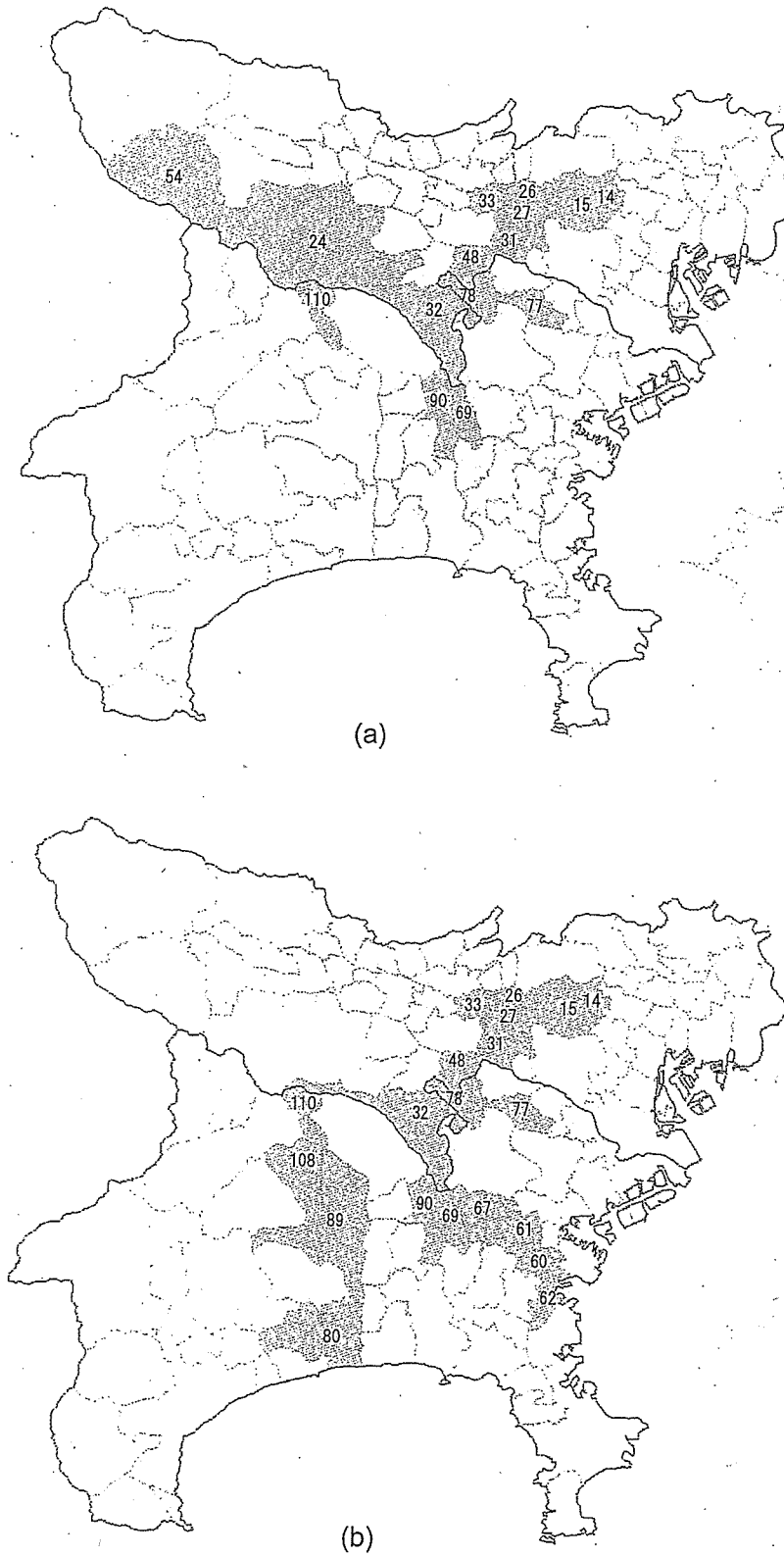

Figure 4
The most likely cluster detected by the Duczmal and Assunção's scan statistic. (a) Detected for $K = 15$ and (b) for $K = 20$, when applied to a random sample from the cluster model $\mathbf{C} = \{14, 15, 26, 27\}$.

Table 1: Regions detected as the most likely cluster by three procedures. Regions detected as the most likely cluster by the circular scan, the flexible scan and Duczmal and Assunção's scan, with the maximum length of cluster set to be $K = 15$ for the simulated random sample from the cluster model C where the hot spot cluster is assumed to be the set of connected four regions {14, 15, 26, 27} with the assumed relative risk $\theta = 3.0$. For details, see text.

region no.	population	observed no. cases	expected no. cases	relative risk estimated (true)	Log likelihood ratio (LLR) and estimated relative risk $\hat{\theta}$ for the most likely cluster		
					Circular	Flexible	Duczmal et al.
14	319,687	14	3.794	3.69 (3.0)	*	*	*
15	529,485	21	6.283	3.34 (3.0)	*	*	*
					LLR = 20.1		
					$\hat{\theta}$ = 3.47		
26	139,077	6	1.650	3.64 (3.0)		*	*
27	165,564	6	1.964	3.05 (3.0)		*	*
33	105,899	4	1.257	3.18 (1.0)		*	*
						LLR = 29.7	
						$\hat{\theta}$ = 3.41	
24	466,347	8	5.534	1.44 (1.0)			*
31	197,677	3	2.346	1.27 (1.0)			*
32	349,050	5	4.142	1.20 (1.0)			*
48	58,635	1	0.696	1.43 (1.0)			*
54	3,808	1	0.045	22.12(1.0)			*
69	119,575	3	1.419	2.11 (1.0)			*
77	177,742	5	2.109	2.37 (1.0)			*
78	125,127	2	1.485	1.34 (1.0)			*
90	194,866	5	2.312	2.16 (1.0)			*
110	21,535	1	0.256	3.91 (1.0)			*
							LLR = 31.8
							$\hat{\theta}$ = 2.41

3) Marginal power P(+, s) and its conditional marginal power P(+, s)/P(+, +)

Regarding the marginal power $P(+, s^*)$ at $s = s^*$, the flexible spatial scan statistic is shown to have much higher power than the circular spatial scan statistic for the case of noncircular clusters (Tables 3, 4, 5). Furthermore, the conditional marginal power $P(+, s)/P(+, +)$ of the flexible spatial scan statistic is 964/964 = 1.000, 969/979 = 0.990, 850/890 = 0.955 and 612/673 = 0.909 for the cluster A-D, respectively. These results indicate that the identified MLC by the flexible spatial scan statistic includes the hot-spot cluster with quite high probability. For the noncircular clusters, the mode of $P(+, s)$ of the circular spatial scan statistic is around $s = s^* - 1$ or $s = s^* - 2$.

4) Marginal power distribution P(l, +)

For the flexible spatial scan statistic, the probability that the length of significant MLC is less than $s = s^*$ is shown to be zero or quite small and the maximum length is around 10 to 12. the circular spatial scan statistic, on the other hand, tends to detect a much longer cluster than expected from the hot-spot cluster assumed in the simulation. For example, the probability that the length of MLC for the cluster B with length $s^* = 4$ is greater than or equal

Table 2: Comparison of the circular and the flexible spatial scan statistic for the cluster model A. Comparison of bivariate power distribution $P(l, s) \times 1000$ between the circular spatial scan statistic and the flexible spatial scan statistic for the hot-spot cluster A = {14, 15, 20}. Nominal α-level is set as 0.05 and 1000 trials are carried out. For more details, see text.

