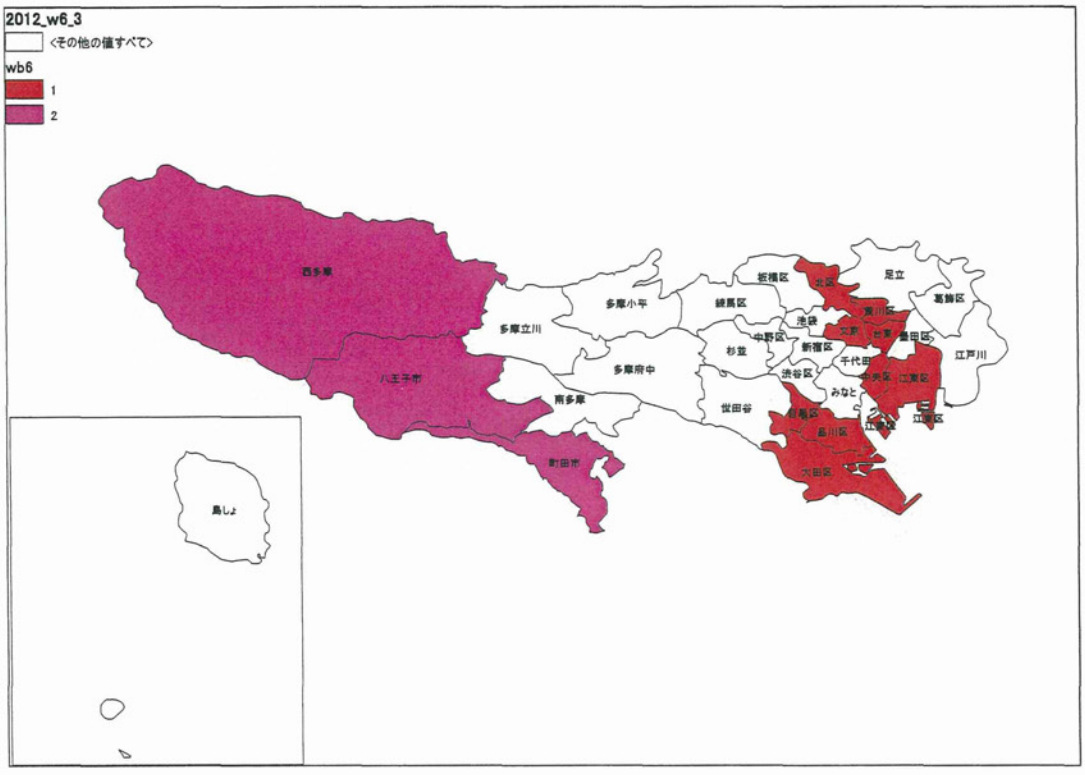
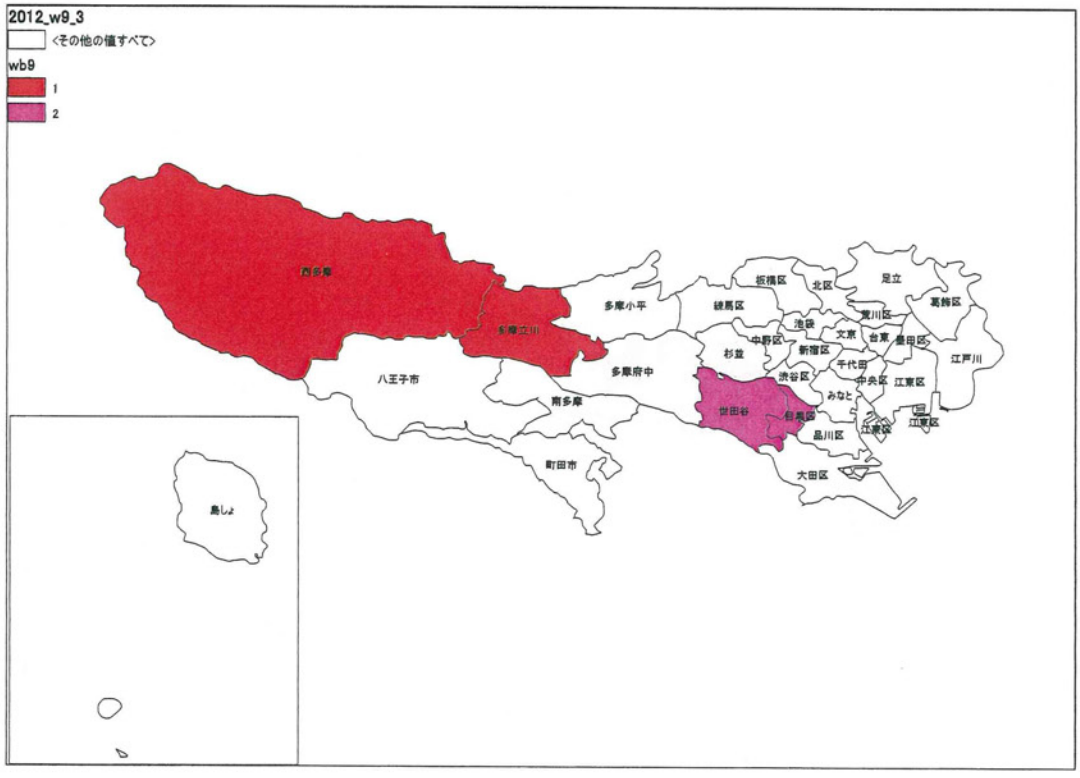
