STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3929–3943
Published online 24 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2490

An elliptic spatial scan statistic

Martin Kulldorff[1,*,†], Lan Huang[2,3,‡], Linda Pickle[3,§] and Luiz Duczmal[4,¶]

[1]*Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, 133 Brookline Avenue, 6th Floor, Boston, MA 02215, U.S.A.*
[2]*Department of Statistics, University of Connecticut, U.S.A.*
[3]*Statistical Applications and Research Branch, Division of Cancer Control and Population Studies, National Cancer Institute, U.S.A.*
[4]*Departamento de Estatistica, Universidade Federal de Minas Gerais, Brazil*

SUMMARY

The spatial scan statistic is commonly used for geographical disease cluster detection, cluster evaluation and disease surveillance. The most commonly used shape of the scanning window is circular. In this paper we explore an elliptic version of the spatial scan statistic, using a scanning window of variable location, shape (eccentricity), angle and size, and with and without an eccentricity penalty. The method is applied to breast cancer mortality data from Northeastern United States and female oral cancer mortality in the United States. Power comparisons are made with the circular scan statistic. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: disease surveillance; clusters; clustering; spatial statistics; eccentricity penalty; statistical power

1. INTRODUCTION

Spatial and space–time scan statistics [1, 2] are commonly used in disease surveillance for geographical cluster detection and evaluation, for which they have been shown to have good statistical power [3, 4]. Recent studies includes among many others its use for the surveillance of lateral sclerosis mortality in Finland [5], breast cancer mortality in Texas [6], late stage breast and colorectal cancer in Minnesota [7], congenital anomalies in Ireland [8], the prevalence of giardiasis intestinal parasites in Ontario [9], listeriosis incidence in New York State

*Correspondence to: Martin Kulldorff, Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, 133 Brookline Avenue, 6th Floor, Boston, MA 02215, U.S.A.
†E-mail: martin_kulldorff@hms.harvard.edu
‡E-mail: huangla@mail.nih.gov
§E-mail: picklel@mail.nih.gov
¶E-mail: duczmal@est.ufmg.br

Contract/grant sponsor: National Cancer Institute

Received 6 May 2004
Accepted 10 November 2005

[10], chronic wasting disease among white-tailed deer in Wisconsin [11], and hospital emergency visits and West Nile virus in New York City [12, 13]. Whether a disease cluster is due to environmental risk factors, the genetic makeup of the population, differences in behavioural risk factors, the spread of an infectious agent or a bioterrorism attack, the scan statistic can assist in the detection of disease clusters. This allows health officials to investigate disease outbreaks in a timely fashion, and if needed, rapidly implement disease prevention and control measures.

With the spatial scan statistic, a window of variable shape and size moves across a geographical region. Each shape, size and location defines a candidate cluster area. For each candidate area, the likelihood is calculated based on the observed and expected number of cases inside and outside that area. The area with the maximum likelihood defines the most likely cluster, that is, the cluster least likely to have occurred by chance. The statistical significance of this cluster is determined by generating a large number of random data sets under the null hypothesis of no clustering, and then calculating the maximum likelihood for each random data set in exactly the same manner as for the real data. If the maximum likelihood for the real data is ranked among the 5 per cent highest, the cluster is significant at the 5 per cent significance level. Important features of the spatial scan statistic are that it adjusts for the uneven spatial distribution of the population at risk, for covariates such as age, and for the multiple testing inherent in the large number of candidate cluster areas considered. Mathematical details, including its derivation as a likelihood ratio test, have been provided by Kulldorff [14].

When applying the spatial scan statistic, a natural choice of window shape is the circle, as it is the most compact shape that can be obtained. This is also the shape that has been used in practice. Other shapes are also possible though, such as ellipses, squares or triangles. These may have higher power if the true cluster shape is non-circular, which one would often expect to be the case. In this paper we evaluate the use of an elliptic spatial scan statistic, with and without an eccentricity (non-compactness) penalty. The method is applied to breast cancer mortality in northeastern United States and female oral cancer mortality in the United States. Power comparisons are made with the circular spatial scan statistic.

2. AN ELLIPTIC SPATIAL SCAN STATISTIC

The elliptic spatial scan statistic is a special case of the spatial scan statistic described by Kulldorff [14], as the mathematical principles behind the spatial scan statistics are identical for circular, elliptic or any other shape of the window, the only difference being the collection of candidate cluster areas considered.

An ellipse can be uniquely defined by five parameters: the x and y coordinates of its centroid, and its shape (eccentricity), angle, and size. We define the shape of the ellipse as the ratio of the length of the semimajor axis to the length of the semiminor axis, that is, it is the ratio of the longest to the shortest axis of the ellipse. A large number indicates a long and narrow ellipse while a shape of 1 gives the circle as a special case. The shape s is related to the eccentricity e of the ellipse through the formula $s = 1/\sqrt{1-e^2}$. The parameter θ is the angle between the horizontal line and the semimajor axis of the ellipse.

For both scientific and computational reasons, we will not consider all possible ellipses. First of all, for computational reasons, we will only consider a finite set of ellipse centroid

Table I. The per cent of the area of one ellipse that is also part of another ellipse with the same centre, shape and size, but with different angles θ° apart. The shape of the ellipse is defined as the ratio of the length of its longest (semimajor) to its shortest (semiminor) axis. The numbers with about three times as many different angles as the ratio of the ellipse length to width, are shown in bold face.

	Ellipse shapes				
# of angles (θ)	1.5	2	4	6	8
2 (90°)	75	59	31	21	16
4 (45°)	**82**	67	41	29	22
6 (30°)	87	**77**	52	38	30
9 (20°)	91	84	64	50	41
12 (15°)	93	88	**71**	59	49
15 (12°)	94	90	76	65	56
18 (10°)	95	92	80	**70**	62
24 (7.5°)	95	94	85	77	**70**
30 (6°)	97	95	88	81	75
36 (5°)	98	95	90	84	79
45 (4°)	98	97	92	87	83
60 (3°)	99	98	94	90	87
90 (2°)	99	98	96	94	91
180 (1°)	100	99	98	97	96
360 (0.5°)	100	100	99	98	98

coordinates. The centroid coordinates could be a regular grid with points that are located a fixed distance apart or, as in the application below, identical to the county centroids.