Flexible (K = 15)						Circular (K = 15)					
Length l	Include s hot-spot regions				Total	Length l	Include s hot-spot regions				Total
	0	1	2	3			0	1	2	3	
1	0	0			0	1	0	0			0
2	0	0	0		0	2	1	0	0		1
3	0	0	0	142	142	3	0	0	0	738	738
4	0	0	0	116	116	4	0	0	0	134	134
5	0	0	0	137	137	5	0	0	0	39	39
6	0	0	0	149	149	6	0	0	0	12	12
7	0	0	0	165	165	7	0	0	0	9	9
8	0	0	0	131	131	8	0	0	0	1	1
9	0	0	0	84	84	9	0	0	2	3	5
10	0	0	0	27	27	10	0	0	0	2	2
11	0	0	0	11	11	11	0	0	0	4	4
12	0	0	0	2	2	12	0	0	0	12	12
13	0	0	0	0	0	13	0	0	0	14	14
14	0	0	0	0	0	14	0	0	0	3	3
15	0	0	0	0	0	15	0	0	0	6	6
Total	0	0	0	964	964	Total	1	0	2	977	980
		usual power = 0.964						usual power = 0.980			

Table 3: Comparison of the circular and the flexible spatial scan statistic for the cluster model B. Comparison of bivariate power distribution $P(l, s) \times 1000$ between the circular spatial scan statistic and the flexible spatial scan statistic for the hot-spot cluster B = {14, 15, 20, 26}. Nominal α-level is set as 0.05 and 1000 trials are carried out. For more details, see text.

Flexible (K = 15)							Circular (K = 15)						
Length l	Include s hot-spot regions					Total	Length l	Include s hot-spot regions					Total
	0	1	2	3	4			0	1	2	3	4	
1	0	0				0	1	0	0				0
2	0	0	0			0	2	0	0	0			0
3	0	0	0	0		0	3	0	0	0	523		523
4	0	0	0	0	127	127	4	0	0	0	65	0	65
5	1	0	0	0	157	158	5	0	0	0	23	0	23
6	0	0	0	0	205	205	6	0	0	0	7	66	73
7	0	0	0	2	198	200	7	0	0	0	0	15	15
8	0	0	0	1	151	152	8	0	0	0	0	32	32
9	0	0	0	5	85	90	9	0	0	0	1	15	16
10	0	0	0	1	24	25	10	0	0	0	0	7	7
11	0	0	0	0	17	17	11	0	0	0	2	3	5
12	0	0	0	0	5	5	12	0	0	0	2	63	65
13	0	0	0	0	0	0	13	0	0	0	0	96	96
14	0	0	0	0	0	0	14	0	0	0	0	30	30
15	0	0	0	0	0	0	15	0	0	0	0	22	22
Total	1	0	0	9	969	979	Total	0	0	0	623	349	972
			usual power = 0.979							usual power = 0.972			

Table 4: Comparison of the circular and the flexible spatial scan statistic for the cluster model C. Comparison of bivariate power distribution $P(l, s) \times 1000$ between the circular spatial scan statistic and the flexible spatial scan statistic for the hot-spot cluster C = {14, 15, 26, 27}. Nominal α-level is set as 0.05 and 1000 trials are carried out. For more details, see text.

Flexible ($K = 15$)							Circular ($K = 15$)						
Length l	Include s hot-spot regions					Total	Length l	Include s hot-spot regions					Total
	0	1	2	3	4			0	1	2	3	4	
1	0	0				0	1	1	0				1
2	0	0	0			0	2	0	0	351			351
3	0	0	0	0		0	3	2	0	4	0		6
4	0	0	0	0	138	138	4	0	0	3	0	0	3
5	0	0	0	3	147	150	5	2	0	2	0	0	4
6	1	0	0	2	200	203	6	1	0	0	0	0	1
7	0	1	0	4	147	152	7	0	0	0	81	0	81
8	0	0	2	9	107	118	8	0	0	10	18	38	66
9	0	0	0	10	71	81	9	0	0	2	0	26	28
10	1	0	2	5	28	36	10	0	0	0	29	3	32
11	0	0	0	0	10	10	11	0	0	1	13	1	15
12	0	0	0	0	2	2	12	0	0	2	4	60	66
13	0	0	0	0	0	0	13	0	0	0	5	62	67
14	0	0	0	0	0	0	14	0	0	0	10	27	37
15	0	0	0	0	0	0	15	0	0	0	6	37	43
Total	2	1	4	33	850	890	Total	6	0	375	166	254	801
		usual power = 0.890							usual power = 0.801				

Table 5: Comparison of the circular and the flexible spatial scan statistic for the cluster model D. Comparison of bivariate power distribution $P(l, s) \times 1000$ between the circular spatial scan statistic and the flexible spatial scan statistic for the hot-spot cluster D = {73, 74, 75, 76, 78}. Nominal α-level is set as 0.05 and 1000 trials are carried out. For more details, see text.

Flexible ($K = 15$)								Circular ($K = 15$)							
Length l	Include s hot-spot regions						Total	Length l	Include s hot-spot regions						Total
	0	1	2	3	4	5			0	1	2	3	4	5	
1	0	0					0	1	6	0					6
2	1	0	0				1	2	3	5	0				8
3	0	0	0	0			0	3	0	0	0	14			14
4	1	0	0	1	0		2	4	1	0	4	5	0		10
5	0	1	0	3	1	242	247	5	0	0	2	1	0	0	3
6	1	0	0	1	2	162	166	6	1	0	0	1	363	0	365
7	2	3	0	5	5	93	108	7	0	0	1	0	56	0	57
8	1	2	1	6	7	53	70	8	0	0	2	2	28	0	32
9	0	2	0	1	5	38	46	9	0	0	2	2	10	0	14
10	0	2	0	1	1	18	22	10	1	0	0	3	3	0	7
11	0	0	0	2	2	5	9	11	0	0	0	0	3	11	14
12	0	0	1	0	0	1	2	12	0	0	0	2	3	8	13
13	0	0	0	0	0	0	0	13	0	0	0	1	1	16	18
14	0	0	0	0	0	0	0	14	0	0	1	0	0	5	6
15	0	0	0	0	0	0	0	15	0	1	0	0	1	7	9
Total	6	10	2	20	23	612	673	Total	12	6	12	31	468	47	576
			usual power = 0.673								usual power = 0.576				

Table 6: Cost comparison Expected number of undetected regions included in the true cluster $E(s^* - S)$, expected number of detected regions not in the true cluster $E(L - S)$ and the ratio of costs C/C_2 ($r = 1, 2$) incurred by incomplete identification of the true cluster. The spatial scan statistic with low values is better.

Hot-spot Cluster	Scan statistic	$E(s^* - S)$	$E(L - S)$	the ratio C/C_2	
				$r = 1$	$r = 2$
A = {14, 15, 20}	Flexible (K = 15)	0.108	2.951	3.059	3.167
	Circular (K = 15)	0.065	0.722	0.787	0.852
B = {14, 15, 20, 26}	Flexible (K = 15)	0.097	2.548	2.645	2.742
	Circular (K = 15)	0.735	2.525	3.260	3.995
C = {14, 15, 26, 27}	Flexible (K = 15)	0.492	2.243	2.735	3.227
	Circular (K = 15)	1.736	3.153	4.889	6.625
D = {73, 74, 75, 76, 78}	Flexible (K = 15)	1.774	1.088	2.862	4.636
	Circular (K = 15)	2.770	1.709	4.479	7.249

to 12 is 213/1000 = 0.213 compared with 5/1000 = 0.005 for the flexible spatial scan statistic. The probability that the length of MLC for the cluster C with length $s^* = 4$ is greater than or equal to 12 is 213/1000 = 0.213 compared with 2/1000 = 0.002 for the flexible spatial scan statistic. This tendency is shown even in the circular cluster A where the same probabilities are 0.035 vs. 0.002.

