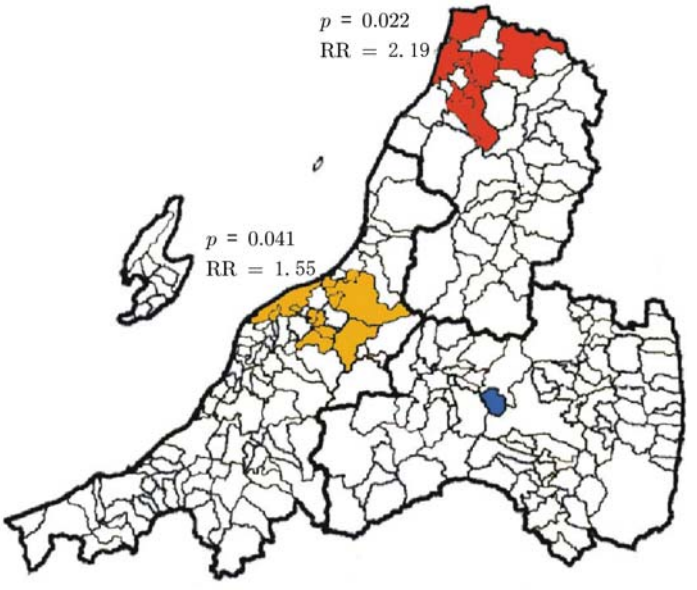

Figure 17.3. The most likely cluster (shaded area) and the secondary cluster (a lighter shaded area) detected by FleXScan for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan.

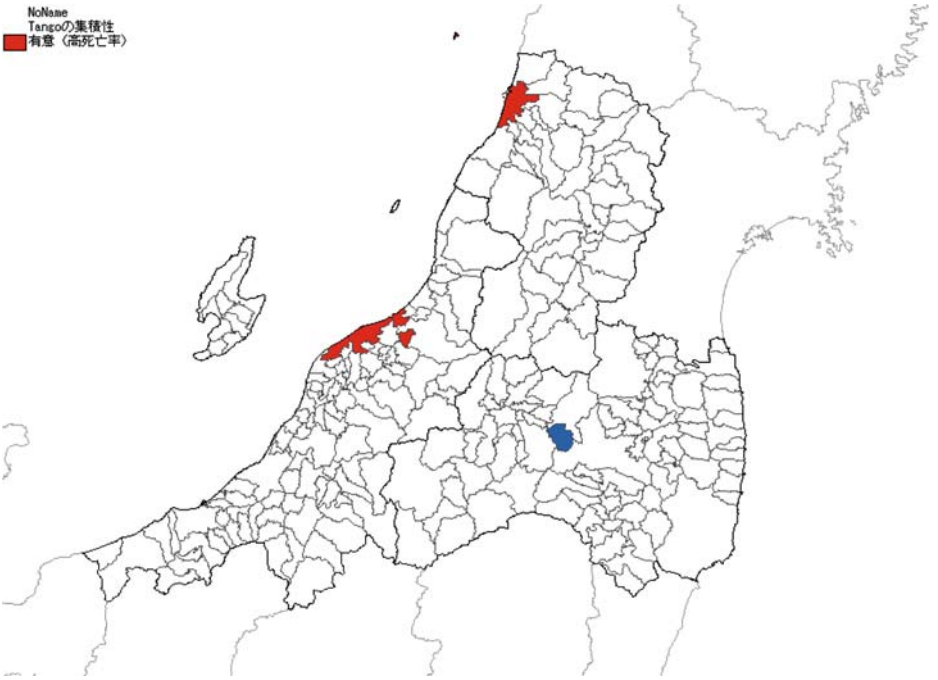

Figure 17.4. Two centers of clustering areas (shaded area) detected by Tango's spatial clustering index for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan.