Next, we need to select a collection of ellipse shapes. An ellipse with shape equal to one defines the circle as a special case and we recommend always including the circle in any elliptic analysis. One possible collection of ellipse shapes is 1, 1.5, 2, 3, 4, 5, 6, 8, 10, 15, 20, 30, 60 and 120. Long and narrow ellipses may not always be of interest for scientific reasons. As an example, with spatially aggregated data, a very long and narrow cluster could include one county in Florida, one in Missouri, one in Montana and one in Alaska, and no other, and that is usually not what we consider to be a geographical cluster. In the subsequent analyses, we will test different subsets of the above-mentioned shapes, by selecting one of them as the maximum and including all smaller (less eccentric) shapes on the list as well.

For each ellipse shape a specific number of different angles is considered. For a compact elliptic shape, such as for example 1.5, the county centroids included in the ellipse do not change much with a slight change in the angle. In the extreme, the county centroids included in an ellipse with a shape of 1 (= circle) do not depend on the angle at all. For a long and narrow ellipse though, a small change in the angle will result in a very different set of county centroids. Hence, it is logical to use a larger collection of angles when the ellipse shape is long and narrow. To evaluate the suitable parameter values, we calculated the per cent overlap between two ellipses with the same ellipse centroid, shape and size but a specific angle apart. These numbers are presented in Table I.

We recommend using at least three times as many different angles as the ratio of the ellipse length to width. These numbers are shown in bold face in Table I, and we can see that it leads to an overlap between neighbouring ellipses of about 70 per cent. Other choices are also possible.

For each centroid, shape and angle, we will consider an infinite number of continuously increasing ellipse sizes from zero up to an upper limit such that for example at most 50 per cent of the total population is included within the ellipse. If the cluster is larger than 50 per cent, then we are not really evaluating a cluster with excess disease within the ellipse, but rather a cluster outside the ellipse with fewer disease cases than expected. This area outside the ellipse could be highly irregular, creating one 'cluster' consisting of for example, Alaska, Hawaii, Maine and Florida in a United States analysis. If clusters with fewer cases than expected are of interest, it is better to do a two-sided test, looking for ellipses with either a high or low rate of disease.

All this means that an infinite number of overlapping ellipses of different location, shape, angle and size are considered as candidate cluster areas. The exact choice is by nature somewhat arbitrary and will depend on the data being analysed, the questions being asked from the data and computational resources. It is critical though that the choice of the collection of ellipses to use is decided before looking at the observed case data. If the choice is made to fit an already observed cluster, then there will be pre-selection bias and the statistical inference will be invalid.

3. BREAST CANCER MORTALITY IN NORTHEASTERN UNITED STATES

To test the elliptic scan statistic, we first applied it to breast cancer mortality data from Northeastern United States, 1988–1992. This data has been previously analysed using the circular spatial scan statistic [15] as well as the circle-based isotonic spatial scan statistic [16]. The data set encompasses the years 1988–1992 and covers the 245 counties and county equivalents in Connecticut, Delaware, District of Columbia, Maine, Maryland, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. There were a total of 58 943 breast cancer deaths among a population of 29 535 210 women. The annual mortality rate was 39.9 per 100 000 women. Following Kulldorff *et al.* [15], all analyses in this paper use the Poisson model, where the observed number of deaths in a county is modelled according to Poisson distribution. All analyses are adjusted for age using indirect standardization and 18 different five year age groups: 0–4, 5–9, ..., 80–84, and 85+.

We analysed the data using ellipses with the set of centroid co-ordinates equal to the county centroids and a maximum cluster size of 50 per cent of the total population. Different maximum shapes were tried, equal to 1, 1.5, 2, 3, 4, 5, 6, 8, 10, 15, 20, 30, 60 and 120, including all smaller shapes on the list with every specific maximum. For example, with 3 as the maximum shape, the collection consisted of the circle and the ellipse shapes 1.5, 2 and 3. For each shape, we used either 3 times as many angles as the shape ratio (Table II) or 6 times as many (Table III). For the former analyses, the most likely clusters detected are shown in Figures 1(b)–(i). For comparison, Figure 1(a) shows the results with the circular scan statistic [15].

We evaluated the sensitivity of the results to the number of ellipse angles chosen. The ellipse angle is defined as the angle between the horizontal east–west line and the longest axis of the ellipse. With one angle, the analysis will only consider horizontal ellipses where the longest axis is in the east–west direction. With two angles, ellipses with a north–south axis are also considered, and so on. For a fixed collection of ellipse shapes set to 1 (circle),

Table II. Analysis of county based breast cancer mortality in the Northeastern United States, 1988–1992, using different maximum ellipse shape ratios. For each analysis, the collection of shapes used consists of all the lower ratios on the list up to and including the maxima. For each ellipse shape, the number of angles considered were *three* times the ellipse shape ratio.

Analysis		Most likely cluster								
Max	Computing	Counties		Ellipse		# Cases				
shape	min:s	Centroid	#	Shape	Angle	Obs	Exp	RR	LLR	*P*
1	0:25	Monmouth, NJ	32	1	n/a	24 044	23 040	1.044	35.70	0.001
1.5	0:36	Ocean, NJ	31	1.5	45°	22 748	21 711	1.048	38.94	0.001
2	0:51	Monmouth, NJ	28	2	30°	22 403	21 350	1.049	40.47	0.001
3	1:10	Camden, NJ	16	3	60°	8342	7585	1.100	42.16	0.001
4	1:40	Atlantic, NJ	33	4	30°	13 181	12 207	1.080	48.05	0.001
5	2:13	Talbot, MD	34	5	36°	13 206	12 231	1.080	48.05	0.001
6	2:52	Camden, NJ	24	6	30°	12 408	11 396	1.089	54.56	0.001
8	3:57	Camden, NJ	24	6	30°	12 408	11 396	1.089	54.56	0.001
10	5:02	Camden, NJ	24	6	30°	12 408	11 396	1.089	54.56	0.001
12	6:19	Camden, NJ	24	6	30°	12 408	11 396	1.089	54.56	0.001
15	7:59	Newcastle, DE	16	15	32°	10 347	9401	1.101	55.16	0.001
20	11:18	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001
30	13:54	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001
40	18:17	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001
60	25:39	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001
120	40:51	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001

Table III. Analysis of county based breast cancer mortality in the Northeastern United States, 1988–1992, using different maximum ellipse shape ratios. For each analysis, the collection of shapes used consists of all the lower ratios on the list up to and including the maxima. For each ellipse shape, the number of angles considered were *six* times the ellipse shape ratio.