Cost comparison

Based upon the bivariate power function $P(l, s)$, we can compute the following expected total cost incurred by incomplete identification of the true cluster:

$$C = C_2\{rE(s^* - S) + E(L - S)\}, r = C_1/C_2 \quad (8)$$

where C_1 and C_2 denote the average cost of missing one region in the true cluster and that of incorrectly detecting one region not in the true cluster, respectively. L and S denote the random variable of l and s, respectively. Two expected numbers $E(s^* - S)$ and $E(L - S)$ for four kinds of clusters A-D are shown in Table 6. In general, we can assume $r > 1$. For example, the ratio C/C_2 is shown for the case of $r = 1$ and $r = 2$, respectively, in Table 6. However, in this example, irrespective of the value of $r(> 1)$, the circular spatial scan statistic is shown to have lower cost for detecting circular cluster A but to have higher cost for detecting non-circular clusters B-D.

Limitations of current work

Needless to say, the results derived here are based upon a small Monte Carlo simulations study and thus the characteristic observed in the current work could change a little bit depending on the cluster model adopted. We assumed here only one hot spot cluster and did not consider the case of several hot spot clusters. Therefore, we need a further simulation study to compare the performance of the two spatial scan statistics under several different clusters.

Regarding the algorithm adopted for the flexible spatial scan statistic, we set the restriction that irregularly shaped windows Z with length $k(\leq K)$ are constructed from members of the $(K - 1)$-nearest neighbours to the starting region. It seems that this restriction plays an important role in preventing the flexible spatial scan statistic from reaching out for and absorbing faraway regions with non-elevated risk. However, to avoid undertaking computationally infeasible searches, the flexible spatial scan statistic has to be set with an upperbound for K. This depends on the disease map under study and the capability of the computer. The current practical upperbound is around $K = 30$ for the reason that the execution time of our current algorithm will take more than a week if $K > 30$ for the number of regions $m = 200 \sim 300$. However, it seems to be unlikely that the length of the true cluster would be larger than $10 \sim 15$ percent of the total number of regions. So, we think that our current algorithm can be applied to many epidemiological studies with small to moderate cluster sizes. However, for larger cluster sizes, a more sophisticated algorithm to increase the upperbound for K is needed.

Regarding data type, the proposed spatial scan statistic can only be applied to regional count data whereas the circular spatial scan statistic can be applied to not only count data but also individual point data. However, at least in disease surveillance, most of the data that people

analyze is aggregated, so the method covers most real-world situations.

Finally, one of the reviewers commented that using small areas as basis for clustering without any attempt to incorporate heterogeneity in background rates is a fundamental flaw of all existing scanning methods. In general, we know that disease risks over study regions are heterogeneous to a certain extent and the null hypothesis of complete spatial randomness is not true. However, statistical hypothesis testing is based upon the null hypothesis which is not true. Likewise, we will use complete spatial randomness as the null hypothesis as indicated in equation (1) since we are interested in rejecting the null hypothesis and detecting the local clusters with excess risk. If we are interested in estimating a clustering mechanism, we should use some modeling approach rather than spatial scan statistics.

Discussion

In this paper, we proposed a flexibly shaped spatial scan statistic to detect arbitrarily shaped clusters by amalgamating administrative units. The flexible spatial scan statistic is, via Monte Carlo simulation, shown to have reasonably high powers compared with the circular spatial scan statistic when examined by a newly introduced bivariate power distribution $P(l, s)$. The simulation reveals that the circular spatial scan statistics shows a high level of accuracy in detecting circular clusters exactly and reasonably good power for including some hot-spot regions into the most likely cluster. The flexible spatial scan statistic exhibits no such high power regarding exact identification of clusters but the support of the power distribution is shown to be concentrated in a relatively narrow range of length l on the line $s = s^*$, indicating that an observed significant most likely cluster contains the true cluster with quite high probability. The circular spatial scan statistic, on the other hand, is shown to have zero powers for detecting exactly noncircular clusters that cannot be captured by any circular window. The circular spatial scan statistic is also shown to have a tendency to detect a larger cluster than the true cluster assumed in the simulation even for the case when the true cluster is circular. Furthermore, by introducing the two kinds of cost due to incomplete detection of the true cluster, we could summarize these characteristics in terms of minimizing expected total cost. One of the reviewers suggested a similar cost comparison using the number of people that are incorrectly classified rather than the number of regions since the cost of misclassifying a large region is at least for disease surveillance purposes higher than that of misclassifying a region with smaller population. We think that would be an interesting additional simulation study worth conducting. However, since it can be expected that the result of such a cost comparison strongly depends on the spatial configuration of regions with different population size in the neighborhood of and within the true cluster and thus it requires careful design for creating suitable cluster models from which we can intuitively infer the result to a certain extent, we would like to leave such a simulation study in our future work.

The surprising result that Duczmal and Assunção's scan statistic detected quite large and unlikely peculiar shaped clusters that had the largest likelihood ratio among the three scan statistics might cast a doubt on the validity of the model selection based upon maximizing the likelihood ratio (5). Such a doubt can also be seen in some simulation results of the circular spatial scan statistic that had non-negligible probabilities of detecting much longer clusters than the true cluster. The flexible spatial scan statistic, on the other hand, is shown not to detect such an unexpected long cluster probably because it has the restriction that our windows are constructed only from members of the $(K - 1)$-nearest neighbours to the starting region. Nevertheless, these undesirable properties produced by maximum likelihood ratio might suggest the use of a different criterion for model selection. For example, we might consider a penalized likelihood where we consider a penalty for the *complexity of the cluster shape*, which is also worth future research.

All the computations and simulations have been conducted on a PC with Windows XP. For users who are interested in applying the flexible spatial scan statistic, we can provide the software FleXScan [20].

Conclusion

The circular spatial scan statistics shows a high level of accuracy in detecting circular clusters exactly and reasonably good power for including some hot-spot regions into the most likely cluster. The proposed flexible spatial scan statistic is shown to have good usual powers plus the ability to detect the noncircular hot-spot clusters more accurately than the circular spatial scan statistic. However, the proposed spatial scan statistic work well for small to moderate cluster size, say up to 30. For larger cluster sizes, the method is not practically feasible and a more efficient algorithm is needed.

Appendix: algorithm to find the set Z_2 defined in equation (3)

There are probably several procedures to find the set Z_2 that is defined as the set of arbitrarily shaped windows Z within a pre-specified maximum length K. The algorithm that we used is described as follows:

Step 1. First we set an $m \times m$ matrix $A = (a_{ij})$ such as

$$a_{ij} = \begin{cases} 1 & \text{(regions } i \text{ and } j \text{ are connected)} \\ 0 & \text{(otherwise),} \end{cases}$$

and set $Z_2 = \phi$ and $i_0 = 0$

Step 2. Let $i_0 \leftarrow i_0 + 1$ and $i_0 (= 1, 2, \ldots, m)$ be the starting region. Then we create the set W_{i_0} consisting of $(K - 1)$-nearest neighbours to the starting region i_0 and i_0 itself, i.e.,

$$W_{i_0} = \{i_0, i_1, i_2, \ldots, i_{K-1}\},$$

where i_k is the k-th nearest to i_0.

Step 3. We consider all the set $Z \subset W_{i_0}$, which includes the starting region i_0. For any given such set Z, repeat the following steps 4–7.

Step 4. We divide the set Z into two disjoint sets: $Z_0 = \{i_0\}$ and Z_1 which contains the other regions of Z.

Step 5. We make two new sets $\mathbf{Z}_0'$ and $\mathbf{Z}_1'$. $\mathbf{Z}_0'$ consists of the regions of Z_1 that are connected to some regions of Z_0. On the other hand, $\mathbf{Z}_1'$ consists of the regions of Z_1 that are not connected to any regions of Z_0. Then we replace Z_0 and Z_1 by $\mathbf{Z}_0'$ and $\mathbf{Z}_1'$, respectively.

Step 6. We repeat the step 5 recursively until either Z_0 or Z_1 becomes null first.