values of cluster size λ as $\lambda = 0.1, 5, 10, 15, ..., 100$ (km) to obtain the test statistic

$$P_{min} = \min_{\lambda \in \{0.1,5,10,...,100\}} \Pr\{C > c \mid H_0, \lambda\}$$

and obtained $P_{min} = 0.00004$ at $\lambda = 45$. This P_{min} value is the second largest among 999 Monte Carlo replicates and, therefore, the adjusted p-value of P_{min} was $2/(999+1) = 0.002$. As possible centers of clusters, regions with standardized $U_i \geq 2.0$ are indicated in Figure 17.4, and these regions are found to be included in the most likely cluster and secondary cluster detected by both the circular scan statistic and flexible scan statistic.

17.4 Discussion

Many different test statistics have been designed for detecting disease clustering in time and in space. Most tests proposed before 1995, however, suffer from multiple testing problems due to one or two unknown parameters that must

be set prior to their applications. For example, Naus's scan statistic (1965) for individual time points data has an unknown length d of the scanning window, the procedure by Turnbull *et al.* (1990) has an unknown parameter regarding the common size of the population at risk R, Cuzick and Edwards's test (1995) has an unknown number of k-nearest-neighbors, and Besag and Newell's test (1991) has an unknown number of cases k for the size of the cluster. However, tests proposed in recent years tend to take such multiple testing into account. For example, such tests include Nagarwalla's scan statistic with variable window (1996), Kulldorff's spatial scan statistic (1997), Tango and Takahashi's flexible spatial scan statistic (2005), and Tango's clustering index (2000), where we have only to specify the maximum possible cluster size.

In recent power comparisons of disease clustering tests including CDTs and GCTs by Kulldorff *et al.* (2003) and Song and Kulldorff (2003), 1) Kulldorff's circular spatial scan statistic is shown to be the most powerful for detecting localized clusters, and 2) Tango's clustering index is the most powerful for general clustering throughout the study area. Note, however, that the power estimates provided reflect only the "power to reject the null hypothesis for whatever reason" and that the probability of both rejecting the null hypothesis and detecting the true cluster correctly is a different matter. To investigate *the performance of power* of the CDT, Tango and Takahashi (2005) proposed a new bivariate power distribution $P(l, s)$, which is the probability that the significant MLC has length $l(\geq 1)$ and includes s regions within the true cluster with length s^*. The usual power is defined by $\sum_l \sum_{s=1}^{s^*} P(l, s)$. Our simulation study using $P(l, s)$ revealed that the circular spatial scan statistic shows a high level of accuracy in detecting circular clusters exactly and reasonably good power for including some true cluster regions into the MLC. However, the circular spatial scan statistic is also shown to have a tendency to detect a cluster much larger than the true cluster assumed in the simulation, even when the true cluster is circular. The flexible spatial scan statistic, on the other hand, exhibits no such high power regarding exact identification of clusters, but the support of the power distribution is shown to be concentrated in a relatively narrow range of length l on the line $s = s^*$, indicating that an observed significant MLC contains the true cluster with quite high probability.

Tango and Takahashi (2005) have also shown examples which cast a doubt on the validity of the model selection based on maximizing the likelihood ratio: Duczmal and Assunção's procedure (2004) detected a quite large and peculiar shaped MLC that had the largest likelihood ratio among the three different MLCs, identified by three different spatial scan statistics, Kulldorff's (1997), Duczmal and Assunção's (2004), and Tango and Takahashi's (2005). Such a doubt can also be seen in the above-stated simulation results of the circular spatial scan statistic that had nonnegligible probabilities of detecting much longer clusters, than the true cluster. The flexible spatial scan statistic, on the other hand, is shown not to detect such an unexpected long cluster, probably

because it has the restriction that our windows are constructed only from members of the $(K-1)$-nearest neighbors to the starting region. Nevertheless, these undesirable properties produced by the maximum likelihood ratio might suggest the use of a different criterion for model selection.