Analysis		Most likely cluster								
Max	Computing	Counties		Ellipse		# Cases				
shape	min:s	Centroid	#	Shape	Angle	Obs	Exp	RR	LLR	*P*
1	0:25	Monmouth, NJ	32	1	n/a	24 044	23 040	1.044	35.70	0.001
1.5	0:48	Ocean, NJ	34	1.5	60°	24 482	23 410	1.046	40.54	0.001
2	1:13	Ocean, NJ	34	1.5	60°	24 482	23 410	1.046	40.54	0.001
3	1:52	Camden, NJ	16	3	60°	8342	7585	1.100	42.16	0.001
4	2:49	Atlantic, NJ	33	4	30°	13 181	12 207	1.080	48.05	0.001
5	3:53	Burlington, NJ	24	5	30°	12 633	11 612	1.088	54.76	0.001
6	5:58	Burlington, NJ	24	5	30°	12 633	11 612	1.088	54.76	0.001
8	8:23	Burlington, NJ	24	5	30°	12 633	11 612	1.088	54.76	0.001
10	9:29	Gloucester, NJ	18	10	33°	11 028	10 054	1.097	55.53	0.001
12	12:22	Gloucester, NJ	19	12	32.5°	11 133	10 155	1.096	55.53	0.001
15	16:06	Gloucester, NJ	19	12	32.5°	11 133	10 155	1.096	55.53	0.001
20	20:30	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001
30	27:10	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001
40	35:40	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001
60	49:44	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001
120	60:15	Monmouth, NJ	15	20	30°	9679	8731	1.109	58.68	0.001

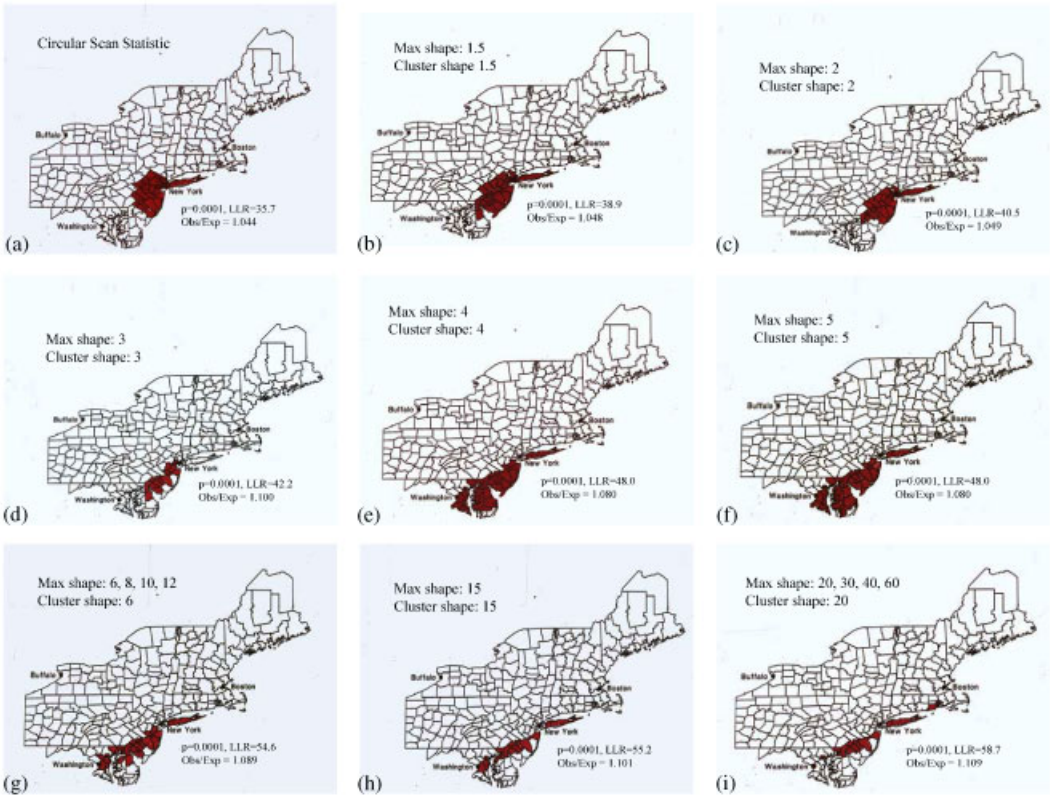

Figure 1. Breast cancer mortality in northeastern United States, 1988–1992, and the most likely clusters found using the circular (a) and elliptic (b)–(i) spatial scan statistic with different upper limits on the ellipse shape.

1.5, 2 and 3, and with the same number of angles for each of the ellipse shapes, the results for different number of angles are provided in Table IV.

4. ECCENTRICITY PENALTY

In Tables II and III, the shape of the most likely cluster is the maximum shape allowed in 15 of the 30 elliptic analyses. One possible reason for this is that the true cluster is very eccentric (long and narrow). Even under the null hypotheses though, the most likely cluster is more likely to be eccentric. To see this, note that with a less eccentric shape such as 1:1.1, all ellipses with the same centroid will irrespectively of their angle include approximately the same counties, with only minor variations at the edges. On the other hand, for eccentric ellipses with the same centroid, the counties included will be very different even if the angle only differs a little. In effect, an eccentric ellipse is more apt at 'picking cherries', setting the

Table IV. Most likely clusters for different number of angles used. The potential cluster shapes were set to 1 (circle), 1.5, 2 and 3. For each analysis, the number of angles were set to either 1, 2, 4, 6, ..., 90 or 180, with the same number of angles used for each shape. The number of degrees between neighbouring angles are given in the second column.