Step 7. We make a decision as follows. Z is said to be "connected" when Z_1 becomes null first and "disconnected" when Z_0 becomes null first. If we can find Z "connected", Z is added to the set Z_2. If we find Z "disconnected", Z is discarded.

Step 8. Repeat the steps 2–7 until we finally get the set Z_2 consisting of arbitrarily shaped windows Z whose maximum length is K.

Now we shall give an example using regions in the Tokyo Metropolitan area shown in Figure 1. Let the starting region $i_0 = 14$. Then, the regions in the set of $(K - 1)$-nearest neighbours to the region 14 are listed as follows in the ascending order of distance to the region 14, i.e.,

$$W_{14} = \{14, 15, 20, 4, 16, 13, 19, 12, 5, \ldots\}.$$

Suppose that we take a subset $Z = \{14, 15, 20, 26\}$. In the first step, we have

$$Z_0 = \{14\}, Z_1 = \{15, 20, 26\}.$$

Since $a_{14,15} = a_{14,20} = 1$ and $a_{14,26} = 0$, we then have

$$Z_0 = \{15, 20\}, Z_1 = \{26\}.$$

Further, because $a_{15,26} = a_{20,26} = 1$, these sets are replaced by

$$Z_0 = \{26\}, Z_1 = \phi.$$

So, we can find that the set $Z = \{14, 15, 20, 26\}$ is "connected" and can be a member of Z_2.

If we take a subset $Z = \{14, 15, 20, 5\}$, we can find Z is "disconnected" because $a_{14,5} = a_{15,5} = a_{20,5} = 0$, $Z_0 = \phi$ and $Z_1 = \{5\}$ at the final stage.

Authors' contributions

TT proposed the flexibly shaped spatial scan statistic and the bivariate power distribution. KT considered the algorithm given in the appendix, programmed the C++ code and carried out the power simulations. TT wrote the first draft of the manuscript. Both authors interpreted the results and wrote the final version of the paper.

Acknowledgements

This research was supported in part by Grant-in-Aid for Scientific Research (Grant No. 16300091) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Marshall RJ: **A review of the statistical analysis of spatial patterns of disease.** *Journal of Royal Statistical Society, Series A* 1991, **154**:421-441.
2. Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R, (Eds): *Disease Mapping and Risk Assessment for Public Health* London: John Wiley & Sons; 1999.
3. Lawson A, Denison D: *Spatial Cluster Modelling* Boca Raton: CRC Press; 2002.
4. Waller LA, Gotway CA: *Applied Spatial Statistics for Public Health Data* New York: John Wiley & Sons; 2004.
5. Cuzick J, Edwards R: **Spatial clustering for inhomogeneous populations (with discussion).** *Journal of the Royal Statistical Society, Series B* 1990, **52**:73-104.
6. Besag J, Newell J: **The detection of clusters in rare diseases.** *Journal of the Royal Statistical Society, Series A* 1991, **154**:143-155.
7. Kulldorff M, Nagarwalla N: **Spatial disease clusters: detextion and inference.** *Statistics in Medicine* 1995, **14**:799-810.
8. Kulldorff M: **A spatial scan statistic.** *Communications in Statistics* 1997, **26**:1481-1496.
9. Tango T: **A class of tests for detecting 'general' and 'focused' clustering of rare diseases.** *Statistics in Medicine* 1995, **14**:2323-2334.
10. Tango T: **A test for spatial disease clustering adjusted for multiple testing.** *Statistics in Medicine* 2000, **19**:191-204.
11. Viel JF, Arveux P, Baverel J, Cahn JY: **Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters and a municipal solid waste incinerator with high dioxin emission levels.** *American Journal of Epidemiology* 2000, **152**:13-19.
12. Sankoh OA, Ye Y, Sauerborn R, Muller O, Becher H: **Clustering of childhood mortality in rural Burkina Faso.** *International Journal of Epidemiology* 2001, **30**:485-492.
13. Perez-AM, Ward MP, Torres P, Ritacco V: **Use of spatial scan statistics and monitoring data to identifyclustering of bovine**

tuberculosis in Argentina. *Preventive Veterinary Medicine* 2002, **56**:63-74.
14. Kulldorff M, Tango T, Park PJ: **Power comparisons for disease clustering tests.** *Computational Statistics and Data Analysis* 2003, **42**:665-684.
15. Song C, Kulldorff M: **Power evaluation of disease clustering tests.** *International Journal of Health Geographics* 2003, **2(9)**:1-8.
16. Kulldorff M, Information Management Services Inc: **SaTScan v4.0: Software for the spatial and space-time scan statistics.** 2004 [http://www.satscan.org/].
17. Patil GP, Taillie C: **Upper level set scan statistic for detecting arbitrarily shaped hotspots.** *Environmental and Ecological Statistics* 2004, **11**:183-197.
18. Duczmal L, Assunção R: **A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters.** *Computational Statistics & Data Analysis* 2004, **45**:269-286.
19. Dwass M: **Modified randomization test for nonparametric hypotheses.** *Annals of Mathematical Statistics* 1957, **28**:181-187.
20. Takahashi K, Yokoyama T, Tango T: **FleXScan: Software for the flexible spatial scan statistic.** National Institute of Public Health, Japan; 2004.

環境分析のためのGISの現状と展望

浅見泰司

２００５
『環境管理』Vol ４１, No. ８

環境分析のための GIS の現状と展望

浅見泰司
東京大学空間情報科学研究センター
副センター長/教授

地理情報システム（GIS）は環境分析の重要なツールになりつつある。ただし，利用の際には，空間データとして，対象とする環境現象に関して，適切なスケール，属性を選択する必要がある。今後，形式相互の適切な変換が自動的にできる技術開発，社会的に環境関連データが共有化されるようにすること，様々な誤差や曖昧性の空間操作による伝播の自動計算の技術開発，対話的な処理をより容易にしていく開発などが求められる。今後，GIS の利用によって，より的確な分析に基づく環境対策が行われることを期待する。

はじめに

環境 GIS とは狭義には，主として自然，大気，水，騒音など環境に関する情報を GIS を用いてわかりやすく表示したものをさし，例えば，環境省，国立環境研究所などがモニターした環境状態を公表している[*1]。しかし，より広義に考えれば，我々の身の回りに関する様々な環境について空間的に表示，解析，予測などができるシステムをさす。本稿では，必ずしも狭義の意味にとらわれず，広く解釈して論を進める。

地理情報システム（GIS）は，空間的な情報を操作し，地図化して表示することを目的としたソフトウェアである。環境問題においては現状や将来値についての空間的な影響に大きな関心があるため，GIS は，現在では環境分析を行うに当たって重要なツールになりつつある。

環境分野において，比較的よく使われているのは GIS の表示機能であろう。対象とする環境の空間的な状況を地図上に表示することで，環境の悪いところを一目で理解したり，特定の場所での活動に対して注意を促したり，また，環境対策のヒントを得ることができる。例えば，大気汚染の場合には，汚染状況が悪い地域を把握して要対策地域をしぼったり，汚染原因の把握に役立てたりできる。

GIS はまた，解析や予測にも利用できる。ただし，通常，市販されている汎用的な GIS は一般的な空間情報の操作に対応できるよう，基礎的な操作機能を持っているだけである。付加的な汎用機能は，徐々に拡大されてきているが，特定現象のモデル化をするには，いまだ完全ではない。そのため対象とする現象によっては必要な機能を付加することが必要になる。ただ，そのような機能を付加することで極めて有力なツールとなる。

そもそも地理情報と環境分野のつながりは古くからある。1854 年のロンドンにおけるコレラ流行が空気伝染ではなく，井戸水の汚染によるものであることを示したのは，John Snow である。それを示すために彼はコレラで死亡した患者の家の位置と公共井戸ポンプの地理的な分布を示す地図を示し，明らかにポンプの近隣で死亡者が多いことを示した[8]。これが疫学の