In this chapter, we did not include tests for space-time disease clustering due to the limitation of space. As far as I know, Kulldorff (2001) proposed a procedure for prospective time periodic geographical disease surveillance using a scan statistic for the first time. In the aftermath of the World Trade Center attacks on September 11, 2001 and the anthrax-laden letters that followed in October 2001, a syndromic surveillance has been poised for deployment across the USA [Lawson and Kleinman, (2005)]. Therefore, statistical methods for timely detection of an outbreak threat, which are closely related to tests for space-time clustering, will be increasingly needed.

References

1. Assunção R., Costa M., Tavares A. and Ferreira S. (2006). Fast detection of arbitrarily shaped disease clusters, *Statistics in Medicine*, **25**, 723–742.

2. Bailar III J.C., Eisenberg H. and Mantel N. (1970). Time between pairs of leukemia cases, *Cancer*, **25**, 1301–1303.

3. Besag J.E. and Newell J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society, Series A*, **154**, 143–155.

4. Bonetti M. and Pagano M. (2005). The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering, *Statistics in Medicine*, **24**, 753–773.

5. Cuzick J.C. and Edwards R. (1990). Spatial clustering for inhomogeneous populations, *Journal of the Royal Statistical Society, Series B*, **52**, 73–104.

6. Duczmal L. and Assunção R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped clusters, *Computational Statistics and Data Analysis*, **45**, 269–286.

7. Dwass M. (1957). Modified randomization test for nonparametric hypotheses, *Annals of Mathematical Statistics*, **28**, 181–187.

8. Ederer F., Myers M.H. and Mantel N. (1964). A statistical problem in space and time: do leukemia cases come in clusters? *Biometrika*, **20**, 626–638.

9. Geary R.C. (1954). The contiguity ratio and statistical mapping, *The Incorporated Statistician*, **5**, 115–145.

10. Grimson R.C., Wang K.C. and Johnson P.W.C. (1981). Searching for hierarchical clusters of disease: spatial patterns of sudden infant death syndrome, *Social Science & Medicine*, **15D**, 287–293.

11. Knox G. (1959). Secular pattern of congenital oesophageal atresia, *British Journal of Preventive Social Medicine*, **13**, 222–226.

12. Knox E.G. and Lancashire R. (1982). Detection of minimal epidemics, *Statistics in Medicine*, **1**, 183–189.

13. Kulldorff M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

14. Kulldorff M. (1998). Statistical methods for spatial epidemiology: tests for randomness. In *GIS and Health*, (Ed., Gatrell A. and Loytonen M.), 49–62, Taylor & Francis, London.

15. Kulldorff M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society, Series A*, **164**, 61–72.

16. Kulldorff M. (2006). Tests for spatial randomness adjusted for an inhomogeneity: a general framework, *Journal of American Statistical Association*, **101**, 1289–1305.

17. Kulldorff M. and Nagarwalla N. (1995). Spatial disease clusters: detection and inference, *Statistics in Medicine*, **14**, 799–810.

18. Kulldorff M., Tango T. and Park P.J. (2003). Power comparisons for disease clustering tests, *Computational Statistics and Data Analysis*, **42**, 665–684.

19. Kulldorff M. and Information Management Services, Inc. (2007). SaTScan v7.0: Software for the spatial and space-time scan statistics, http://www.satscan.org/

20. Larsen R.J., Holmes C.L. and Heath C.W. (1973). A statistical test for measuring unimodal clustering: a description of the test and of its application to cases of acute leukemia in metropolitan Atlanta, Georgia, *Biometrics*, **29**, 301–309.

21. Lawson A.B., Browne W.J. and Vidal Rodeiro C.L. (2003). *Disease Mapping with WinBUGS and MLwiN*, John Wiley & Sons, Chichester.

22. Lawson A.B. and Kleinman K. (eds.) (2005). *Spatial & Syndromic Surveillance for Public Health*, John Wiley & Sons, New York.

