

Table 7: Sensitivity and positive predictive value (PPV) of the flexible and cylindrical space-time scan statistics.

		traditional power	zip codes sensitivity (%)	PPV (%)	population sensitivity (%)	PPV (%)
Cluster A; $s^* = 1$; high risk						
Day 31	flexible (K = 20)					
	cylindrical (K = 20)	0.860	85.30	89.45	85.30	91.50
	cylindrical (50% pop)	0.862	85.90	88.79	85.90	90.84
Day 32	flexible (K = 20)	0.988	98.80	84.50	98.80	87.80
	cylindrical (K = 20)	0.996	99.50	97.44	99.50	98.18
	cylindrical (50% pop)	0.996	99.50	97.33	99.50	98.08
Day 33	flexible (K = 20)	0.999	99.90	96.27	99.90	97.32
	cylindrical (K = 20)	0.999	99.90	99.48	99.90	99.65
	cylindrical (50% pop)	0.999	99.90	99.48	99.90	99.65
Cluster A5; $s^* = 5$; high risk						
Day 31	flexible (K = 20)	0.797	66.08	55.93	67.29	63.00
	cylindrical (K = 20)	0.850	69.62	80.35	71.65	84.21
	cylindrical (50% pop)	0.847	70.62	78.17	71.62	82.02
Day 32	flexible (K = 20)	0.994	92.22	70.17	92.94	76.73
	cylindrical (K = 20)	0.996	88.78	85.14	90.86	89.41
	cylindrical (50% pop)	0.996	88.96	84.81	91.05	89.11
Day 33	flexible (K = 20)	1.000	95.88	80.02	96.64	85.25
	cylindrical (K = 20)	1.000	91.42	87.32	93.66	91.67
	cylindrical (50% pop)	1.000	91.42	87.30	93.66	91.65
The Rockaways; $s^* = 5$; high risk						
Day 31	flexible (K = 20)	0.769	60.68	72.09	69.04	73.58
	cylindrical (K = 20)	0.855	62.32	91.63	74.45	91.76
	cylindrical (50% pop)	0.840	61.40	91.04	73.65	91.15
Day 32	flexible (K = 20)	0.992	86.76	87.36	94.17	89.86
	cylindrical (K = 20)	0.997	77.00	96.84	92.75	97.46
	cylindrical (50% pop)	0.997	77.00	96.84	92.75	97.46
Day 33	flexible (K = 20)	1.000	92.16	93.81	97.15	95.97
	cylindrical (K = 20)	1.000	78.50	98.06	94.51	98.59
	cylindrical (50% pop)	1.000	78.50	98.06	94.51	98.59
Hudson River; $s^* = 20$; high risk						
Day 31	flexible (K = 20)	0.656	20.07	64.99	26.00	69.72
	cylindrical (K = 20)	0.597	14.23	61.10	18.16	65.18
	cylindrical (50% pop)	0.632	26.15	50.70	31.26	53.71
Day 32	flexible (K = 20)	0.964	32.17	73.59	41.81	78.36
	cylindrical (K = 20)	0.933	24.13	61.55	31.58	66.69
	cylindrical (50% pop)	0.949	42.90	50.27	51.50	53.96
Day 33	flexible (K = 20)	0.998	34.91	79.39	46.27	84.17
	cylindrical (K = 20)	0.994	27.23	60.56	36.75	66.20
	cylindrical (50% pop)	0.995	48.14	47.34	58.54	51.34

with 999 Monte Carlo replications, the flexible space-time scan statistic took 87.7 minutes to run on a 3.06-GHz Pentium 4 computer, while the cylindrical space-time scan statistic took only 9.8 minutes.

A limitation of length may also prevent the analysis to present large clusters of unlikely and very peculiar shapes. These undesirable properties produced by maximum likelihood ratio might suggest the use of different criterion for model selection, including some penalized likelihood [20,29]. Also, for larger cluster seizures, the method is not

practically feasible and a more efficient algorithm is needed.

In this paper, we considered the *right* cylinder or *right* prism of the cluster model, as an expansion of the cylindrical space-time scan statistic for a prospective disease surveillance by Kulldorff [10]. This does not allow the scanning window to adjust itself as the disease outbreak grows or shrinks geographically over time. Recently, Iyengar has suggested using a *square pyramid shape* window

which can model either growth (or shrinkage) and movement of the disease cluster [30]. For the proposed flexible space-time scan statistic, if we could consider the flexibility in both space and time, that is, evaluating all connected subsets within a cylinder instead of \mathcal{W} in (4), we can detect more arbitrarily shaped clusters in space-time. For such an expansion, an efficient computational algorithm will be needed for the scanning process, as well as a more sophisticated mechanism for the interpretation of such complicatedly shaped clusters. The implementation and importance of such methods for disease surveillance and monitoring is an issue for future research.

Authors' contributions

KT, MK and TT developed the statistical methodology and designed the study. KT, MK and KY analyzed and interpreted the syndromic surveillance data. KT programmed the methods, did the power calculations and wrote the first draft of the manuscript. All authors participated in the interpretation of the results, revised the manuscript, and approved the final version.

Acknowledgements

The authors thank Allyson Abrams for comments concerning the syndromic surveillance data from Massachusetts, and Dr. Tetsuji Yokoyama for advice about C++ programming.

This research was partly funded by a Modeling Infectious Disease Agent Study (MIDAS) grant (No. U01GM076672) from the National Institute of General Medical Science, National Institutes of Health, USA, and a scientific grant (No. H16-Kenkou-039) from the Ministry of Health, Labour and Welfare, Japan.