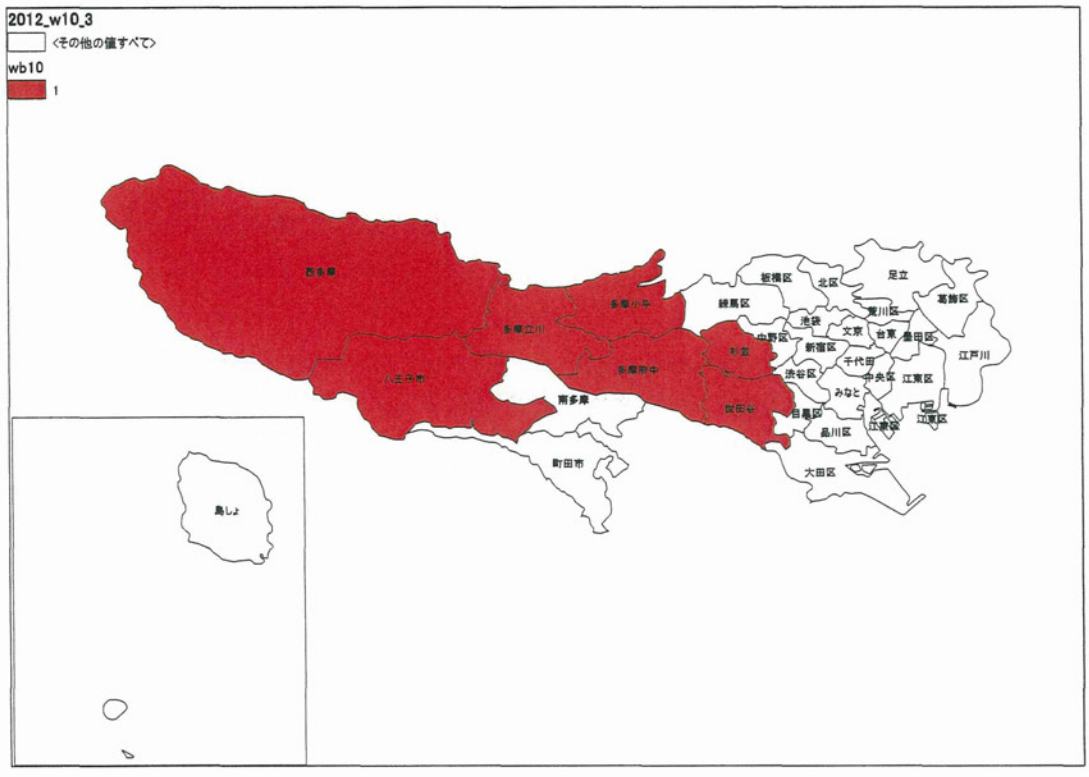