Analysis			Most likely cluster								
#		Computing	Counties		Ellipse		# Cases				
Angles	Closeness	min:s	Centroid	#	Shape	Angle	Obs	Exp	RR	LLR	*P*
1	n/a	0:33	Monmouth, NJ	32	1	n/a	24 044	23 040	1.044	35.70	0.001
2	90°	0:41	Monmouth, NJ	32	1	n/a	24 044	23 040	1.044	35.70	0.001
4	45°	0:54	Sussex, DE	48	3	45°	25 646	24 569	1.044	40.34	0.001
6	30°	1:08	Camden, NJ	16	3	60°	8342	7585	1.100	42.16	0.001
9	20°	1:26	Camden, NJ	16	3	60°	8342	7585	1.100	42.16	0.001
12	15°	1:49	Camden, NJ	16	3	60°	8342	7585	1.100	42.16	0.001
15	12°	2:08	Worcester, MD	35	3	36°	13 242	12 275	1.079	47.23	0.001
18	10°	2:27	Camden, NJ	16	3	60°	8342	7585	1.100	42.16	0.001
24	7.5°	3:01	Worcester, MD	33	3	37.5°	12 425	11 522	1.078	43.11	0.001
30	6°	3:44	Worcester, MD	35	3	36°	13 242	12 275	1.079	47.23	0.001
36	5°	4:18	Cape May, NJ	34	3	35°	13 110	12 138	1.080	48.06	0.001
45	4°	5:27	Worcester, MD	35	3	36°	13 242	12 275	1.079	47.23	0.001
60	3°	7:55	Worcester, MD	35	3	36°	13 242	12 275	1.079	47.23	0.001
90	2°	11:06	Worcester, MD	35	3	36°	13 242	12 275	1.079	47.23	0.001
180	1°	20:33	Cape May, NJ	34	3	35°	13 110	12 138	1.080	48.06	0.001

centroid, shape and angle so that only the counties with the highest rates are included even though they may not even be neighbouring counties. This can be seen in Figure 1(i).

One solution to this problem is to exclude the more eccentric ellipses from the analysis, as done in the previous section. Another possible solution is to adjust the cluster likelihood with a penalty function that discourages eccentric clusters without excluding their possibility. We decided to use the eccentricity penalty function $(4s/(s+1)^2)^a$ so that the adjusted log likelihood is

$$\mathrm{LLR}_{\mathrm{adj}} = \mathrm{LLR} * \left(\frac{4s}{(s+1)^2}\right)^a$$

where $\mathrm{LLR}_{\mathrm{adj}}$ is the adjusted log likelihood ratio, LLR is the original log likelihood ratio, and s is the cluster shape defined as the length of the longest axis divided by the length of the shortest axis of the ellipse and $a \geqslant 0$ is a tuning parameter. (*Note*: For $a = 1$, this is the inverse ratio of the area of the smallest rectangular box containing the ellipse with the area of a square with the same circumference as that box.) The eccentricity penalty is stronger for large values of the tuning parameter a. In the extreme, when $a = 0$, there is no penalty, and when $a \to \infty$, the penalty is so strong that only circular clusters are allowed. Note that (i) for the circle $s = 1$, so that the adjusted log likelihood ratio is equal to the true log likelihood ratio, (ii) that the amount of penalty increases in a monotone fashion as s increases, and (iii) that $\mathrm{LLR}_{\mathrm{adj}} \to 0$ as $s \to \infty$.

Using the above penalty function we reanalysed the breast cancer mortality data with the same set of maximum shapes. The results are presented in Table V. For shape 1.5, the most

Table V. Analysis of county based breast cancer mortality in the Northeastern United States, 1988–1992, using different maximum ellipse shape ratios, and the $4s/(s+1)^2$ penalty function, where s is the length of longest axis of the ellipse divided by the length of the shortest axis. For each ellipse shape, the number of angles considered was three times the ellipse shape. The most likely cluster was the same for shape maxima of 1.5, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 30, 40, 60 and 120.

	Most likely cluster									
Max	Counties		Ellipse		# Cases					
shape	Centroid	#	Shape	Angle	Obs	Exp	RR	LLR	LLR_{adj}	P
1	Monmouth, NJ	32	1	n/a	24 044	23 040	1.044	35.70	35.70	0.001
1.5–12	Ocean, NJ	31	1.5	45°	22 748	21 711	1.048	38.94	37.38	0.001

likely cluster had a log likelihood value of 38.94 (Table II), and the adjusted log likelihood is then $38.94 * (4 * 1.5/(1.5+1)^2) = 37.38$ which is higher than the maximum log likelihood for the circular shape, which was 35.70. For shape 2, the most likely cluster had a log likelihood value of 40.47 (Table II), so that the adjusted log likelihood is $40.47 * (4 * 2/(2+1)^2) = 35.97$. This is also larger than the maximum for the circles but not larger than the maximum among the 1.5 shape ellipses. In fact, no other shape has an adjusted log likelihood higher than for 1.5, so that is the shape of the eccentricity adjusted most likely cluster.

5. ORAL CANCER MORTALITY IN THE UNITED STATES

We applied the circular and elliptic spatial scan statistic to the classical oral cancer mortality data for white females in the United States, 1950–1969, collected by the National Center for Health Statistics. In 1975, these data were shown to have high rates in southeastern United States [17]. This led to a subsequent case-control study in North Carolina identifying snuff dipping (smokeless tobacco use) as a primary risk factor for oral cancer, with relative risks ranging as high as 50 for cancers of the oral tissue that comes in direct contact with the tobacco [18].