References

- Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D: **Syndromic surveillance in public health practice**, New York City. *Emerging Infectious Diseases* 2004, **10**:858-864.
- Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W, Sari J, Sniegowski C, Wojcik R, Pavlin J: **A systems overview of the electronic surveillance system for the early notification of community-based epidemics (ESSENCE II)**. *Journal of Urban Health* 2003, **80**(2 suppl.1):i32-i42.
- Lazarus R, Kleinman K, Dashevsky I, Adams C, Kludt P, DeMaria A, Platt R: **Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events**. *Emerg Infect Dis* 2002, **8**(8):753-760.
- Platt R, Bocchino C, Caldwell B, Harmon R, Kleinman K, Lazarus R, Nelson AF, Nordin JD, Ritzwoller P: **Syndromic surveillance using minimum transfer of identifiable data: the example of the National Bioterrorism Syndromic Surveillance Demonstration Program**. *Journal of Urban Health* 2003, **80**(2 suppl.1):i25-i31.
- Sonesson C, Bock D: **A review and discussion of prospective statistical surveillance in public health**. *Journal of the Royal Statistical Society, Series A* 2003, **166**:5-21.
- Lawson AB, Kleinman K, Eds: *Spatial & Syndromic Surveillance for Public Health* Chichester: Wiley; 2005.
- Naus J, Wallenstein S: **Temporal surveillance using scan statistics**. *Statistics in Medicine* 2006, **25**:311-324.
- Kulldorff M: **A spatial scan statistic**. *Communications in Statistics - Theory and Methods* 1997, **26**:1481-1496.
- Rogerson PA, Yamada I: **Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches**. *Statistics in Medicine* 2004, **23**:2195-2214.
- Kulldorff M: **Prospective time periodic geographical disease surveillance using a scan statistic**. *Journal of the Royal Statistical Society, Series A* 2001, **164**:61-72.
- Kulldorff M, Heffernan R, Hartman J, Assunção R, Mostashari F: **A space-time permutation scan statistic for disease outbreak detection**. *PLoS Medicine* 2005, **2**(3):e59.
- Lawson AB, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R, Eds: *Disease Mapping and Risk Assessment for Public Health* New York: Wiley; 1999.
- Lawson AB: *Statistical Methods in Spatial Epidemiology* 2nd edition. Chichester: Wiley; 2006.
- Kulldorff M, Information Management Services, Inc: **SaTScan version 7.0: software for the spatial and space-time scan statistics**. 2007 [<http://www.satscan.org/>].
- Kulldorff M: **Scan statistics for geographical disease surveillance: an overview**. In *Spatial & Syndromic Surveillance for Public Health* 2nd edition. Edited by: Lawson AB, Kleinman K. Chichester: Wiley; 2005:115-131.
- Tango T, Takahashi K: **A flexibly shaped spatial scan statistic for detecting clusters**. *International Journal of Health Geographics* 2005, **4**(11):.
- Duczmal L, Assunção R: **A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters**. *Computational Statistics & Data Analysis* 2004, **45**:269-286.
- Patil GP, Taillie C: **Upper level set scan statistic for detecting arbitrarily shaped hotspots**. *Environmental and Ecological Statistics* 2004, **11**:183-197.
- Assunção R, Costa M, Tavares A, Ferreira S: **Fast detection of arbitrarily shaped disease clusters**. *Statistics in Medicine* 2006, **25**:723-742.
- Kulldorff M, Huang L, Pickle L, Duczmal L: **An elliptic spatial scan statistic**. *Statistics in Medicine* 2006, **25**:3929-3943.
- Takahashi K, Yokoyama T, Tango T: **FleXScan version 2.0: Software for the Flexible Spatial Scan Statistic**. *Japan* 2007 [http://www.niph.go.jp/soshiki/gijyutsu/index_e.html].
- Kulldorff M, Zhang Z, Hartman J, Heffernan R, Huang L, Mostashari F: **Benchmark data and power calculations for evaluating disease outbreak detection methods**. *Morbidity and Mortality Weekly Report* 2004, **53**(Supplement 1):144-151.
- Dwass M: **Modified randomization tests for nonparametric hypotheses**. *The Annals of Mathematical Statistics* 1957, **28**:181-187.
- Lazarus R, Kleinman K, Dashevsky I, DeMaria A, Platt R: **Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection**. *BMC Public Health* 2001, **1**:9.
- Kleinman K, Lazarus R, Platt R: **A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism**. *American Journal of Epidemiology* 2004, **159**:217-224.
- Takahashi K, Tango T: **An extended power of cluster detection tests**. *Statistics in Medicine* 2006, **25**:841-852.
- Forsberg L, Bonetti M, Jeffery C, Ozonoff A, Pagano M: **Distance-based methods for spatial and spatio-temporal surveillance**. In *Spatial & Syndromic Surveillance for Public Health* 2nd edition. Edited by: Lawson AB, Kleinman K. Chichester: Wiley; 2005:115-131.
- Huang L, Kulldorff M, Gregorio D: **A spatial scan statistic for survival data**. *Biometrics* 2007, **63**:109-118.
- Duczmal L, Kulldorff M, Huang L: **Evaluation of spatial scan statistics for irregularly shaped clusters**. *Journal of Computational and Graphical Statistics* 2006, **15**(2):428-442.
- Iyengar VS: **Space-time clusters with flexible shapes**. *Morbidity and Mortality Weekly Report* 2005, **54**(Supplement):71-76.

Original Article

A Spatial Scan Statistic with a Restricted Likelihood Ratio

Toshiro Tango

Department of Technology Assessment and Biostatistics
National Institute of Public Health,
3-6 Minami 2 chome, Wako, Saitama, 351-0197, Japan
e-mail:tango@niph.go.jp

Kulldorff (1997) developed a circular spatial scan statistic for identifying the most likely cluster of disease that maximizes the likelihood ratio and his software SaTScan has been widely used for geographical disease cluster detection and disease surveillance. To detect non-circular clusters which cannot be detected by Kulldorff's circular spatial scan statistic, several non-circular spatial scan statistics have been proposed. However, it does not seem to be well recognized that these spatial scan statistics tend to detect the most likely cluster much larger than the true cluster by swallowing neighbouring regions with non-elevated risk. This paper proposes a new spatial scan statistic free from such an undesirable property by modifying the likelihood ratio so that it scans only the regions with elevated risk. Monte Carlo Simulation study shows that the proposed circular spatial scan statistic is shown to have better ability to identify the true cluster compared with Kulldorff's one in all the cluster models considered. The proposed circular spatial scan statistic is illustrated with mortality data from cerebrovascular disease in Tokyo Metropolitan area, Japan.

Key words: Cluster detection; hot-spot cluster; likelihood ratio test; Monte Carlo testing; spatial epidemiology.

1. Introduction

Many different test statistics have been proposed for detecting disease clustering (Lawson *et al.*, 1999; Waller and Gotway, 2004; Kulldorff, 2006). Especially, the spatial scan statistic proposed by Kulldorff and Nagarwalla (1995) and Kulldorff (1997) 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.*, 2006a). The spatial scan statistic tries to identify the *most likely cluster* (MLC) defined as the set of connected regions that attains the maximum likelihood ratio. However, since it uses a circular window to scan the potential cluster areas, it has difficulty in correctly detecting actual non-circular clusters. To detect arbitrarily shaped clusters which cannot be detected by the circular spatial scan statistic, Duczmal and

Assunção (2004), Patil and Taillie (2004), Tango and Takahashi (2005) and Assunção *et al.* (2006) have proposed different spatial scan statistics. It should be noted that all of these scan statistics are based on maximizing the likelihood ratio.

Tango (2000) showed an interesting example that Kulldorff's circular spatial scan statistic detected an unrealistically large MLC consisting of 70 regions, much larger than expected from an observed disease map, by absorbing neighbouring regions with non-elevated risk of disease occurrence in his simulated data. Furthermore, Tango and Takahashi (2005) have shown examples in which Duczmal and Assunção's procedure detected 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, Duczmal and Assunção's and Tango and Takahashi's. These results casted a doubt on the validity of the model selection based on maximizing the likelihood ratio.

In this paper, we shall propose a new spatial scan statistic free from such an undesirable property by modifying the likelihood ratio so that it scans only the regions with elevated risk. The performance of the circular spatial scan statistic with the modified likelihood ratio is compared with Kulldorff's original one via Monte Carlo Simulations. The proposed circular spatial scan statistic is illustrated with mortality data from cerebrovascular disease in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan.

2. Motivating Example

As a motivating example, we shall introduce here a part of Tango and Takahashi's (2005) results of the application of three spatial scan statistics — Kulldorff's, Duczmal and Assunção's, and Tango and Takahashi's — to a simulated disease map in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan wherein there are $m = 113$ regions that comprise wards, cities, and villages (Figure 1). The variability of regional populations is: 25 percentile = 56,704, median = 142,320 and 75 percentile = 200,936. On this map, they simulated a random sample of $n = 235$ cases by assuming the hot-spot cluster regions $\{14, 15, 26, 27\}$ in which each region's relative risk was set to constant $\theta = 3.0$ and the cases to be Poisson distributed. Under the null hypothesis of no clustering, the expected number of cases for each region was set proportional to the population (see equation (2)). For comparison purpose, they set the maximum number of regions for the MLC to be $K = 15$ and used 999 replications for the Monte Carlo hypothesis testing (see section 3 for details). The results are shown in Table 1 and are summarized as follows.