*1　例えば，生物多様性センターの生物多様性情報システム（http://www.biodic.go.jp/kiso/map/survey_map_f.html），国立環境研究所の環境 GIS（http://www-gis.nies.go.jp/）参照。

祖となっている。

本稿では，今後環境分野でますます活用されると思われるGISの特性を概観し，環境問題に適用するときの課題や展望を述べてみたい。

1　GISのデータと表示

GISでは，地理上の事物や状態を数値データ（あるいは文字列や記号列）で記述する。また，データは有限でなければならない。そのために，デジタル情報にできるものや，関心の高いものに限定してデータ化することが避けられない。このことは，対象とする環境現象に関して，適切なスケール，属性を選択しなければならないことを意味する。

空間データは，ラスターデータとベクターデータに大別される。ラスターデータとは，規則的に並んだ格子点（ないし，メッシュ）に属性データを重ねたデータであり，点の属性を表すことも，一定の範囲の属性を表すこともある。例えば，衛星写真は，ラスターデータの典型であるが，これは厳密にいうと，格子点のデータではなく，それを中心とするカメラの測定範囲の重み付けの平均的なデータとなっている。そのため，ラスターデータは，たとえ位置が同じであっても，同じ領域を表現したものであるとは限らないという意外な問題点がある。一方，ベクターデータとは，座標点ないしその集合で位置や形状を記述するもので，図形は折れ線によって記述される。このため，形状はあくまで近似されたものであり，データソースに応じて，その精度はまちまちである。環境関連データを表すときに，重ね合わせ（オーバーレイ）て表現することが多いが，このように空間精度や空間的な代表性の異なるデータを一つの図面にしてしまうときには，その表示や解釈において注意が必要である。

また，地図では，小縮尺の地図に本当の精度で表示するとかえって見づらいことがあり，ある程度簡単な形に省略して描画すること（generalization）が多い。例えば，細街路を省略したり，地区をまとめてしまったり，記号を大きめに描いてしまうなどということはしばしばである。GISでは拡大縮小が自由にできるが，他方で，データ自体は特定のスケールでの利用を前提にしていることも多く，この点も重ね合わせで注意しなければならない。

GISデータを扱う上で，時間の扱いも重要である。地理的なデータは，あくまである時点でのデータであり，その後の状況の変化は反映されていない。時に，データの古さが決定的な事故につながることもあるだけに，その扱いは慎重であるべきである。

マーク・モンモニア[9]は地図の表示方法の技術を使って，かなり都合のよい地図を作れてしまうことを指摘し，その具体的な手法を解説している。環境問題はかなり政治問題化しやすい対象であるだけに，この点についても十分に注意しなければならない。

2　空間解析・モデル

GISを表示するだけでなく，環境の分析に用いることで，空間的な関係を解明できることが期待される。汎用的なGISで使える空間操作機能は，空間データの蓄積・表示機能，重ね合わせ機能（オーバーレイ），空間オブジェクトの共通集合などの代数的処理，空間オブジェクトのカウント・面積・周長などの算出，距離計測，バッファリング，等値線作成，三次元地表面の作成・斜度などの計測，可視領域の算出などであり，環境解析で必要な機能としてはかなり限られている。

例えば，環境影響範囲を求めたいというときに，一般的には距離によって減衰する現象が多い。そのため，GIS機能を用いて，影響評価を行う場合には，影響領域をバッファリングと呼ばれる操作で求めることが多い。バッファリングとは，ある空間オブジェクトから一定距離だけ広げた範囲を求める操作である。確かに，基本的な状況を知るには，バッファリングが役に立つ。しかし，実際には距離以外の要因でも影響領域が変わる。例えば卓越風により特定の方向に影響が広がりやすかったり，地形によって影響範囲が異なる場合には，単純なバッファリングという操作は適切とはいえない。ところが，GISを用いて環境モデリングを行う際には，なるべく汎用的な機能を有効利用しがちなので，

本来の影響領域を求めるという目的からははずれてしまいかねない。Albrecht[1]は環境モデリングでGIS機能だけに限定して，どの程度まで利用可能かを調べているが，概して分析面では弱いことを述べている。

環境を分析する上で，特定のサンプリングポイントにおける環境値が計測されていて，それを，面的に広げて検討したいということは多々ある。これには，三次元的な曲面（サーフェス）の標高点が有限個与えられたときのサーフェスの推定の手法が役に立つ（あるいは，有限個のデータ点から，他の点の値を推定する問題と考えてもよい）。このための空間解析の手法としては，クリギング（kriging）と呼ばれる手法がある。クリギングにおいては，近くの値は余り変わらず，遠くになるほど値が違う可能性があるという空間的な性質を用いて，なるべく近くにあるサンプル点との距離によって，重み付けをして新たな点の環境値を推定するものである[2]。ただし，この手法は，他の内挿的な手法と同様，サンプル点以外に特異点があっても，それは再現できないし，また，非連続な変化がある場合には適切な手法とはならなくなる。そのため，市街地における騒音レベルのように，空間的に短距離で減衰してしまうような現象については，適切なサンプリング密度を確保しないと意味ある分析ができない可能性がある。

一定の確率で現象が起きていることを発見したサンプルポイントを得たときに，もとの現象生起確率分布を求めたいというときには，密度推定手法が役に立つ。よく用いられる方法としては，カーネル（kernel）関数法[14]がある。得られたサンプル点に，一定の密度分布を割り当てていき，順次空間的に加えていくことで，サンプル点が集まっているところは密度が高く，逆にほとんどないときには密度が小さく確率密度を推定することができる。カーネル関数としては，等方的な関数（方向によって，密度の多寡がない）を仮定されることが多く，環境現象によっては，工夫が必要な可能性もある。

二つの事象が生起したことを示す点位置データがあるときに，二つの現象に関連があるかどうかを調べたいことがある。このための方法としては，クロスK関数法（ないし，クロスL関数法）がある[2]。この方法では二つの事象が関連しているならば，より近くなる傾向があるか，逆により離れる傾向があるということを利用している。一つの事象が起きた点から他方の事象が起きた点までの最も近い距離を計算して，それが，ランダムな場合と比べて有意に小さければ近接傾向があると判断する。例えば，環境汚染が疑われる施設が複数あり，それと環境影響を受けた生体の位置が既知であった場合に，それらが有意に近接しているとすれば，影響がある可能性が高いことになる。この方法では，因果関係は直接的にはわからないし，見かけの相関のような現象（例えば，施設は駅に近接し，人口も駅を中心に分布しているなど）も大いにありうるので，直ちに関係があるとはいえない。しかし，上で紹介したコレラの例のように，影響が疑われる施設の種類を絞り込むには有効な方法である。

実際の環境解析で頻繁に用いられるのは，メッシュデータである。これは，空間を有限な要素に分解することで，記述すべき変数の量をコントロールできるだけでなく，隣接関係が比較的簡単なため，モデル自体を単純に記述できるためである。しかし，空間データをすべてメッシュデータに合わせることで，位置同定の精度，空間的偏在性の記述力の限界などの問題が生じる。また，せっかくの原データにおける情報量をメッシュデータ化することで減らしてしまう。

例えば，地理上のデータをメッシュデータ化する際には，対象となる地区なり，空間オブジェクトが複数のメッシュにわたることがあるにもかかわらず，対象の代表点をとってそれが含まれるメッシュにすべてを帰属させてしまうことが多い。その際には，メッシュが小さくなるほど，この手法による精度は粗くなってしまう。仮に，対象を複数のメッシュに分割するとしても，単純に面積按分するのがよいのかは疑問が残る。特に空間的に偏在しているような環境現象の場合には，単純な面積按分は問題は矮小化しかねない。このように空間単位を変換することで，位置同定の問題が必ず発生してしまうの