Plate 1 shows the original age-adjusted mortality data remapped to quintiles of the rate distribution at the county level, along with the most likely circular and elliptical clusters. We analysed these data using circles and ellipses based on county centroid co-ordinates with a maximum cluster size of 50 per cent of total population, ellipse shapes of 1.5, 2, 3, and 4 and using the eccentricity penalty. The optimum ellipse has an axis ratio of 1.5 and is centred in the southwestern corner of Georgia whereas the circle is centred about 100 miles to the east in Valdosta, GA. It is interesting to note that the circular cluster includes eastern North Carolina where the original etiological study was conducted, but the elliptical cluster more closely follows the concentration of high rates from the Mississippi River to South Carolina.

6. POWER EVALUATION

In this section, the power of the elliptic spatial scan statistic is estimated for different alternative hypothesis, where the true cluster is either circular or elliptic in shape. Comparisons

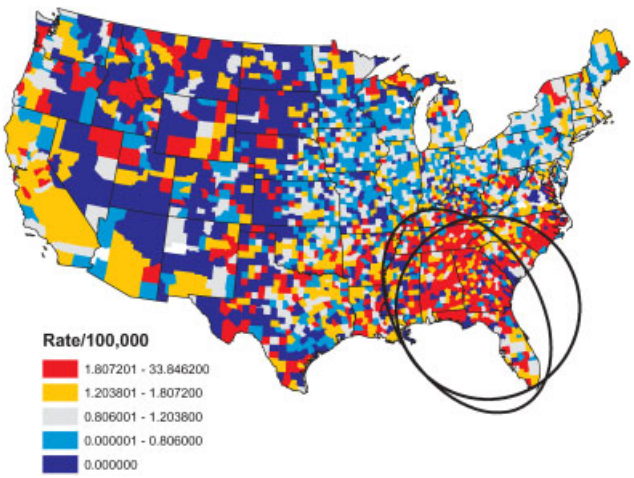

Plate 1. Oral cancer mortality rates in women 1950–1969, by county. Shown are also the locations of the most likely cluster for the circular and elliptic spatial scan statistics, the latter being used with the eccentricity penalty.

Table VI. Power of the circular and elliptic spatial scan statistics for true circular clusters with 1, 2, 4, 8 or 16 counties, at four different locations and for an $\alpha = 0.05$ significance level. Different maximum elliptic shapes were used, set at 2, 4, 8 and 20, respectively.

True cluster		Type of scan statistic (number = max shape)								
(circular)		Circular	Elliptic, without penalty				Elliptic, with penalty			
# counties		1	2	4	8	20	2	4	8	20
Rural	1	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00
(Grand Isle, VT)	2	0.99	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99
	4	0.97	0.96	0.96	0.96	0.95	0.97	0.97	0.97	0.97
	8	0.97	0.97	0.96	0.96	0.95	0.97	0.97	0.97	0.97
	16	0.97	0.97	0.96	0.96	0.95	0.97	0.97	0.97	0.97
Mixed	1	0.94	0.93	0.91	0.92	0.92	0.92	0.92	0.92	0.92
(Allegheny, PA)	2	0.94	0.92	0.91	0.92	0.92	0.92	0.92	0.92	0.92
	4	0.94	0.93	0.92	0.92	0.92	0.93	0.93	0.93	0.93
	8	0.94	0.93	0.92	0.92	0.91	0.93	0.93	0.93	0.93
	16	0.95	0.94	0.94	0.93	0.92	0.94	0.94	0.94	0.94
Mixed	1	0.99	0.99	0.98	0.98	0.98	0.99	0.99	0.99	0.99
(Delaware, NY)	2	0.98	0.97	0.97	0.97	0.96	0.98	0.98	0.98	0.98
	4	0.97	0.97	0.96	0.96	0.94	0.97	0.97	0.97	0.97
	8	0.96	0.95	0.94	0.93	0.91	0.95	0.95	0.95	0.95
	16	0.95	0.95	0.94	0.92	0.91	0.95	0.95	0.95	0.95
Urban	1	0.92	0.88	0.86	0.87	0.89	0.90	0.90	0.90	0.90
(Manhattan, NY)	2	0.90	0.88	0.86	0.80	0.88	0.88	0.88	0.88	0.88
	4	0.89	0.87	0.86	0.86	0.87	0.88	0.87	0.87	0.87
	8	0.91	0.90	0.89	0.89	0.88	0.91	0.91	0.91	0.91
	16	0.93	0.92	0.91	0.91	0.90	0.92	0.92	0.92	0.92

are made with the power of the circular scan statistic. When the true cluster is elliptic, the question is how much we will gain by using the elliptic scan statistic. When the true cluster shape is circular, the hope is that we do not loose too much power when using the elliptic spatial scan statistic as opposed to the circular one.

For all subsequent power estimates, 99 999 random data sets were generated under the null hypothesis with 600 cases in each, where each case is assigned to a county with probability in proportion to its population. Among the maximum likelihood ratios for each of these data sets, the one ranked as number 5000, starting with the highest, is the critical value needed to reject the null hypothesis at the $\alpha = 0.05$ significance level. We then generated 10 000 random data sets under each of the alternative hypotheses, and the estimated power is the per cent of these 10 000 data sets that has a maximum likelihood ratio larger than the previously described critical value.

6.1. Circular cluster alternatives

Kulldorff *et al.* [3] have constructed a set of benchmark data sets for evaluating the power of spatial clustering tests. Three of these are hot spot clusters consisting of a single county

located in a rural area (Grand Isle, VT), a mixed urban/rural area (Allegheny, PA) and an urban area (Manhattan, NY). Centred on the same three counties, but also including the closest surrounding counties, they constructed additional hot-spot clusters with 2, 4, 8 and 16 counties, respectively. The relative risk was the same in all counties within the same cluster, and defined so that one would have 99.9 per cent power to detect the cluster if one knew the exact geographical extent *a priori*. Hence, this is an upper limit on the power that any test statistic could obtain.