- Kulldorff's circular spatial scan statistic detected 2 regions $\{14, 15\}$ as the MLC with log likelihood ratio $LLR = 20.1$ and Monte Carlo p -value $p = 1/(999 + 1) = 0.001$, and the estimated relative risk $\hat{\theta} = 3.47$.
- Tango and Takahashi's flexible spatial scan statistic detected 5 regions $\{14, 15, 26, 27, 33\}$ as

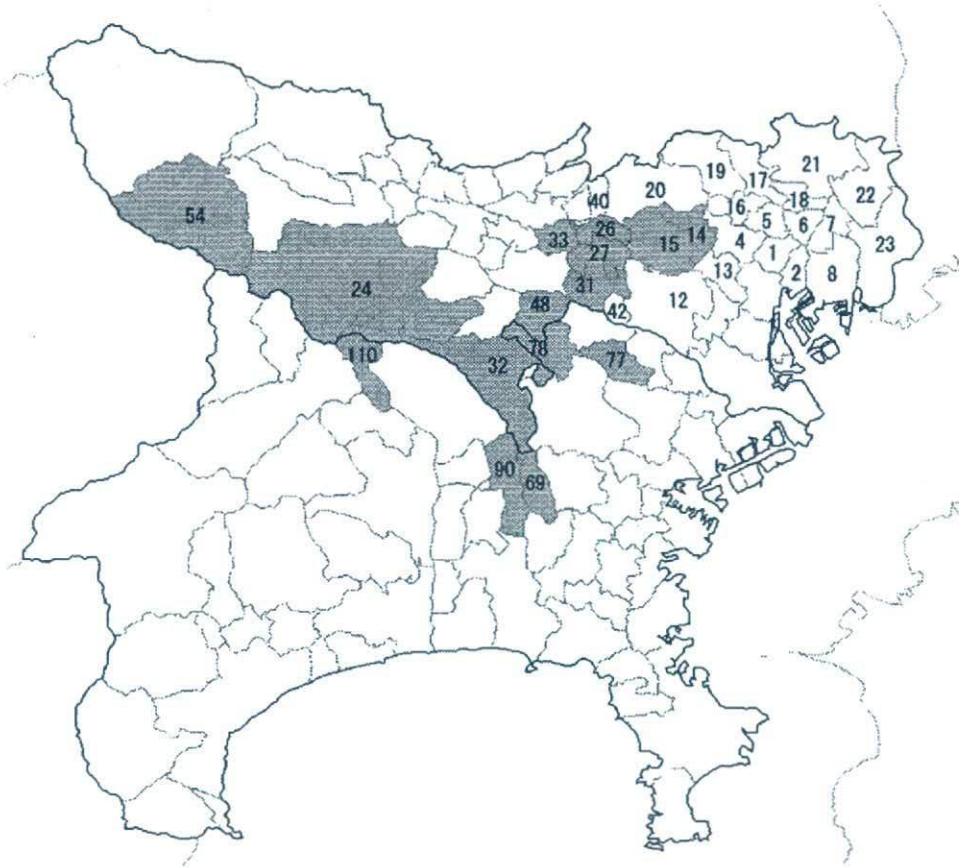


Fig. 1. An entire study population for simulation studies. The 113 regions that comprise wards, cities, and villages in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan. The region number used in the text is shown. The shaded area was detected as the *most likely cluster* by Duczmal and Assunção's spatial scan statistic (see section 2).

the MLC with $LLR = 29.7$ and $p = 0.001$, and $\hat{\theta} = 3.41$.

- Duczmal and Assunção's spatial scan statistic detected 15 connected regions $\{14, 15, 24, 26, 27, 31, 32, 33, 48, 54, 69, 77, 78, 90, 110\}$ as the MLC (shaded area in Figure 1) with $LLR = 31.8$ and $p = 0.001$, and $\hat{\theta} = 2.41$.

All these spatial scan statistics rejected the null hypothesis of " H_0 : there are no clusters," but detected different sets of connected regions as their MLC. In this example, Kulldorff's circular spatial scan statistic detected a small cluster consisting of only two regions out of four hot-spot cluster regions. In terms of maximizing the likelihood ratio, Duczmal and Assunção's spatial scan statistic is the best. However, their spatial scan statistic detected the MLC of a peculiar shape that was considerably larger than the true cluster, which includes eight regions $\{31, 48, 78, 32, 90, 69, 24, 110\}$ where each region's relative risk is not statistically significantly larger than 1 at 0.05 level. Table 1 shows the individual region's one-tailed *mid p-value*

Table 1. Regions detected as the most likely cluster by applying three procedures, Kulldorff's circular spatial scan, Tango and Takahashi's flexible spatial scan and Duczmal and Assuncao's spatial scan, to a simulated random sample $n = 235$ from the hot-spot cluster model. In the simulation, the true cluster is assumed to be $\{14, 15, 26, 27\}$ with constant relative risk $\theta = 3.0$. The maximum length (=number of regions) of cluster was set to be $K = 15$ (Tango and Takahashi, 2005).

No.	region no.	observed no. cases	expected no. cases	relative risk (true)	One-tailed p_i	Cumulative statistics	
						LLR	O/E
1	14	14	3.794	3.69(3)	0.000027	8.3	3.69
2	15	21	6.283	3.34(3)	0.000002	20.1	3.47
Circular's MLC = $\{14, 15\}$, LLR = 20.1, $\hat{\theta} = 3.47$							
3	26	6	1.650	3.64(3)	0.004	24.1	3.50
4	27	6	1.964	3.05(3)	0.010	27.3	3.43
5	33	4	1.257	3.18(1)	0.024	29.7	3.41
Flexible's MLC = $\{14, 15, 26, 27, 33\}$, LLR = 29.7, $\hat{\theta} = 3.41$							
6	31	3	2.346	1.28(1)	0.313	28.1	3.12
7	48	1	0.696	1.44(1)	0.328	27.8	3.06
8	78	2	1.485	1.35(1)	0.312	27.2	2.93
9	32	5	4.142	1.21(1)	0.318	25.2	2.63
10	77	5	2.109	2.37(1)	0.042	27.2	2.60
11	90	5	2.312	2.16(1)	0.057	29.0	2.57
12	69	3	1.419	2.11(1)	0.114	30.0	2.55
13	24	8	5.534	1.45(1)	0.152	30.0	2.37
14	110	1	0.256	3.91(1)	0.127	30.7	2.38
15	54	1	0.045	22.20(1)	0.022	31.8	2.41
Duczmal et al.'s MLC = $\{14, 15, 26, 27, 33, 31, \dots, 54\}$, LLR = 31.8, $\hat{\theta} = 2.41$							

defined later in (7). This surprising result casts a doubt on the validity of the model selection based upon maximizing the likelihood ratio. Such a doubt can also be seen in Tango and Takahashi's simulation results of Kulldorff's circular spatial scan statistic that had non-negligible probabilities of detecting much longer clusters than the true cluster.

3. A New Spatial Scan Statistic

Consider the situation where an entire study area is divided into m regions (for example, county, enumeration districts, etc.). The number of cases in the region i is denoted by the random variable N_i with observed value n_i ($i = 1, \dots, m$) and $n = n_1 + \dots + n_m$. Under the null hypothesis H_0 of no clustering, the N_i are independent Poisson variables such that

$$H_0: E(N_i) = \xi_i, N_i \sim \text{Pois}(\xi_i), \quad i = 1, \dots, m \quad (1)$$

where $\text{Pois}(\xi)$ denotes Poisson distribution with mean ξ and the ξ_i are the expected number of cases in the region i under the null hypothesis. For the calculation of the expected number of cases adjusted for the potential confounders such as age, we can use an indirect standardization or a

Poisson mixed-effects regression model (Kulldorff *et al.*, 2006a). If we can ignore the confounders, the ξ_i can be calculated as

$$\xi_i = n \frac{w_i}{\sum_{k=1}^m w_k}, \quad i = 1, \dots, m \quad (2)$$

where w_i denotes the population size in the region i . To specify the geographical position of each region, we will use the coordinates of the administrative population centroid.

Under this situation, Kulldorff's circular spatial scan statistic imposes a circular window \mathbf{Z} on each centroid. For any of those centroids, the radius of the circle varies continuously from zero upwards until 50 percent of the population at risk is covered, which is the standard option for SaTScan. 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} ($k = 1, \dots, K_i$) denote the window composed by the $(k - 1)$ -nearest neighbours to region i . Then, all the windows to be scanned by the circular spatial scan statistic are included in the set

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

Under the alternative hypothesis, there is at least one window $\mathbf{Z} \in \mathcal{Z}$ for which the underlying risk is higher inside the window when compared with outside. In other words, we are considering the following hypothesis testing:

$$\begin{aligned} H_0: E(N(\mathbf{Z})) &= \xi(\mathbf{Z}), \quad \text{for all } \mathbf{Z} \in \mathcal{Z} \\ H_1: E(N(\mathbf{Z})) &> \xi(\mathbf{Z}), \quad \text{for some } \mathbf{Z} \in \mathcal{Z} \end{aligned} \quad (3)$$

where $N()$ and $\xi()$ denote the random variable for the 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 assumption of Poisson distribution (1), Kulldorff's likelihood ratio test statistic is given by

$$\sup_{\mathbf{Z} \in \mathcal{Z}} \lambda_K(\mathbf{Z}) = \sup_{\mathbf{Z} \in \mathcal{Z}} \left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} \right)^{n(\mathbf{Z})} \left(\frac{n - n(\mathbf{Z})}{n - \xi(\mathbf{Z})} \right)^{n - n(\mathbf{Z})} I \left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} > \frac{n - n(\mathbf{Z})}{n - \xi(\mathbf{Z})} \right) \quad (4)$$

where $n()$ 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 ratio is defined as the *most likely cluster* (MLC). However, it seems something wrong to me that the likelihood ratio defined above does not take individual observed relative risk n_i/ξ_i into account even if the MLC includes some regions with non-elevated risk, such as a region with $n_i/\xi_i \approx 1$. This property leads to the following proposition.