23. Mantel N., Kryscio R.J. and Myers M.H. (1976). Tables and formulas for extended use of the Ederer-Myers-Mantel disease clustering procedure, *American Journal of Epidemiology*, **104**, 576–584.

24. Molinari N., Bonaldi, C. and Daures, J.P. (2001). Multiple temporal cluster detection. *Biometrics*, **57**, 577–583.

25. Nagarwalla N. (1996). A scan statistic with a variable window. *Statistics in Medicine*, **15**, 845–850.

26. Naus J.I. (1965). The distribution of the size of the maximum cluster of points on a line, *Journal of the American Statistical Association*, **60**, 532–538.

27. Naus J.I. (1966). A power comparison of two tests of non-random clustering, *Technometrics*, **8**, 493–517.

28. Ohno Y., Aoki K. and Aoki N. (1979). A test of significance for geographic clusters of disease, *International Journal of Epidemiology*, **8**, 273–281.

29. Ohno Y. and Aoki K. (1981). Cancer deaths by city and county in Japan: a test of significance for geographic clustering of disease, *Social Science & Medicine*, **15D**, 251–258.

30. Openshaw S., Craft A.W., Charlton M. and Birth J.M. (1988). Investigation of leukemia clusters by use of a geographical analysis machine, *Lancet*, **1(8580)**, 272–273.

31. Patil G. P. and Taillie C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics*, **11**, 183–197.

32. Song C. and Kulldorff M. (2003). Power evaluation of disease clustering tests, *International Journal of Health Geographics*, **2**, 9.

33. Song C. and Kulldorff M. (2005). Tango's maximized excess events test with different weights, *International Journal of Health Geographics*, **4**, 32.

34. Takahashi K., Yokoyama T. and Tango T. (2007). FleXScan: Software for the Flexble Scan Statistic. v2.0. http://www.niph.go.jp/soshiki/gijutsu/index_e.html/.

35. Tango T. (1984). The detection of disease clustering in time, *Biometrics*, **40**, 15–26.

36. Tango T. (1990). Asymptotic distribution of an index for disease clustering, *Biometrics*, **46**, 351–357.

37. Tango T. (1995). A class of tests for detecting "general" and "focused" clustering of rare diseases, *Statistics in Medicine*, **14**, 2323–2334.

38. Tango T. (1999). Comparison of general tests for disease clustering, In *Disease Mapping and Risk Assessment for Public Health*, (Ed., A.B. Lawson *et al.*), pp. 111–117, Wiley & Sons, New York.

39. Tango T. (2000). A test for spatial disease clustering adjusted for multiple testing, *Statistics in Medicine*, **19**, 191–204.

40. Tango T. and Takahashi K. (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, 11. http://www.ij-healthgeographics.com/content/4/1/11.

41. Tango T. (2007). A class of multiplicity adjusted tests for spatial clustering based on case-control point data, *Biometrics*, **63**, 119–127.

42. Turnbull B.W., Iwano E.J., Burnnett W.S., Howe H.L. and Clark LC. (1990). Monitoring for clusters of disease: application to leukemia incidence in upstate New York, *American Journal of Epidemiology*, **132**, suppl. S136–143.

43. Wallenstein S. (1980). A test for detection of clustering over time, *American Journal of Epidemiology*, **111**, 367–372.

44. Wallenstein S. and Neff N. (1987). An approximation for the distribution of the scan statistic. *Statistics in Medicine*, **6**, 197–207.

45. Weinstock M.A. (1981). A generalized scan statistic test for the detection of clusters, *International Journal of Epidemiology*, **10**, 289–93.

46. Whittemore A. and Keller J.B. (1986). A letter to the editor. On Tango's index of disease clustering in time, *Biometrics*, **42**, 218.

47. Whittemore A.S., Friend N., Brown B.W. and Holly E.A. (1987). A test to detect clusters of disease, *Biometrika*, **74**, 631–635.