Kulldorff *et al.* [3] calculated the power of the circular scan statistic for each of these 15 alternative models. Those numbers are reproduced in Table VI, together with the power estimates for the elliptic scan statistics with different upper limits on the ellipse shape, and with and without the eccentricity penalty. Using the same procedure, we also created an additional set of five hot-spot clusters in a mixed urban/rural area centred on Delaware County, NY. As expected, the circular scan statistic has the highest power for circular alternatives. It is worth noting though, that the loss in power for the elliptic scan statistics is modest, and especially so when using the eccentricity penalty.

6.2. *Elliptic cluster alternatives*

For the elliptic cluster alternative models, we used the same four central counties as for the circular alternative hypotheses. Instead of picking the nearest neighbours to add to the cluster we drew an ellipse around the county centroid with the longest axis at 45° (southwest to northeast). We then included additional counties into the cluster in order of their inclusion into the ellipse as the size of the ellipse was increased. We did this for three different ellipse shapes: 2, 4 and 8. The twelve clusters with 16 counties are depicted in Figure 2.

With four centroids, four cluster sizes and four cluster shapes, we evaluated a total of 64 different elliptic alternative hypotheses. The results are presented in Tables VII–IX for true cluster shapes of 2, 4 and 8, respectively. As expected, the power is often highest for the elliptic scan statistic with maximum shape close to the true shape. The loss in power when specifying a large range of shapes is rather modest though. Likewise, the circular scan statistic is still competitive for elliptic clusters. Results were similar for true cluster shapes of 1.5 and 3 (data not shown).

7. DISCUSSION

The elliptic spatial scan statistic performs well when the maximum shape is not too large. For very large shape maxima the elliptic spatial scan statistic performs poorly, as it may select a narrow string of noncontiguous census areas, with many areas in between left outside of the cluster (Figure 1(I)). This may be less of a problem with data that is less aggregated, but even then there is a question of why one would expect a true cluster to follow a very narrow but straight line rather than a slightly curved one. Hence, we recommend that if the elliptic spatial scan statistic is used without an eccentricity penalty, there should be a fairly restrictive upper limit on the shape ratio. This recommendation also makes sense in terms of computing time. While more computer intensive than the circular scan statistic, it is not overly problematic when the maximum shape ratio is low. Long and narrow ellipses need a much higher number of angles for each shape though, leading to considerable increase in

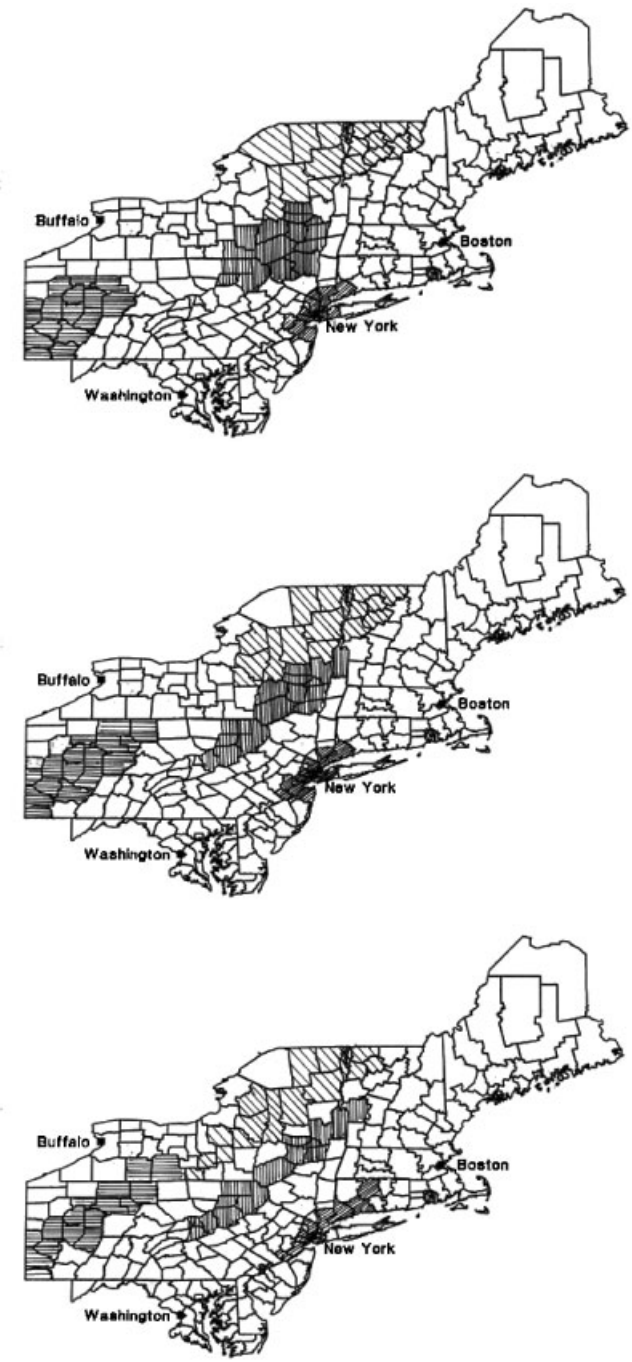

Figure 2. The 16 county elliptic clusters used for the power evaluations, with a shape of two (top), four (middle) and eight (bottom), respectively.

Table VII. Power of the circular and elliptic spatial scan statistics for true elliptic clusters with shape 2, with 1, 2, 4, 8 or 16 counties, at four different locations and for an $\alpha = 0.05$ significance level. Different maximum elliptic shapes were used, set at 2, 4, 8 and 20, respectively.