2012_w6_3
<その他の値すべて>
wb6
1
2
西多摩
多摩小平
多摩立川
八王子市
多摩府中
南多摩
町田市
島しょ
板橋区
北区
足立
練馬区
荒川区
葛飾区
池袋
文京
台東
墨田区
中野区
杉並
新宿区
千代田
江戸川
渋谷区
中央区
江東区
みなと
世田谷
目黒区
品川区
大田区


2012_w7_3
<その他の値すべて>
wb7
1
2
3
4
西多摩
多摩小平
多摩立川
八王子市
多摩府中
南多摩
町田市
島しょ
板橋区
北区
足立
練馬区
荒川区
葛飾区
池袋
文京
台東
墨田区
中野区
杉並
新宿区
千代田
江戸川
渋谷区
中央区
江東区
みなと
世田谷
目黒区
品川区
大田区


2012_w8_3
<その他の値すべて>
wb8
1
2
西多摩
多摩小平
多摩立川
八王子市
多摩府中
南多摩
町田市
板橋区
北区
足立
練馬区
荒川区
葛飾区
池袋
文京
台東
豊田区
中野区
杉並
新宿区
千代田
江戸川
渋谷区
中央区
江東区
みなと
世田谷
目黒区
江東区
品川区
大田区
島しょ


2012_w9_3
<その他の値すべて>
wb9
1
2
西多摩
多摩小平
多摩立川
八王子市
多摩府中
南多摩
町田市
板橋区
北区
足立
練馬区
荒川区
葛飾区
池袋
文京
台東
豊田区
中野区
杉並
新宿区
千代田
江戸川
渋谷区
中央区
江東区
みなと
世田谷
目黒区
江東区
品川区
大田区
島しょ


2012_w10_3
<その他の値すべて>
wb10
1
西多摩
多摩小平
多摩立川
八王子市
多摩府中
南多摩
町田市
島しょ
板橋区
北区
足立
練馬区
葛飾区
荒川区
豊島区
中野区
杉並
文京
台東
墨田区
新宿区
千代田
江戸川
渋谷区
中央区
江東区
みなと
目黒区
品川区
大田区

厚生労働科学研究費補助金（健康安全・危機管理対策総合研究事業）
分担研究報告書

東京都におけるインフルエンザ学級閉鎖状況の地理情報システムを用いた可視化に関する研究（2011-2012年シーズン）

研究協力者　東京都健康安全研究センター　高橋琢理
研究協力者　東京都健康安全研究センター　杉下由行
研究分担者　東京都健康安全研究センター　灘岡陽子

研究要旨：東京都におけるインフルエンザによる学級閉鎖実施状況の地理的な分布とその規模を地理情報システム（GIS）によって地図上に描画することで可視化し、動画・静止画によってその流行状況を把握することを試みた。また、従来から実施されているインフルエンザ定点医療機関からの報告と比較を行った。その結果、学級閉鎖の時系列変化を示した動画の分析から、学級閉鎖の実施状況が定点報告と関連することが示唆された。学級閉鎖状況からインフルエンザの流行状況を推測する方法を開発し、公衆衛生行政の疫学情報として還元することが今後の課題である。

A．研究目的

地理情報システム（GIS）による情報活用と公開が様々な分野で進められている。自治体においても、行政データの分析や住民への情報提供などへの地理情報の活用が求められている1)。公衆衛生行政では、地理情報に基づいた疫学情報の分析とその提供が課題であり、例えば、季節性インフルエンザについては、発生状況の把握と分析、流行拡大への対策、住民への情報提供などに地理情報を活用することが求められる。そこで、本研究ではインフルエンザによる学級閉鎖実施状況の地理的な分布とその規模をGISによって地図上に描画することで可視化し、動画・静止画によってその流行状況を把握することを試みた。