Proposition:

In the process of scanning the window based on $\sup_{\mathbf{Z} \in \mathcal{Z}} \lambda_K(\mathbf{Z})$, there is a possibility that there exists two disjoint windows \mathbf{Z}_1 and \mathbf{Z}_2 and several regions $\{i_1\}, \dots, \{i_r\}$ such that

$$\lambda_K(\{\mathbf{Z}_1, \mathbf{Z}_2, \{i_1\}, \dots, \{i_r\}\}) > \max\{\lambda_K(\mathbf{Z}_1), \lambda_K(\mathbf{Z}_2)\} \quad (5)$$

where

$$\frac{n(\mathbf{Z}_1)}{\xi(\mathbf{Z}_1)} > 1, \quad \frac{n(\mathbf{Z}_2)}{\xi(\mathbf{Z}_2)} > 1 \quad \text{and} \quad \frac{n_i}{\xi_i} \leq 1 \quad (i = 1, \dots, r)$$

For example, we can easily consider the following example:

An example: Suppose that $n = 200$ and two windows \mathbf{Z}_1 ($n(\mathbf{Z}_1) = 30$, $\xi(\mathbf{Z}_1) = 12$, $\log \lambda_K(\mathbf{Z}_1) = 8.88$) and \mathbf{Z}_2 ($n(\mathbf{Z}_2) = 29$, $\xi(\mathbf{Z}_2) = 13$, $\log \lambda_K(\mathbf{Z}_2) = 7.97$) were created in the process of scanning windows. Let 5-nearest neighbours to \mathbf{Z}_1 be $\{i_1\}, \{i_2\}, \dots, \{i_4\}$, \mathbf{Z}_2 where four regions have non-elevated risk such as $(n_{i_1} = 4, \xi_{i_1} = 5)$, $(n_{i_2} = 6, \xi_{i_2} = 6)$, $(n_{i_3} = 0, \xi_{i_3} = 3)$, $(n_{i_4} = 8, \xi_{i_4} = 9)$. In this case, we have

$$\log \lambda_K(\{\mathbf{Z}_1, \mathbf{Z}_2, \{i_1\}, \dots, \{i_4\}\}) = 9.58$$

which satisfies equation (5).

The above proposition and example mean that if we allow any window and/or region to be a candidate for the MLC it causes the possibility of detecting an unrealistically large MLC by swallowing up neighboring regions with non-significantly elevated risk due to random fluctuation or with non-elevated risk as is shown in section 2. Therefore, to avoid such undesirable phenomena, we propose the following restricted likelihood ratio test statistic by taking individual region's risk into account:

$$\sup_{\mathbf{Z} \in \mathcal{Z}} \lambda_T(\mathbf{Z}) = \sup_{\mathbf{Z} \in \mathcal{Z}} \left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} \right)^{n(\mathbf{Z})} \left(\frac{n - n(\mathbf{Z})}{n - \xi(\mathbf{Z})} \right)^{n - n(\mathbf{Z})} I \left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} > \frac{n - n(\mathbf{Z})}{n - \xi(\mathbf{Z})} \right) \prod_{i \in \mathbf{Z}} I(p_i < \alpha_1) \quad (6)$$

where p_i is the one-tailed p -value of the test for $H_0: E(N_i) = \xi_i$ given by the *mid-p value*

$$p_i = \Pr\{N_i \geq n_i + 1 \mid N_i \sim \text{Pois}(\xi_i)\} + \frac{1}{2} \Pr\{N_i = n_i \mid N_i \sim \text{Pois}(\xi_i)\} \quad (7)$$

and α_1 is the pre-specified significance level for the individual region. The reason why the *mid-p value* was used is to adjust for conservatism of the ordinary definition of p -value for small ξ_i . In this formulation, we devised $I(p_i < \alpha_1)$ as a screening criterion and we do not mean that we are performing multiple hypothesis tests. Therefore, as in the case of the Kulldorff's circular spatial scan statistic, the restricted likelihood ratio test $\sup_{\mathbf{Z} \in \mathcal{Z}} \lambda_T(\mathbf{Z})$ of the nominal α_0 level of 0.05, say, can be based on its distribution derived from a large number of Monte Carlo replications of the data set generated under the null hypothesis. Under this framework, the p -value is obtained

through Monte Carlo hypothesis testing (Dwass, 1957), by comparing the rank of the maximum likelihood from the real data set with the maximum likelihoods from the random data sets. If this rank is R , then $p = R/(1 + \#(\text{replications}))$.

Well, how is α_1 chosen? If α_1 is chosen too small then clusters might be missed or be too small, however, if chosen too large then it becomes equivalent to the Kulldorff's spatial scan statistic. So, we shall examine how sensitive the proposed method is to the selection of α_1 and the optimal range of α_1 , by using five different values, 0.05, 0.10, 0.20, 0.30, 0.40, via Monte Carlo simulations in section 5, which might depend on both the cluster size and risk size.

It should be noted that the restricted likelihood ratio test statistic (6) is set to scan only for clusters with high relative risk. It also can be set to scan only for clusters with low relative risk by replacing the inequality ">" of the indicator function $I()$ which compares the observed relative risks within and outside the window \mathbf{Z} with "<" and using the following *mid-p* value:

$$p_i = \Pr\{N_i \leq n_i - 1 \mid N_i \sim \text{Pois}(\xi_i)\} + \frac{1}{2} \Pr\{N_i = n_i \mid N_i \sim \text{Pois}(\xi_i)\}. \quad (8)$$

When we would like to scan for either high or low relative risk, we have only to delete the indicator function regarding the comparison of the observed relative risks and replace α_1 with $\alpha_1/2$ in equation (6) using either *mid-p* value. Needless to say, maximizing the restricted likelihood ratio $\lambda_T(\mathbf{Z})$ can be applied not only to the circular spatial scan statistic but also to non-circular spatial scan statistics such as Duczmal and Assunção's (2004) and Tango and Takahashi's (2005).

4. Illustrations

As an illustration, we shall apply the proposed circular spatial scan statistic and Kulldorff's circular spatial scan statistic to the mortality data from cerebrovascular disease (female, 1993-1997) in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan. Total observed number of deaths from female cerebrovascular disease for five years was 45,700 in this area. Regarding the restriction of the maximum length for the MLC, we shall select a standard option of SaTScan that the radius of the circle varies continuously from zero upwards until 50 percent of the population at risk is covered. P -value of the spatial scan statistics is calculated using 9999 replications for the Monte Carlo hypothesis testing. Significance levels considered are $\alpha_0 = 0.05$ and $\alpha_1 = 0.05, 0.10, 0.20, 0.30, 0.40$. Results are shown in Table 2 and summarized as follows: Irrespective of significance levels α_1 chosen, the proposed circular spatial scan statistic detected the same MLC consisting of six regions $\{23, 22, 7, 8, 6, 18\}$ (these region numbers are shown in Figure 1) with $\log \lambda_T (= \log \lambda_K) = 126.6$, $\hat{\theta} = 1.23$, and $p = 1/(999 + 1) = 0.001$. All of these six regions within the MLC are shown to have highly significantly elevated risk. In contrast, Kulldorff's circular spatial scan statistic added six more regions $\{5, 1, 17, 21, 2, 16\}$ to the MLC detected by the proposed circular spatial scan statistic and finally detected a larger MLC consisting of 12 regions with $\log \lambda_K = 140.6$ ($\log \lambda_T = 0$), $\hat{\theta} = 1.17$, and $p = 0.001$. However, three regions $\{5, 1, 2\}$ do not have significantly elevated risks. Especially, the region $\{5\}$, called

Table 2. The most likely clusters detected at $\alpha_0 = 0.05$ by the proposed circular spatial scan statistic and Kulldorff's circular spatial scan statistic in their application to the mortality data from cerebrovascular disease (female, 1993-1997) in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan. The MLC detected by the proposed procedure was the same irrespective of values of $\alpha_1 = 0.05, 0.10, 0.20, 0.30, 0.40$.