True cluster		Type of scan statistic (number = max shape)								
(shape = 2)		Circular	Elliptic, without penalty				Elliptic, with penalty			
# counties		1	2	4	8	20	2	4	8	20
Rural	2	0.99	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99
(Grand Isle, VT)	4	0.97	0.97	0.96	0.96	0.95	0.97	0.97	0.97	0.97
	8	0.97	0.97	0.96	0.96	0.95	0.97	0.97	0.97	0.97
	16	0.95	0.95	0.95	0.94	0.93	0.95	0.95	0.95	0.95
Mixed	2	0.94	0.92	0.91	0.91	0.92	0.92	0.92	0.92	0.92
(Allegheny, PA)	4	0.93	0.92	0.92	0.92	0.91	0.92	0.92	0.92	0.92
	8	0.94	0.93	0.93	0.92	0.92	0.92	0.93	0.93	0.93
	16	0.94	0.94	0.93	0.93	0.92	0.94	0.94	0.94	0.94
Mixed	2	0.96	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98
(Delaware, NY)	4	0.95	0.96	0.96	0.96	0.94	0.97	0.97	0.97	0.97
	8	0.85	0.92	0.92	0.91	0.91	0.92	0.92	0.92	0.92
	16	0.93	0.94	0.94	0.93	0.91	0.94	0.94	0.94	0.94
Urban	2	0.91	0.89	0.87	0.87	0.89	0.89	0.89	0.89	0.89
(Manhattan, NY)	4	0.89	0.87	0.86	0.86	0.87	0.88	0.88	0.88	0.88
	8	0.91	0.90	0.89	0.89	0.89	0.90	0.90	0.90	0.90
	16	0.91	0.91	0.91	0.91	0.89	0.91	0.91	0.91	0.91

Table VIII. Power of the circular and elliptic spatial scan statistics for true elliptic clusters with shape 4, with 1, 2, 4, 8 or 16 counties, at four different locations and for an $\alpha = 0.05$ significance level. Different maximum elliptic shapes were used, set at 2, 4, 8 and 20, respectively.

True cluster		Type of scan statistic (number = max shape)								
(shape = 4)		Circular	Elliptic, without penalty				Elliptic, with penalty			
# counties		1	2	4	8	20	2	4	8	20
Rural	2	0.99	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99
(Grand Isle, VT)	4	0.96	0.97	0.96	0.96	0.96	0.97	0.97	0.97	0.97
	8	0.94	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95
	16	0.91	0.93	0.94	0.94	0.93	0.92	0.93	0.93	0.93
Mixed	2	0.94	0.92	0.91	0.91	0.92	0.92	0.92	0.92	0.92
(Allegheny, PA)	4	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	8	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	16	0.92	0.92	0.92	0.92	0.91	0.92	0.92	0.92	0.92
Mixed	2	0.96	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98
(Delaware, NY)	4	0.87	0.92	0.94	0.94	0.95	0.91	0.92	0.92	0.92
	8	0.86	0.91	0.94	0.93	0.92	0.90	0.91	0.91	0.91
	16	0.86	0.88	0.92	0.93	0.92	0.88	0.89	0.89	0.89
Urban	2	0.91	0.87	0.86	0.86	0.88	0.89	0.89	0.89	0.89
(Manhattan, NY)	4	0.87	0.87	0.86	0.86	0.88	0.88	0.88	0.88	0.88
	8	0.89	0.88	0.88	0.88	0.88	0.92	0.91	0.91	0.91
	16	0.91	0.91	0.91	0.91	0.90	0.91	0.91	0.91	0.91

Table IX. Power of the circular and elliptic spatial scan statistics for true elliptic clusters with shape 8, with 1, 2, 4, 8 or 16 counties, at four different locations and for an $\alpha = 0.05$ significance level. Different maximum elliptic shapes were used, set at 2, 4, 8 and 20, respectively.

True cluster		Type of scan statistic (number = max shape)								
(shape = 8)		Circular	Elliptic, without penalty				Elliptic, with penalty			
# counties		1	2	4	8	20	2	4	8	20
Rural	2	0.99	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99
(Grand Isle, VT)	4	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.97	0.97
	8	0.90	0.91	0.92	0.94	0.93	0.90	0.91	0.91	0.91
	16	0.87	0.89	0.91	0.93	0.92	0.88	0.89	0.89	0.89
Mixed	2	0.94	0.92	0.91	0.91	0.92	0.92	0.92	0.92	0.92
(Allegheny, PA)	4	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	8	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	16	0.89	0.90	0.91	0.91	0.91	0.90	0.90	0.90	0.90
Mixed	2	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(Delaware, NY)	4	0.90	0.93	0.94	0.95	0.95	0.92	0.93	0.93	0.93
	8	0.84	0.85	0.89	0.93	0.94	0.86	0.87	0.87	0.87
	16	0.82	0.85	0.90	0.92	0.92	0.84	0.85	0.85	0.85
Urban	2	0.91	0.87	0.86	0.86	0.88	0.89	0.89	0.89	0.89
(Manhattan, NY)	4	0.83	0.79	0.82	0.84	0.85	0.81	0.81	0.81	0.81
	8	0.69	0.72	0.78	0.84	0.85	0.71	0.73	0.73	0.73
	16	0.78	0.81	0.84	0.86	0.86	0.80	0.80	0.80	0.80

computing time as the maximum shape ratio increases. This can be seen from the analyses in Tables II and III.

An alternative is to use a large number of elliptic shapes combined with the eccentricity penalty, so that eccentric clusters will emerge only if there is strong evidence compared to less eccentric clusters, but where the more compact cluster is favoured if the likelihoods are about the same. By incorporating the eccentricity penalty, it is less likely that eccentric clusters are detected when the true cluster is more compact, at the same time as eccentric clusters can be detected if the evidence is sufficiently strong. In most practical situations, we think that this is a reasonable approach to take and better than using the elliptic scan statistic without an eccentricity penalty.

In this paper we only evaluated one particular choice of eccentricity penalty function. By changing the tuning parameter, it is possible to increase or decrease the amount of penalty. While to some extent arbitrary, the choice of this parameter should as much as possible be based on prior views on the type of clusters that are likely to occur or of interest to find in the data set being analysed. It is reassuring to know though that the statistical power is rather robust to this choice. There are also alternative penalty functions that one could use, some of which may be more appropriate or more natural for certain types of applications.

The number of angles to use for each ellipse shape is an arbitrary decision, but it is clear from Table II that a narrower ellipse needs more angles. We tried to set the number of angles to both three and six times the shape ratio, with very similar results. From Table IV it is

clear that we should not use too few though. We recommend using at least three times as many angles as the shape ratio.