である。

メッシュデータを用いると，ポイント，ライン，ポリゴン，ラスターで得られる当初の空間データをすべてメッシュデータに変換する手間が必要となる。例えば，ポイント施設や道路などライン施設からメッシュの代表点までの距離を求めたり，ポリゴンデータで得られている土地利用分類の面積構成をメッシュデータに変換するという操作が頻繁に行われる。このような変換は，試行錯誤において施設位置を変えると，すべて再計算して変換作業を行うなど，データ構築の作業に余計な負担がかかることとなる。このため，GIS側でデータ変換を自動化したり，環境モデルにおいてGIS機能を加味したシステムを構築することが望ましい。

3 空間モデリングとGIS

GISが有益なツールになるためには，環境モデルとの組み合わせが重要である。環境モデルにおいて，騒音の拡散や交通発生予測など，それぞれの対象現象特有の解析を行い，それをGISとリンクすることで，大きな威力を発揮する（例えば，参考文献3～6，10，11，15，16など。）

Rodriguez－Bachiller[12]によれば，影響評価関連文献の6％でしかGISは使われていないという。この大きな原因として以下を指摘できるだろう。1）GIS自体にそなわった機能は限られており，環境分析で必要な機能と大きくかけ離れている。2）環境評価を行う者は，開発業者であったり，環境関連行政体であったりするが，これらの間でのデータ面でのやりとりなどがなく，全体として，環境データが限られた中でしか評価作業ができない。また，仮にGISが使われているとしても，マッピング（図面表示）のみに機能を限定していることも多い。3）GISと環境モデリングソフトウェアとのインターフェースが十分に整備されていない。

以前は，汎用GISしかなかったために，環境モデリングは，現実にはGISとは関係なく行い，最後の表示のみGISを用いることが少なくなかった。そのために，環境モデルソフトとGISとの間のデータの互換性が課題となっていた。実際，沢野[13]も指摘するように，このデータ共有の問題は大きな課題ではある。ただ，最近では，この状況も徐々に改善されつつある。例えば，ESRI社のArcGIS 9.0 ModelBuilderというソフトウェアでは，モデリングのためのインターフェースが用意されている。他のソフトウェアから特定の機能を呼び出すことができるコンピュータの環境が用意されてきているため，今後，モデル構築はかなり進むと期待される。

4 環境モデルの不確実性

環境アセスメントなどで重要なのは，ある事業が個々の環境要素にどのような影響を与えるかを予測することである。ところが，環境モデル自体においても，精度をあげることが困難な要因が存在する。

環境モデルとしては，電波障害や汚染物質の拡散のようにかなり物理的法則を用いて記述できるモデルや，化学物質の反応のような化学的なモデルで記述できるものもある。それらのモデルにおいては，気象が関与するなど，自然現象が介在したり，現象自体に複雑系的な要素を含むなど，モデル自体の精度がある程度以上はあがり得ないものもある。また，そうではないとしても，空間に対して高次元の変数が必要となり，計算の複雑性によって，実質的に精度の高い計算が難しい現象も存在する。

さらに，生息圏や交通モデルのように動物や人間の行動も含む場合には，モデルに行動反応モデルを含めねばならない。動物や人間の行動が必ずしも画一的に予測できないのであるから，モデル精度にも自ずと限界があることとなる。

Huang and Chang[7]は環境モデリングに伴う不確実性の理由を三つあげている。第一に，環境現象が複雑なことである。環境現象の中には要因や非線形な関係性など，数学的に表現することが難しいものがある。第二に，環境情報の中には取得できないものがある。その場合には，ラフな推定値を当てはめたり，その要素を無視せざるを得ない。そのため，環境モデルへの入力情報は本来あるべき情報量に比べて小さくな

りがちである。第三に，計量できる変数についても，多くの場合は確率的な変数となりがちである。このためには，ファジー理論や確率的モデルを導入しなければならないこともある。

不確実性をGISで対処するためには，空間データ自体に精度に関する情報を内包させ，その情報に基づいて空間処理がなされるようなシステムが開発されることが望ましい。空間オブジェクト自体の精度としては，位置精度，属性値精度，属性分類精度，時間的精度など様々な精度の問題があり，それを有効に組み込んだGISは存在しない。汎用GISには精度を擬似的に表現するため，縮尺に応じて表示するかどうかを定義しているものもあるが，本来は，線自体の精度を表現するために，ベクターデータ（座標位置で記述されたデータ）であっても，線が太くなったり，ぼかされて表示されるなどの工夫があるとよい。さらに，空間操作によって，誤差がどのように伝搬するかを算出して，自動的に新たな空間オブジェクトの諸精度が計算されることが望ましい。空間形状に応じた精度の変化や複数の精度の問題がある場合に，どのように相互に影響されるかなどについては，空間解析分野での今後の研究課題である。

5　対話性とダイナミック性

環境モデリングは本来的に，試行錯誤など対話的処理が欠かせない。GISが環境アセスメントや環境計画において，有力視される大きな理由は，その対話性である。GISがないと環境影響をいくつかの環境指標値群でしか表現できなかったのが，GISの利用によって，結果をビジュアルに逐次表示でき，それをもとにさらにパラメータを変えていくなどということが可能となる。このため，環境モデルとGISとのデータ入出力の手間が簡便となれば，試行錯誤による対話的処理がより簡単にできるようになる。

対話的な処理を行うことは，現象理解や代替案の検討などにおいて重要であるが，環境モデルを律するパラメータの数が増えてしまうと，非常に難しくなる。そのため，他の評価値への影響を最小限にして特定の評価値を効果的に高めるにはどのパラメータをどのように変えるべきかを求めるような，試行錯誤自体を支援するシステムの開発も必要となる。

また，対話的処理では，作業を進めていくうちに，最初はパラメータを大きく動かし，次第に小さいパラメータの変化にしていくというような操作になりがちである。そこで，そのような操作に有効な環境モデル側での逐次的評価更新の技術が望まれる。また，場合によっては望まれる精度に応じて，モデル自体の概略的モデルから精緻なモデルへとスイッチすることも許容するシステムの構築が必要となる。

環境現象が動的な現象である以上，環境モデルにおいても時系列的なモデルの構築が望ましい。GISにおいて時系列的な状況を表示することは，離散的な各時点での空間分布図を表示するか，それを動画で表現するなどといった直接的に表現する以外は，通常のGISでの対応が難しい。しかし，時系列的にどのようになるかを一覧できるような表示方法が開発されていくことが望まれる。このためには，時系列的に（環境基準値を上回るなどの意味で）特徴的な地区を取り出して表示したり，空間位置を表示しつつそこでの時系列的変化を重ねて表示することなどができるGIS側の工夫が必要となる。

おわりに

本稿では，環境問題にGISを利用する場合の課題や展望について論考した。GISは，空間的な情報を操作し，地図化して表示することを目的としたソフトウェアで，環境分析を行うに当たっても重要なツールになりつつある。

環境データとの関連でいえば，GISで扱われるデータの性質上，対象とする環境現象に関して，適切なスケール，属性を選択しなければならない。特に，精度や代表性が異なる複数のデータを重ね合わせて表示使用というときには注意が必要である。また，地図では総描規則に応じて簡略化していることもあるため，適切なスケールに配慮してデータを扱う必要がある。

また，環境を分析する際には，環境分析を汎用GISのみで行うことは機能が限られるために難しい。そのため，分析ソフトとGISを両方呼び出しできる環境でのモデリングが必要と

なる。環境評価などでは，操作性にすぐれるために，メッシュデータにしてから分析するものが多い。そのため，ラスターデータ，ベクターデータ，メッシュデータ相互の適切な変換が自動的にできるよう技術開発されることが望ましい。環境評価において，いまだにGISが活用されている割合は低く，GISの技術的な問題のほかに，データの共有化の問題もある。この点の改善も急務であろう。