No.	region no.	observed no. cases	expected no. cases	relative risk	One-tailed p_i
1	23	1297	1072.3	1.21	1.5×10^{-11}
2	22	1266	1013.2	1.25	1.0×10^{-14}
3	18	738	522.7	1.41	$< 1.0 \times 10^{-17}$
4	7	737	620.4	1.19	2.7×10^{-6}
5	8	896	780.6	1.15	2.7×10^{-5}
6	6	678	550.5	1.23	7.7×10^{-8}
The proposed circular scan's MLC = {23, 22, 18, 7, 8, 6} $\log \lambda_T = 126.6, \hat{\theta} = 1.23, p\text{-value} = 0.001$					
7	5	548	566.3	0.97	0.778
8	1	164	144.7	1.13	0.057
9	17	1110	999.3	1.11	0.00029
10	21	1530	1335.1	1.15	9.2×10^{-8}
11	2	267	251.3	1.06	0.161
12	16	798	743.4	1.07	0.024
Kulldorff's circular scan's MLC = {23, 22, 18, 7, 8, 6, 5, 1, 17, 21, 2, 16} $\log \lambda_K = 140.6, \hat{\theta} = 1.17, p\text{-value} = 0.001$					

"Bunkyo-ku," is well known as "healthy district" in this area and has a low relative risk of 0.97 and $p_5 = 0.778$ as expected. Therefore, it seems unacceptable to residents in "Bunkyo-ku" that their region is added to the MLC.

It should be noted that three regions {17, 21, 16} have highly significant risk and are adjacent to the MLC detected by the proposed procedure. Therefore, the window {23, 22, 7, 8, 6, 18, 17, 21, 16} could be a candidate for the MLC. In fact, its log likelihood ratio is $\log \lambda_T = 149.8$ which is larger than that of two MLCs. Needless to say, this non-circular window cannot be detected by circular spatial scan statistics and thus may be detected by such non-circular spatial scan statistic as that proposed by Duczmal and Assunção (2004) or Tango and Takahashi (2005) by applying the restricted likelihood ratio test statistic.

5. Power Comparisons

In this section, we shall compare the power of Kulldorff's circular spatial scan statistic with that of the proposed circular spatial scan statistic via Monte Carlo simulations at significance level $\alpha_0 = 0.05$ and examine the effect of the choice of α_1 by using five different values $\alpha_1 = 0.05, 0.10, 0.20, 0.30, 0.40$. As an entire study population, we will use the area of Tokyo Metropolis and Kanagawa prefecture in Japan (Figure 1), described in section 2. In our simulation study, we consider two extremes regarding the sample size n : one assume rare disease with

total observed number of cases $n = 200$, a similar sample size used in section 2 and the other assumes non-rare disease with $n = 45,700$, the same n for the cerebrovascular disease illustrated in section 4.

The same null hypothesis is used throughout, where the relative risk is set to one for each region and case locations are independent of each other. Although the N_i are independent Poisson random variables, by conditioning on the total number of cases $N_+ = n$, the disease locations are the values of a random sample of size n from a multinomial distribution with parameter $(\xi_1/n, \dots, \xi_m/n)$. We generated 10,000 random data sets with n cases and these are used to estimate the upper $100\alpha_0$ percent point for significance. Under each alternative hypothesis, we generated 1,000 random data sets with n cases.

Alternative hypotheses H_1 considered here assume a single cluster $\mathbf{R} = \{i_1, \dots, i_{s^*}\}$ with values of relative risk $\boldsymbol{\theta}_R = (\theta_{i_1}, \dots, \theta_{i_{s^*}})$, i.e.,

$$H_1: N_{i_k} \sim \text{Pois}(\theta_{i_k} \xi_{i_k}), \quad k = 1, \dots, s^* \tag{9}$$

where s^* denote the size or length of the cluster assumed in the simulation. As cluster models, we considered here two types of clusters, namely, "clinal" and "hot-spot" clusters, defined by Wartenberg and Greenberg (1990). Clinal clusters have a monotone decrease in disease risk as distance from the point source increases. Hot-spot clusters are characterized by a constant elevated disease risk, i.e., $\theta_{i_1} = \dots = \theta_{i_{s^*}} = \theta$. Under H_1 , we generated a random sample (n_1, \dots, n_m) of size n from a multinomial distribution with parameter (q_1, \dots, q_m) where

$$q_i = \frac{\pi_i w_i}{\sum_{k=1}^m \pi_k w_k} \tag{10}$$

where

$$\pi_i = \begin{cases} \theta_i & \text{if } i \in \mathbf{R} \\ 1, & \text{otherwise.} \end{cases}$$

In order to compare the performance of the spatial scan statistic based on Monte Carlo simulation, we shall use the bivariate power distribution $P(l, s)$ proposed by Tango and Takahashi (2005), which is classified according to the length l of the significant MLC and the number s of the assumed hot-spot regions included therein. It is estimated by

$$P(l, s) = \frac{\#\{\text{significant MLC has length } l \text{ and includes } s \text{ true regions}\}}{\#\{\text{trials for each simulation}\}}, \tag{11}$$

where $1 \leq l$ and $0 \leq s \leq \min\{l, s^*\}$. Especially, we are interested in the power around the point $(l = s^*, s = s^*)$ and $P(s^*, s^*)$, the probability of exact detection. Then, the usual power is defined as the sum of $P(l, s)$:

$$P(+, +) = \sum_{l=1}^{l_{\max}} \sum_{s=0}^{\min\{l, s^*\}} P(l, s) = 1 - P(0, 0) \tag{12}$$

where l_{\max} denotes the maximum length l observed in the simulation and $P(0,0)$ denotes the probability that the spatial scan statistic cannot detect any clusters. However, when the support of the bivariate power distribution is widely scattered over the plane (l, s) , we do not think that we can use the usual power as the primary criterion to evaluate the performance because the performance of the spatial scan statistic is not only to reject the null hypothesis but also to identify the cluster accurately.

5.1 Case of Rare Disease

We examined several sets of cluster models with different relative risk θ_R and cluster length s^* and we shall report here the results of the following two hot-spot clusters and two clinal clusters since these models provided us with typical results.

1. Hot-spot cluster **A** = {14, 15, 20} with relative risk $\theta_A = (3.0, 3.0, 3.0)$, ($s^* = 3$).
2. Clinal cluster **B** = {14, 15, 26, 27} with relative risk $\theta_B = (2.5, 3.0, 3.0, 2.5)$, ($s^* = 4$)
3. Hot-spot cluster **C** = {1, 4, 5, 12, 13, 14, 15, 16, 19, 20} with relative risk $\theta_C = (2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0)$, ($s^* = 10$)
4. Clinal cluster **D** = {12, 14, 15, 20, 26, 27, 31, 33, 40, 42} with relative risk $\theta_D = (1.8, 1.8, 1.8, 2.2, 2.4, 2.4, 2.2, 2.2, 2.0, 1.8)$, ($s^* = 10$)

Two clusters **A** and **B** are examples of a very sharp increase in risk in a very small area in which the increase in each region within the cluster has more than 65% chance of being significant at $\alpha_1 = 0.05$ level. The clusters **C** and **D** are, on the other hand, examples of smaller increase in risk in a larger area consisting of $s^* = 10$ regions, in which the increase in each region within the cluster has less than 50% chance of being significant at $\alpha_1 = 0.05$ level. Regarding the shape of cluster, the clusters **A** and **C** are circular clusters that can be in the set of the circular windows and is expected to be identified exactly by the circular spatial scan statistic. The clusters **B** and **D** are non-circular clusters that are not in the set of the circular windows and thus cannot be identified exactly by the circular spatial scan statistic. The estimated bivariate power distribution $P(l, s) \times 1000$ is shown in Tables 3-6 for each of the three cluster models, respectively, in the form of cross table classified by l ("length" in tables) and s ("include" in tables).