Although more flexible than the circle, the elliptic scan statistic still imposes a parametric shape on the potential clusters. This may make it difficult to detect clusters of very irregular shapes, such as a narrow strip on each side of a long and winding river or along the shore of a rough edged coastline. To detect such clusters, it is probably better to use one of the non-parametric spatial scan statistics proposed by Duczmal and Assunção [19], Patil and Taillie [20] or Tango and Takahashi [21].

A major conclusion from this paper is that in terms of power, the elliptic scan statistic performs well for circular clusters, and equally important, that the circular scan statistic performs well for elliptic clusters. One possible advantage of the elliptic *versus* the circular scan statistic is that the former may give a better estimate of the true cluster area. When the true cluster is an elongated one, the geographic area determined by the most likely cluster to be included in follow-up investigations may then be more specifically defined by the elliptic method than the circular scan statistic. Given the scarce resources available to most state and local health departments, a greater specificity of cluster identification would reduce the cost to investigate potential disease outbreaks. It could also be the opposite though. Even if the true cluster is circular, some areas on its border will by chance have fewer cases and some areas just outside the border will by chance have more cases. This may lead to the most likely cluster being elliptic, even when the true cluster is circular.

Whatever the shape of the most likely cluster, it is important to keep in mind that it only indicates the general area of the true underlying cluster, and that the exact borders of the detected clusters are uncertain. This is sufficient for most practical purposes, as the method's main purpose is to generate a signal with a general idea of where the outbreak has occurred. More detailed information about the outbreak, its cause, nature and extent, can only be obtained through detailed epidemiological investigations by public health officials, who should not focus exclusively on the area within the most likely cluster, but also on neighbouring localities. In light of this last comment, the exact choice of shapes and angles is not of critical importance. The key is to use a wide variety of centroid co-ordinates and cluster sizes.

If computing resources allow, better results may sometimes be obtained using an elliptic rather than circular scan statistic. It is reassuring though that the difference in power is marginal, and the elliptic scan statistic will perform only slightly better or worse depending on the shape of the true underlying cluster. Either option fulfills the basic purpose of geographical cluster detection. For valid statistical inference, it is important that the choice is made *a priori* though, before analysing the data, in order to avoid pre-selection bias.

ACKNOWLEDGEMENTS

This work was partially funded by the National Cancer Institute. We thank Don Green, Information Management Services, Inc., Silver Spring, MD, for computer programming support. Valuable comments from three anonymous reviewers are gratefully acknowledged. In particular, we thank one of them for suggesting the addition of a tuning parameter to the eccentricity penalty function.

REFERENCES

1. Glaz J, Naus JI, Wallenstein S. *Scan Statistics*. Springer: New York, 2001.
2. Glaz J, Balakrishnan N. *Scan Statistics and Applications*. Birkhäuser: Boston, 1999.

3. Kulldorff M, Tango T, Park P. Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis* 2003; **42**:665–684.
4. Song C, Kulldorff M. Power evaluation of disease clustering tests. *International Journal of Health Geographics* 2003; **2**:9.
5. Sabel CE, Boyle PJ, Löytönen M, Gatrell AC, Jokelainen M, Flowerdew R, Maasilta P. Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death. *American Journal of Epidemiology* 2003; **157**:898–905.
6. Hsu CE, Jacobson HE, Soto Mas F. Evaluating the disparity of female breast cancer mortality among racial groups—a spatiotemporal analysis. *International Journal of Health Geographics* 2004; **3**:4.
7. Thomas AJ, Carlin BP. Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. *Statistics in Medicine* 2003; **22**:113–127.
8. Boyle E, Johnson H, Kelly A, McDonnell R. Congenital anomalies and proximity to landfill sites. *Irish Medical Journal* 2004; **97**:16–18.
9. Odoi A, Martin SW, Michel P, Middleton D, Holt J, Wilson J. Investigation of clusters of giardiasis using GIS and a spatial statistic. *International Journal of Health Geographics* 2004; **3**:11.
10. Sauders BD, Fortes ED, Morse DL, Dumas N, Kiehlbauch JA, Schukken Y, Hibbs JR, Wiedmann M. Molecular subtyping to detect human listeriosis clusters. *Emerging Infectious Diseases* 2003; **9**:672–680.
11. Joly DO, Ribic CA, Langenberg JA, Beheler K, Batha CA, Dhuey BJ, Rolley RE, Bartelt G, Van Deelen TR, Samual MD. Chronic wasting disease in free-ranging Wisconsin white-tailed deer. *Emerging Infectious Diseases* 2003; **9**:599–601.
12. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice: the New York City emergency department system. *Emerging Infectious Diseases* 2004; **10**:858–864.
13. Mostashari F, Kulldorff M, Hartman JJ, Miller JR, Kulasekera V. Dead bird clustering: a potential early warning system for West Nile virus activity. *Emerging Infectious Diseases* 2003; **9**:641–646.
14. Kulldorff M. A spatial scan statistic. *Communications in Statistics*: *Theory and Methods* 1997; **26**:1481–1496.
15. Kulldorff M, Feuer E, Miller B, Freedman L. Breast cancer in northeast United States: a geographic analysis. *American Journal of Epidemiology* 1997; **146**:161–170.
16. Kulldorff M. An isotonic spatial scan statistic for geographical disease surveillance. *Journal of the National Institute of Public Health* 1999; **48**:94–101.
17. Mason TJ, McKay FW, Hoover R, Blot WJ, Fraumeni JF. *Atlas of Cancer Mortality for U.S. Counties*: *1950–1969*. Washington, D.C., USGPO (DHEW Publ. No. (NIH) 75-780), 1975.
18. Winn DM, Blot WJ, Shy CM, Pickle LW, Toledo A, Fraumeni Jr JF. Snuff dipping and oral cancer among women in the southern United States. *New England Journal of Medicine* 1981; **304**:745–749.
19. Duczmal L, Assunçião RM. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* 2004; **45**:269–286.
20. Patil GP, Taillie C. Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science* 2003; **18**:457–465.
21. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 2005; **4**:11.