環境現象は，極めて複雑で不確実な情報を処理しなければならないこともある。GISにおいては，この面での対応も発展途上であり，様々な意味での誤差や曖昧性の空間操作による伝播の自動計算などの技術発展を今後に期待する。また，環境問題は，対話的な処理を行うことで，試行錯誤を行ったり，環境対策の代替案を検討することが可能となる。この面でのGISの期待は大きい。

今後，広義の環境GISがより広く活用されていくと思われるが，本稿で指摘した課題が解決され，より適切なツールとなって使われ，環境把握，分析，政策対応などが，より的確な分析に基づくようになることを望みたい。

参考文献

1) Albrecht JH, (Goodchild MF, et al, eds)：Universal GIS Operations for Environmental Modeling, Proceedings of the Third International Conference Workshop on Integrating GIS and Environmental Modeling (1996), National Centre for Geographic Information and Analysis, Santa Fe, New Mexico

2) Bailey TC, Gatrell AC：Interractive Spatial Data Analysis (1995), Prentice Hall, Harlow, England

3) Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, Pryl K, Van Reeuwijk H, Smallbone K, Van der Veen A：Mapping Urban Air Pollution Using GIS：A Regression-Based Approach, International Journal of Geographical Information Science (1997), 11 (7)：699～718

4) Dev DS, Venkatachalam P, Natarajan C：Geographic Information Systems for Environmental Impact Assessment (EIS)—A Case Study, International Journal of Environmental Studies (1993), 43：115～122

5) Goodchild MF, Parks BO, Steyaert LT (eds)：Environmental Modeling with GIS, Oxford University Press (1993), Oxford

6) Goodchild MF, Steyaert LT, Parks BO, Johnston C, Maidment D, Crane M, Glendinning S (eds)：GIS and Environmental Modeling：Progress and Research Issues (1996), GIS Eorld Books

7) Huang GH, Chang NB：Perspectives of Environmental Informatics and Systems Analysis, Journal of Environmental Informatics (2003), 1 (1)：1～6

8) Longley PA, Goodchild MF, Maguire DJ, Rhind DW：Geographical Information Systems and Science, 2nd ed. (2005), John Wiley & Sons, Chichester

9) マーク・モンモニア (渡辺潤，訳)：地図は嘘つきである (1995)，晶文社

10) Morris P, (Therivel R, eds)：Methods of Environmental Impact Assessment, 2nd ed., UCL Press (2000), London

11) Parker S, Cocklin C：The Use of Geographical Information Systems for Cumulative Environmental Effects Assessment, Computers, Environment and Urban Systems (1993), 17：393～407

12) Rodriguez-Bachiller A："Geographical Information Systems and Expert Systems for Impact Assessment, Part I：GIS", Journal of Environmental Assessment Policy and Management (2000), 2 (3)：369～414

13) 沢野伸浩：数値シミュレーションとGIS地形データ：公共数値地図情報の活用 (2003)，環境アセスメント学会2003年研究発表会要旨集，p. 83～86

14) Silverman BW：Density Estimation for Statistics and Data Analysis (1986), Chapman and Hall, New York,

15) Stein A, Staritsky I, Bouma J, Van Groenigen JW：Interactive GIS for Environmental Risk Assessment, International Journal of Geographical Information Systems (1995), 9 (5)：509～525

16) 高岸　且，盛田彰宏，秋島重樹，豊永竜二：戦略的環境アセスメントへのGISの有用性の考察 (2003)，環境アセスメント学会2003年研究発表会要旨集，p. 33～38

Statistics in Medicine

An extended power of cluster detection tests

Kunihiko Takahashi and Toshiro Tango

2006

Published online in Wiley InterScience

STATISTICS IN MEDICINE
Statist. Med. 2006; 25:841–852
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2419

An extended power of cluster detection tests

Kunihiko Takahashi*,† and Toshiro Tango

Department of Technology Assessment and Biostatistics, National Institute of Public Health, 2-3-6 Minami, Wako, Saitama 351-0197, Japan

SUMMARY

Several tests have been proposed to detect spatial disease clustering without prior information on their locations. In order to compare the performance of these tests, most authors employ the usual power, i.e. the rejection probability of the null hypothesis of no clustering due to various reasons. However, the usual power is not always appropriate for evaluating the cluster detection tests since their purpose is to both reject the null hypothesis and identify the cluster areas accurately. In this paper, we propose an extended power of the cluster detection tests, which includes the usual power as a special case. Further, we define the profile of the extended power, which can be expected to play an important role in the evaluation and comparison of several cluster detection tests. The proposed extended power and its profile are demonstrated by two tests—Kulldorff's circular spatial scan statistic and a flexible spatial scan statistic proposed by Tango and Takahashi. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: cluster detection; hot-spot clusters; hypothesis testing; power; spatial epidemiology

1. INTRODUCTION

Several test statistics have been proposed to detect for spatial clustering, which have been applied to a wide variety of epidemiological studies for spatial disease cluster detection [1, 2]. In particular, Kulldorff's circular spatial scan statistic [3, 4] has been used extensively along with his software SaTScan [5]. Tests for spatial randomness can be classified according to their purpose. Focused tests have been developed to detect the existence of a local cluster around a predetermined point source, while general tests search for clusters without any preconceived assumptions about their location [6]. There are two types of general tests. *Cluster detection tests* (CDTs) such as that developed by Turnbull *et al.* [7], Besag and Newell [6], Kulldorff and Nagarwalla [3], Tango [8], Duczmal and Assunção [9], Patil and Taillie [10] and Tango and Takahashi [11] are used to both detect local clusters, without any prior information on their

*Correspondence to: Kunihiko Takahashi, Department of Technology Assessment and Biostatistics, National Institute of Public Health, 2-3-6 Minami, Wako, Saitama 351-0197, Japan.
†E-mail: kunihiko@niph.go.jp

Received 23 August 2005
Accepted 1 September 2005

location, and to determine their statistical significance. On the other hand, *global clustering tests*, such as those developed by Moran [12], Whitemore *et al.* [13], Oden [14], Tango [15], Rogerson [16] and Bonetti and Pagano [17], are used to detect the presence of clusters in a study area without determining the statistical significance of individual clusters.

In order to compare the performance of these tests, the usual power has been treated in the same manner as in the usual hypothesis tests [4, 18]. In recent power comparisons of the disease clustering tests, Kulldorff's circular scan statistic was demonstrated to have the best power for detecting localized clusters [19]. However, it should be noted that the power estimates reflected the 'power to reject the null hypothesis for whatever reasons,' while the probability of both rejecting the null hypothesis and accurately identifying the true cluster is a different matter altogether. In order to compare the performance of the CDTs, Tango and Takahashi [11] proposed a bivariate power distribution classified according to the number of regions detected as the *most likely cluster* (MLC) and the number of true hot-spot regions included in the MLC. Since the bivariate power distribution contains all the information on the performance of any procedure for detecting the hot-spot clusters, any summary measures including the usual power for evaluating the CDT can be based on the power distribution.

In this paper, we propose an extended power of the CDTs, which includes the usual power as a special case, and a profile of the extended power for evaluating and comparing several CDTs. The proposed extended power and its profile will be demonstrated by two tests, namely, Kulldorff's circular spatial scan statistic [3, 4] and a flexible spatial scan statistic proposed by Tango and Takahashi [11] along with their software FleXScan [20].

2. MOTIVATING EXAMPLE

As a motivating example of our study, we present the results of the application of three CDTs—the scan statistics of Kulldorff's, Duczmal and Assunção's, and Tango and Takahashi—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). In this map, we simulated a random sample of $n=235$ cases by assuming the hot-spot cluster regions $\mathbf{C}=\{14,15,26,27\}$ whose relative risk was set to 3.0 and the cases to be Poisson distributed (see Reference [11] for details). In this application, we selected a maximum length of $K=15$ for the MLC in order to achieve a reasonably accurate comparison and a number $B=999$ for the Monte Carlo replications. The results are summarized as follows.