Table 3 shows a good characteristic of the proposed circular spatial scan statistic. Namely, it could detect the hot-spot circular cluster **A** with length $s^* = 3$ considerably accurately with powers $P(3, 3) = 0.834, 0.864, 0.767, 0.755, 0.748$ for $\alpha_1 = 0.05, 0.10, 0.20, 0.30, 0.40$, respectively. When we also evaluate the power of detecting the true cluster plus additional one region, its power was also high without reference to the value of α_1 , i.e., $P(3, 3) + P(4, 3) = 0.882, 0.976, 0.957, 0.918, 0.906$ for $\alpha_1 = 0.05, 0.10, 0.20, 0.30, 0.40$, respectively. The maximum length of the MLCs among all the results was only $l = 8$. On the other hand, Kulldorff's circular spatial scan statistic also had relatively high powers such as $P(3, 3) = 0.672$ and $P(3, 3) + P(4, 3) = 0.819$. However, the estimated bivariate power distribution had a long tail to the right on the line $s = 3$ and the maximum length was 47. Usual power of the proposed circular spatial scan statistic was larger

Table 3. Estimated bivariate power distributions $P(l, s) \times 1000$ of Kulldorff's circular spatial scan statistic and the proposed circular spatial scan statistic for the hot-spot circular cluster $A = \{14, 15, 20\}$ with relative risk $\theta_A = (3.0, 3.0, 3.0)$ and $n = 200$. $\alpha_0 = 0.05$ and 1000 trials were carried out.

Kulldorff's circular scan					Proposed circular scan						
Length l	Include s hot-spot regions				α_1	Length l	Include s hot-spot regions				usual power
	0	1	2	3			0	1	2	3	
1	1	0			0.05	1	1	0			0.885
2	0	0	0			2	1	0	0		
3	0	0	0	672		3	0	0	0	834	
4	0	0	0	147		4	0	0	0	48	
5	0	0	0	38		5	0	0	0	1	
6	0	0	0	26	0.10	1	1	0			0.984
7	0	0	0	8		2	0	0	0		
8	0	0	0	3		3	1	0	0	864	
9	0	0	0	1		4	0	0	0	112	
10	0	0	0	3		5	0	0	0	5	
11	0	0	0	3		6	0	0	0	1	
12	0	0	0	13	0.20	1	1	0			0.978
13	0	0	0	15		2	0	0	0		
14	0	0	0	4		3	1	0	0	767	
15	0	0	0	2		4	0	0	0	190	
16-20	0	0	0	13		5	0	0	0	16	
21-25	0	0	0	6		6	0	0	0	3	
26-30	0	0	0	1	0.30	1	1	0			0.978
31-35	0	0	0	2		2	0	0	0		
36-40	0	0	0	3		3	1	0	0	755	
41-45	0	0	0	1		4	0	0	0	163	
46-47	0	0	0	1		5	0	0	0	42	
						6	0	0	0	16	
					0.40	1	1	0			0.978
						2	0	0	0		
						3	0	0	0	748	
						4	0	0	0	158	
						5	0	0	0	45	
						6	0	0	0	22	
						7-8	0	0	0	4	
Total	1	0	0	962							

usual power = 0.963

than that of Kulldorff's circular spatial scan statistic except for the case of $\alpha_1 = 0.05$.

Table 4, on the other hand, shows that both Kulldorff's and the proposed circular spatial scan statistics had zero probability $P(4, 4) = 0/1000$ for the exact detection since the clinal cluster **B** is non-circular. In this case, the support of the estimated bivariate power distribution of Kulldorff's circular spatial scan statistic was scattered over the plane $\{(l, s) : l \geq s, l = 1, 2, \dots, 64, s = 0, 2, 3, 4\}$ with the highest $P(2, 2) = 0.311$. This type of simulation results are quite typical of

Table 4. Estimated bivariate power distributions $P(l, s) \times 1000$ of Kulldorff's circular spatial scan statistic and the proposed circular spatial scan statistic for the clinal non-circular cluster $\mathbf{B} = \{14, 15, 26, 27\}$ with relative risk $\theta_B = (2.5, 3.0, 3.0, 2.5)$ and $n = 200$. $\alpha_0 = 0.05$ and 1000 trials were carried out.

Kulldorff's circular scan						Proposed circular scan							
Length l	Include s clinal regions					α_1	Length l	Include s clinal regions					usual power
	0	1	2	3	4			0	1	2	3	4	
1	5	0				0.05	1	9	0				0.748
2	0	0	311				2	4	0	709			
3	0	0	6	0			3	0	1	25	0		
4	0	0	5	0	0	0.10	1	8	0				0.754
5	0	0	2	0	0		2	3	0	689			
6	0	0	1	0	0		3	2	2	41	0		
7	0	0	1	15	0		4	1	0	8	0	0	
8	0	0	5	7	18		5	1	0	0	0	0	
9	0	0	1	1	9	0.20	1	7	0				0.685
10	0	0	2	15	5		2	3	0	602			
11	0	0	1	1	3		3	2	1	36	0		
12	0	0	0	3	32		4	1	1	26	0	0	
13	0	0	1	2	35		5	1	0	3	0	0	
14	0	0	0	4	16		6	0	0	0	2	0	
15	0	0	0	1	2	0.30	1	6	0				0.689
16-20	0	0	2	2	53		2	3	0	588			
21-25	0	0	0	0	43		3	2	1	32	0		
26-30	0	0	0	0	16		4	1	1	30	0	0	
31-35	0	0	0	0	15		5	1	0	4	0	0	
36-40	0	0	0	0	5		6	0	0	5	11	0	
41-45	0	0	0	0	5		7	0	0	0	0	0	
46-50	0	0	0	0	7		8	0	1	1	1	2	
51-55	0	0	0	1	1	0.40	1	5	0				0.693
56-60	0	0	0	0	0		2	2	0	583			
61-64	0	0	0	1	1		3	1	1	30	0		
							4	1	1	29	0	0	
							5	1	0	7	0	0	
							6	0	0	4	0	0	
							7	0	0	1	3	0	
							8	0	0	2	4	2	
							9	0	0	0	2	2	
							10-12	0	0	0	2	0	
Total	5	0	338	53	266								

usual power = 0.662

Kulldorff's circular spatial scan statistic when applied to non-circular clusters in our experience. In contrast, the proposed circular spatial scan statistic was shown to have higher powers for detecting a half of the true cluster, i.e., $P(2, 2) + P(3, 2) = 0.733, 0.730, 0.638, 0.620, 0.613$, for $\alpha_1 = 0.05, 0.10, 0.20, 0.30, 0.40$, respectively. Incidentally, the results that both spatial scan statistics had the highest detection probability at $(l, s) = (2, 2)$ seem to be due to the geograph-

Table 5. Estimated bivariate power distributions $P(l, s) \times 1000$ of Kulldorff's circular spatial scan statistic for the hot-spot circular cluster $C = \{1, 4, 5, 12, 13, 14, 15, 16, 19, 20\}$ with relative risk $\theta_C = (2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0)$ and $n = 200$. $\alpha_0 = 0.05$ and 1000 trials were carried out.