B．研究方法

東京都内の高校、中学校、小学校、幼稚園、保育園、各種学校（以下、学校）で2011-2012年シーズン（2011年9月1日～2012年8月31日）に報告された学級閉鎖・学年閉鎖の状況を対象とした。

学級閉鎖・学年閉鎖の実施状況に関しては学校から管轄保健所に対し以下の内容が報告された。市区町村、学校種別、学校名、臨時休業の種別（学級閉鎖、あるいは学年閉鎖）、閉鎖する学年・年齢、閉鎖学年数、閉鎖学級数、在籍者数、インフルエンザ様疾患による欠席者数、インフルエンザ様疾患罹患登校者数、閉鎖期間開始日、閉鎖期間終了日。

学級閉鎖・学年閉鎖の実施された学校の緯度経度情報を、市区町村、学校種別および学校名をもとに、国土交通省国土政策局の国土数値情報公共施設データ（東京都、平成18年度）から得た2)。新設や統廃合、あるいは名称変更などがなされた学校では、上記データで緯度経度が得られないため、インターネットから検索した住所をもとにGeocodingサービス3)から緯度経度情報を得た。これらの情報によりGISソフト（

ArcGIS 10 ESRI社製）によって学校の位置を地図上に表示した。また、在籍者の数を色で分類し、学級閉鎖・学年閉鎖の期間と規模を表示する動画として可視化した。また、従来から実施されているインフルエンザ定点医療機関からの報告と比較を行った。

図1 2011-2012年シーズンの都内学級閉鎖状況例

C．研究結果

2011-2012年シーズンに東京都内で報告されたインフルエンザによる学級閉鎖・学年閉鎖の分布状況を静止画として可視化した。また、その実施状況の時系列変化を動画として示した。これにより、学校における学級閉鎖・学年閉鎖の状況が地理情報として把握可能となった。図1に学級閉鎖状況の例を示す。点は学級閉鎖等の実施された学校の所在地であり、2012年第3週に都内全域で学級閉鎖が実施されたことが示された。

また、時系列変化の動画から、2011-2012年シーズンの学級閉鎖は、2012年1月の第3週に都内全域で一斉に増加し、3月20日ごろに終息したことが明らかとなった。これはインフルエンザ定点医療機関からの報告の増加および終息の傾向と同じであった。資料に時系列動画表示を分解して示した。

D．考察

学級閉鎖の時系列変化を示した動画の分析からは、学級閉鎖の実施状況が定点報告と関連することが示唆された。公立の小学校、中学校、幼稚園は各地域で均等に配置されるように学区が設けられている。これらの学校に通う子供たちは、学区住民の子供たちである。そのため、学区内のインフルエンザ流行状況が子供たちの学級閉鎖にあらわれていると仮定すると、地域におけるインフルエンザの流行状況をある程度反映していると考えることもできる。また、成人における職場を中心とした流行は、定点医療機関からの報告や、薬局サーベイランスなどが実施されている。しかしながら、基礎的なデータは不足しており、地理情報についても十分なデータが得られていない。一方、学校は、子供の集団生活の場であり、学校における学級閉鎖の状況を分析することで、職場を介した成人におけるインフルエンザ流行状況を推測する方法も得られる可能性が考えられる。