- Kulldorff's scan statistic detected $l=2$ regions $\{14,15\}$ as the MLC with a log likelihood ratio of 20.1 and $p=1/(999+1)=0.001$, and the estimated relative risk was $\hat{\theta}=3.47$.
- Tango and Takahashi's scan statistic detected $l=5$ regions $\{14,15,26,27,33\}$ as the MLC with a log likelihood ratio of 29.7 and $p=0.001$, and the estimated relative risk was $\hat{\theta}=3.41$.
- Duczmal and Assunção's scan statistic detected $l=15$ connected regions $\{14,15,24,26,27,31,32,33,48,54,69,77,78,90,110\}$ as the MLC with a log likelihood ratio of 31.8 and $p=0.001$, and the estimated relative risk was $\hat{\theta}=2.40$.

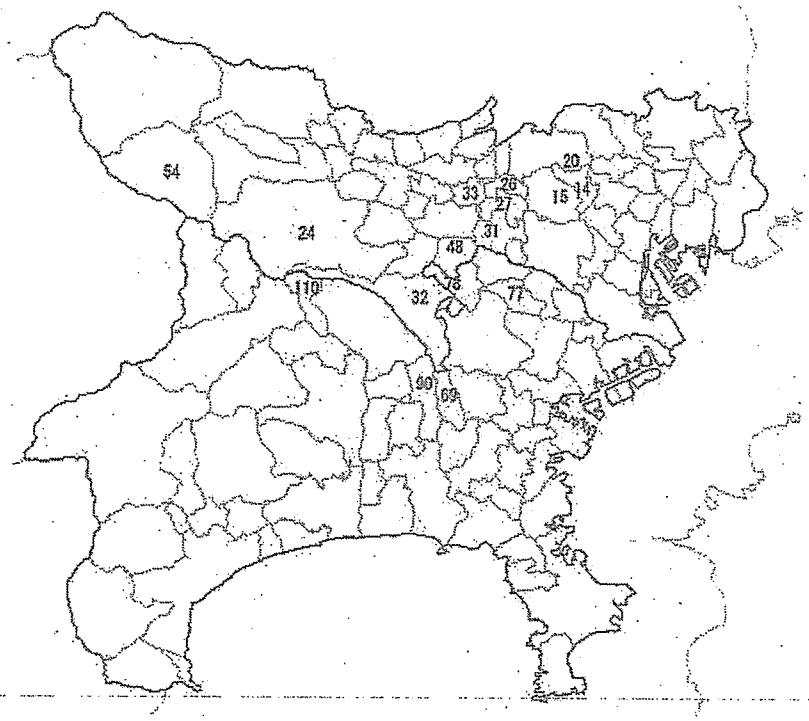

Figure 1. The 113 regions that comprise wards, cities, and villages in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan. The region number used in the text is also indicated. In particular, the region numbers of two hot-spot clusters considered in this paper are $\mathbf{A}=\{14,15,20\}$ (circular) and $\mathbf{C}=\{14,15,26,27\}$ (non-circular) (see Reference [11] for details).

All these tests rejected the null hypothesis of 'H_0: there are no clusters,' but detected different regions as MLC. In particular, Duczmal and Assunção's scan statistic detected a cluster of a peculiar shape that was considerably larger than the true cluster. This example is sufficient to show that the usual power is an inappropriate measure of the performance of a CDT. Therefore, a different measure that can reflect the extent of misclassification is required.

We consider two types of misclassifications when applying the CDT. One is a *false negative test result* (FN) in which the CDT misses a region included in the true cluster. The other is a *false positive test result* (FP) in which the CDT incorrectly detects a region that is not present in the true cluster. In the above example, Kulldorff's scan statistic has the regions $\{26,27\}$ as the FNs, but it does not have any FPs. Tango and Takahashi's scan statistic has the region $\{33\}$ as the FP, and Duczmal and Assunção's scan statistic has 11 regions $\{24,31,32,33,48,54,69,77,78,90,110\}$ as the FPs. However, both these tests do not have any FNs.

In the next section, we propose the extended power of the CDTs, which is based on the bivariate power distribution proposed by Tango and Takahashi [11] and the newly introduced penalties for the FPs and FNs.

3. EXTENDED POWER

In order to compare the performance of the CDTs, Tango and Takahashi [11] proposed a bivariate power distribution based upon Monte Carlo simulation. They introduced a bivariate power distribution $P(l,s)$ classified according to the length l of the significant MLC and the number s of the assumed hot-spot regions included therein

$$P(l,s)=\frac{\#\{\text{significant MLC has length } l \text{ and includes } s \text{ true regions}\}}{\#\{\text{trials for each simulation}\}} \tag{1}$$

where $l \geqslant 1$ and $s \geqslant 0$. Based on $P(l,s)$, we examined the following powers:

1. The usual power, i.e. $P(+,+)=\sum_{l\geqslant 1}\sum_{s\geqslant 0}P(l,s)$.
2. The joint power $P(l,s)$, especially $P(s^*,s^*)$ where s^* is the length of the hot-spot cluster assumed in the simulation.
3. The marginal power distribution of s $(\geqslant 0)$, $P(+,s)=\sum_{l\geqslant 1}P(l,s)$ and its conditional power $P(+,s)/P(+,+)$.
4. The marginal power distribution of l $(\geqslant 1)$, $P(l.+)=\sum_{s\geqslant 0}P(l,s)$.

They prepared tables of $P(l,s)$ for the following two hot-spot cluster models:

- $\mathbf{A}=\{14,15,20\}$ (circular cluster; $s^*=3$);
- $\mathbf{C}=\{14,15,26,27\}$ (non-circular cluster; $s^*=4$).

The powers are calculated for tests of nominal α levels of 0.05 and for the expected total number of cases 200 under the null hypothesis, which are based on Monte Carlo simulation using Poisson random numbers. For each simulation, 1000 trials were carried out. The resultant power distributions $P(l,s)\times 1000$ are reproduced in Tables I and II for each of the cluster models, respectively, in the form of cross table classified by l ('*length*' in tables) and s ('*include*' in tables).

Both tests have high usual powers for the circular cluster **A** (Table I), while Tango and Takahashi's scan statistic has higher power for the non-circular cluster **C** (Table II). Table I shows that Kulldorff's scan statistic detects the circular cluster **A** considerably accurately with power $P(3,3)=738/1000$, while Table II shows that it exhibits zero power $P(4,4)=0/1000$ for detecting the non-circular cluster **C** accurately. On the other hand, Tango and Takahashi's scan statistic does not exhibit such high power for identifying the clusters accurately, $P(3,3)=142/1000$ for the cluster **A** and $P(4,4)=138/1000$ for the cluster **C**. However, the power distribution appears to be concentrated in a relatively narrow range of the length l on the line $s=s^*$, thereby indicating that the observed significant MLC contains the true cluster with a considerably high probability. In particular, for the cluster **C** with length $s^*=4$, the marginal power of Tango and Takahashi's scan statistic $P(+,s^*)=850/1000$ and its conditional marginal power $P(+,s^*)/P(+,+)=850/890$ are much higher than that of Kulldorff's scan statistic, $P(+,s^*)=254/1000$ and $P(+,s^*)/P(+,+)=254/801$, respectively. Furthermore, Kulldorff's scan statistic exhibits a greater tendency to detect a larger cluster than the true cluster as compared with that of Tango and Takahashi. For example, the probability that the length of MLC for the cluster **C** ($s^*=4$) is greater than or equal to 12 is 213/1000 compared with 2/1000 for Tango and Takahashi's scan statistic. This tendency is shown even in the circular cluster **A** where the same probabilities are 35/1000 *versus* 2/1000. Therefore, the bivariate power distribution can be considered to provide very