Length l	Kulldorff's circular scan										
	Include s hot-spot regions										
	0	1	2	3	4	5	6	7	8	9	10
1	5	0									
2	0	0									
3	0	0	0	0							
4	0	0	0	0	0						
5	0	0	0	0	0	0					
6	0	0	0	0	1	0	0				
7	0	0	0	0	0	0	0	0			
8	0	0	0	0	0	0	0	0	0		
9	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	0	0	0	323
11	0	0	0	0	0	0	0	0	9	0	56
12	0	0	0	0	0	0	0	0	12	0	23
13	0	0	0	0	0	0	0	0	5	4	97
14	0	0	0	0	0	0	0	0	5	1	16
15	0	0	0	0	0	0	0	0	0	3	32
16	0	0	0	0	0	0	0	0	0	5	19
17	0	0	0	0	0	0	0	0	0	1	28
18	0	0	0	0	0	0	0	0	0	0	28
19	0	0	0	0	0	0	0	0	0	0	12
20	0	0	0	0	0	0	0	0	0	0	9
21-25	0	0	0	0	0	0	0	0	0	0	38
26-30	0	0	0	0	0	0	0	0	13	0	21
31-35	0	0	0	0	0	0	0	0	4	0	11
36-40	0	0	0	0	0	0	0	0	1	6	8
41-45	0	0	0	0	0	0	0	0	0	0	5
46-50	0	0	0	0	0	0	0	0	1	0	4
51	0	0	0	0	0	0	0	0	0	0	1
Total	5	0	0	0	1	0	0	0	50	20	731

usual power = 0.807

ical configuration of the non-circular cluster **B**. In this case, the proposed circular spatial scan statistic had larger usual powers than Kulldorff's circular spatial scan statistic without reference to the value of α_1 . Judging from the results shown in Tables 3-4, the result with $\alpha_1 = 0.10$ could be the best, however, the results with $\alpha_1 = 0.20 \sim 0.40$ were also shown to be better than Kulldorff's results.

Tables 5-6 show the results for the hot-spot circular cluster **C** where the affected area of regions is larger and the relative risk is smaller. In this instance, Kulldorff's circular spatial scan statistic was expected to detect changes in larger area that may not be obvious from each region individually and had surely high usual power 0.807 with the exact detection probability

Table 6. Estimated bivariate power distributions $P(l, s) \times 1000$ of the proposed circular spatial scan statistic for the hot-spot circular cluster $C = \{1, 4, 5, 12, 13, 14, 15, 16, 19, 20\}$ with relative risk $\theta_C = (2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0)$ and $n = 200$. $\alpha_0 = 0.05$ and 1000 trials were carried out.

		Proposed circular scan										usual power				
α_1	Length l	Include s hot-spot regions														
		0	1	2	3	4	5	6	7	8	9		10			
0.05	1	10	0													0.019
	2	6	0	0												
	3	1	0	0	0											
	4	0	0	0	0	2										
0.10	1	8	0													0.161
	2	7	0	0												
	3	2	0	0	0											
	4	0	0	0	0	110										
	5	0	0	0	0	0	33									
	6	0	0	0	0	0	1	0								
0.20	1-4	12	0	0	0	0										0.548
	5	0	0	0	0	0	403									
	6	0	0	0	0	0	58	0								
	7	0	0	0	0	0	0	3	0							
	8	0	0	0	0	0	0	0	0	0						
	9	0	0	0	0	0	0	0	0	0	0	71				
	10-12	0	0	0	0	0	0	0	0	0	0	1	0			
0.30	1-5	9	0	0	0	0	0									0.741
	6	0	0	0	0	1	0	0								
	7	0	0	0	0	0	0	3	0							
	8	0	0	0	0	0	0	0	0	138						
	9	0	0	0	0	0	0	0	2	0	575					
	10	0	0	0	0	0	0	0	1	0	0	0				
	11-12	0	0	0	0	0	0	0	0	9	3	0				
0.40	1-5	9	0	0	0	0	0									0.744
	6	0	0	0	0	1	0	0								
	7	0	0	0	0	0	0	2	0							
	8	0	0	0	0	0	0	0	0	135						
	9	0	0	0	0	0	0	0	6	0	568					
	10	0	0	0	0	0	0	0	1	0	0	0				
	11	0	0	0	0	0	0	0	0	8	0	0				
	12	0	0	0	0	0	0	0	0	4	7	0				
	13-14	0	0	0	0	0	0	0	0	3	0	0				

$P(10, 10) = 0.323$. However, the support of the estimated bivariate power distribution was again scattered over the broad area on the plane $\{(l, s): l \geq s, l = 1, 2, \dots, 51, s = 0, 4, 8, 9, 10\}$ by swallowing up many additional regions. In contrast, Table 6 shows that the support of the proposed circular spatial scan statistic was distributed in a relatively confined area on the plane (l, s) . However, the results with $\alpha_1 = 0.05$ and 0.10 were miserable, i.e., their usual power was quite low 0.039 and 0.161, respectively. As the value of α_1 increases, the usual power increased to

0.548, 0.741 and 0.744 for $\alpha_1 = 0.20, 0.30$ and 0.40 , respectively. Although the probability of detecting the cluster exactly was zero without reference to the value of α_1 chosen, the proposed circular spatial scan statistic was shown to have high probability of pinpointing eight or nine regions out of $s^* = 10$ regions, i.e., $P(8, 8) + P(9, 9) = 0.713, 0.703$ for $\alpha_1 = 0.30, 0.40$, respectively.

Tables 7-8 shows the results for the clinal non-circular cluster **D** where the affected area of regions is larger and the relative risk is smaller. Due to the non-circular shape, 1) both Kulldorff's and the proposed circular spatial scan statistic had zero probability of detecting the

Table 7. Estimated bivariate power distributions $P(l, s) \times 1000$ of Kulldorff's circular spatial scan statistic for the clinal non-circular cluster $\mathbf{D} = \{12, 14, 15, 20, 26, 27, 31, 33, 40, 42\}$ with relative risk $\theta_D = (1.8, 1.8, 1.8, 2.2, 2.4, 2.4, 2.2, 2.2, 2.0, 1.8)$ and $n = 200$. $\alpha_0 = 0.05$ and 1000 trials were carried out.

Length <i>l</i>	Kulldorff's circular scan										
	Include <i>s</i> clinal regions										
	0	1	2	3	4	5	6	7	8	9	10
1	3	0									
2	0	0	0								
3	0	0	0	0							
4	0	0	0	0	24						
5	0	0	0	0	0	0					
6	0	0	0	0	0	12	0				
7	0	0	0	0	0	0	0	0			
8	0	0	0	0	0	0	29	0	0		
9	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	32	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	6	0	0
13	0	0	0	0	0	0	0	0	0	1	0
14	0	0	0	0	0	0	0	1	0	0	0
15	0	0	0	0	0	0	0	0	0	13	0
16	0	0	0	0	0	0	0	2	0	7	0
17	0	0	0	0	0	0	0	6	0	4	0
18	0	0	0	0	0	0	0	1	0	6	40
19	0	0	0	0	0	0	0	3	0	1	90
20	0	0	0	0	0	0	0	0	1	1	41
21-25	0	0	0	0	0	0	0	2	7	1	147
26-30	0	0	0	0	0	0	1	1	0	4	63
31-35	0	0	0	0	0	0	0	1	0	3	46
36-40	0	0	0	0	0	0	0	0	1	1	28
41-45	0	0	0	0	0	0	1	0	0	0	7
46-50	0	0	0	0	0	0	0	0	0	0	7
51-55	0	0	0	0	0	0	0	0	0	0	6
56-60	0	0	0	0	0	0	0	0	0	0	2
61-62	0	0	0	0	0	0	0	0	0	0	3
Total	3	0	0	0	24	12	31	49	15	42	480

usual power = 0.656

Table 8. Estimated bivariate power distributions $P(l, s) \times 1000$ of the proposed circular spatial scan statistic for the clinal non-circular cluster $D = \{12, 14, 15, 20, 26, 27, 31, 33, 40, 42\}$ with relative risk $\theta_D = (1.8, 1.8, 1.8, 2.2, 2.4, 2.4, 2.2, 2.2, 2.0, 1.8)$ and $n = 200$. $\alpha_0 = 0.05$ and 1000 trials were carried out.

α_1	Length l	Proposed circular scan										usual power	
		Include s clinal regions											
		0	1	2	3	4	5	6	7	8	9	10	
0.05	1	12	0										0.039
	2	6	3	0									
	3	0	0	0	0								
	4	0	0	0	0	18							
0.10	1	6	0										0.285
	2	4	1	0									
	3	2	0	0	0								
	4	0	0	0	0	266							
	5	0	0	0	0	6	0						
0.20	1-3	9	1	0	0								0.496
	4	0	1	0	1	410							
	5	0	0	0	3	68	0						
	6	0	0	0	0	0	1	0					
	7	0	0	0	0	0	0	0	0				
	8	0	0	0	0	0	0	2	0	0			
0.30	1-3	7	0	0	0								0.513
	4	0	1	0	1	356							
	5	0	0	0	3	0	0						
	6	0	0	0	0	0	101	0					
	7	0	0	0	0	0	17	0	0				
	8	0	0	0	0	1	0	23	0	0			
	9	0	0	0	0	0	0	3	0	0	0		
0.40	1-3	5	0	0	0								0.512
	4	0	1	0	0	353							
	5	0	0	0	2	0	0						
	6	0	0	0	0	0	97	0					
	7	0	0	0	0	1	0	0	0				
	8	0	0	0	0	4	0	41	0	0			
	9	0	0	0	0	1	0	4	0	0	0		
	10	0	0	0	0	0	0	0	0	0	0	0	
	11	0	0	0	0	0	0	2	0	0	0	0	
	12	0	0	0	0	0	0	0	0	0	0	0	
	13	0	0	0	0	0	0	0	0	0	1	0	

cluster D exactly and 2) the support of the Kulldorff's estimated bivariate power distribution was scattered over the broader area than that for the cluster C on the plane $\{(l, s): l \geq s, l = 1, 2, \dots, 62, s = 0, 4, 5, \dots, 10\}$ by swallowing up many additional regions. Similar to the results for the cluster C , the support of the proposed circular spatial scan statistic was distributed in a relatively confined area on the plane (l, s) but the result with $\alpha_1 = 0.05$ was also miserable, i.e., the usual power was quite low 0.039. As the value of α_1 increases, the usual power increased

to 0.285, 0.490, 0.513, and 0.512 for $\alpha_1 = 0.10, 0.20, 0.30$ and 0.40 , respectively. The results that the highest probability was observed at $(l, s) = (4, 4)$ irrespective of α_1 also seems to be due to the geographical configuration of the cluster **D**.

The results shown in Tables 5-8 suggest that, for larger clusters with smaller relative risk, 1) the usual power of Kulldorff's circular spatial scan statistic was higher than that of the proposed circular spatial scan statistic, 2) but the Kulldorff's MLCs tended to be much larger than the true cluster to try to include the true cluster within the MLC, and 3) the proposed circular spatial scan statistic had higher powers of detecting a part of the true cluster when $\alpha_1 = 0.30$ or 0.40 .

5.2 Case of Non-rare Disease

As an example of cluster models of non-rare diseases, we adopted the most likely cluster **E** = $\{6, 7, 8, 18, 22, 23\}$ with $s^* = 6$ to be detected by the proposed scan statistic in section 5. The total observed number of cases in the study area is assumed to be $n = 45,700$ and then the expected number of cases per region is roughly on the average 400. As the relative performance of the proposed scan statistic compared with Kulldorff's scan statistic has been quite similar regardless of the cluster type in our experience, we considered here a hot-spot cluster and changed the value of constant relative risk to examine the effect of the size of relative risk on the performance of both scan statistics. As the value of constant relative risk θ for the cluster **E**, we have chosen five values $\theta = 1.04, 1.06, 1.08, 1.10, 1.12$ so that the probability of being significant at 0.05 level under individual region's test within the cluster has a range from $24 \sim 37\%$ ($\theta = 1.04$) to $85 \sim 99\%$ ($\theta = 1.12$). In Table 9, we omitted the estimated bivariate power distributions for ease of comparison and, instead, showed three measures, 1) the exact detection probability $P(6, 6)$, 3) usual power and 4) $\max l$ extracted from the estimated bivariate power distributions. As to the proposed scan statistic, we selected three results with $\alpha_1 = 0.1, 0.2, 0.3$. Table 9 shows that the proposed scan statistic performed better than Kulldorff's scan statistic for $\alpha_1 = 0.20$ and 0.30 .

5.3 Choice of α_1

Tables 3-9 indicate that the probability of detecting irrelevant regions outside the true cluster tends to get larger as the value of α_1 becomes larger in all the results. Tables 3-4 show that the case of $\alpha_1 = 0.10$ could be the best but the case of $\alpha_1 = 0.20 \sim 0.40$ were also good. Tables 6, 8 show that the case of $\alpha_1 = 0.30$ or 0.40 could be the best. Table 9 shows that the case of $\alpha_1 = 0.20$ could be the best for $\theta = 1.04 \sim 1.08$ and the case of $\alpha_1 = 0.10$ could be the best for $\theta = 1.10 \sim 1.12$. However, the results with $\alpha_1 = 0.20$ or 0.30 were also shown to have good results for all the values of θ considered.

These results suggest that, for $\alpha_0 = 0.05$, we might have the following guidance:

1. $\alpha_1 = 0.10 \sim 0.20$ to detect small clusters with a sharp increase in risk.
2. $\alpha_1 = 0.20 \sim 0.30$ to detect small to middle-sized clusters with a moderate increase in risk.
3. $\alpha_1 = 0.30 \sim 0.40$ to detect larger clusters with a slight increase in risk.

Table 9. Three sorts of powers extracted from the estimated bivariate power distributions $P(l, s) \times 1000$ of Kulldorff's circular spatial scan statistic and the proposed circular spatial scan statistic for the hot spot circular cluster $E = \{6, 7, 8, 18, 22, 23\}$ with the common relative risk θ , where $\theta = 1.04, 1.06, 1.08, 1.10, 1.12$ and $n = 45,700$. $\alpha_0 = 0.05$ and 1000 trials were carried out. Bold numbers indicate the optimal values.

θ	Prob.* (%)	Kulldorff's circular scan			Proposed circular scan			
		$P(6,6)$	Usual power	max l	α_1	$P(6,6)$	Usual power	max l
1.04	24 ~ 37	0.047	0.167	30	0.10	0.061	0.111	7
					0.20	0.106	0.208	9
					0.30	0.090	0.205	10
1.06	39 ~ 62	0.245	0.556	29	0.10	0.337	0.448	8
					0.20	0.421	0.634	10
					0.30	0.351	0.630	10
1.08	57 ~ 83	0.492	0.897	27	0.10	0.729	0.857	8
					0.20	0.678	0.932	10
					0.30	0.573	0.931	10
1.10	73 ~ 95	0.641	0.992	18	0.10	0.875	0.985	8
					0.20	0.757	0.996	10
					0.30	0.672	0.996	10
1.12	85 ~ 99	0.722	1.000	18	0.10	0.901	0.999	8
					0.20	0.791	1.000	9
					0.30	0.737	1.000	10

*: The range of probability of being significant at 0.05 level under individual region's test within the cluster.

Other simulation studies considering other cluster models resulted in a similar finding (data not shown). Although the risk size and cluster size assumed in a geographical cluster detection study or disease surveillance might depend on their purposes, it seems to me that we can set $\alpha_1 = 0.20$ as a default based on our simulation results.

6. Discussion

Our work is motivated by a surprising result observed in a simulation study done by Tango and Takahashi (2005) where Duczmal and Assunção's spatial scan statistic (2004) detected quite large and unlikely peculiar shaped clusters that had the largest likelihood ratio among the three spatial scan statistics. To avoid such an undesirable property of the likelihood ratio test statistic proposed by Kulldorff (1997), Tango and Takahashi discussed a possibility of using penalized likelihood approach where a penalty is considered for the *complexity of the cluster shape*. Kulldorff *et al.* (2006b) explored an elliptic version of the spatial scan statistic by introducing the *eccentricity penalty* that discourages eccentric clusters. Regarding the penalized likelihood approach, Assunção *et al.* (2006) made an important comment that this approach is a possible solution but certainly plagued with a large dose of subjectivity in the penalty parameters and