E．結論

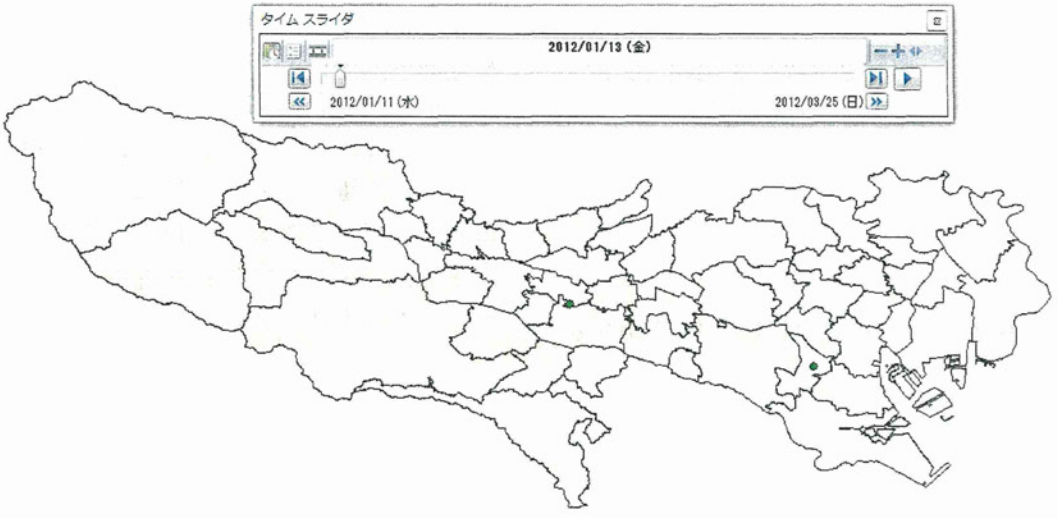
インフルエンザによる学級閉鎖・学年閉鎖の状況を、GISによって地理的な分布と規模、さらに時間変化を静止画と動画により可視化した。これらの結果から、学校における学級閉鎖状況が、インフルエンザ定点報告の傾向を反映する可能性が示された。本研究では、学級閉鎖と定点報告との関連性の数理的な検討が不十分である。また、地域の流行状況の反映程度についての分析も限定されており、これらは制限として挙げられる。今後、数理的な検討によって、学級閉鎖と定点報告との関連性を明らかにするとともに、学級閉鎖状況から流行状況を推測する方法を開発し、疫学情報として還元することが、今後の課題である。

参考

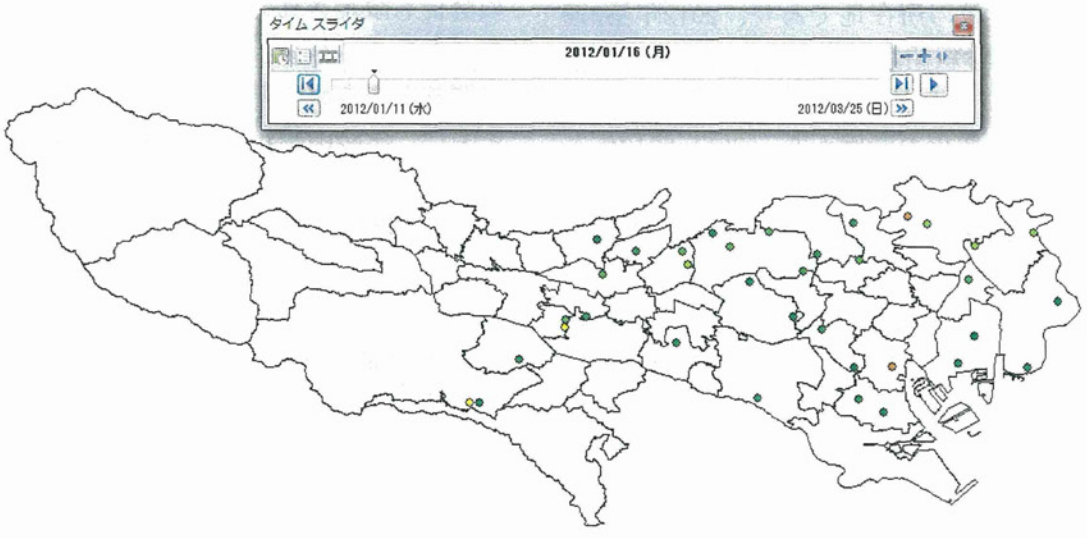
1) 地理空間情報活用推進基本法
（平成十九年五月三十日法律第六十三号）
2) 国土交通省国土政策局 国土数値情報ダウンロードサービスhttp://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-P02-v4_0.html
(2013年4月10日閲覧確認)
3) Geocoding http://www.geocoding.jp/
(2013年4月10日閲覧確認)

資料

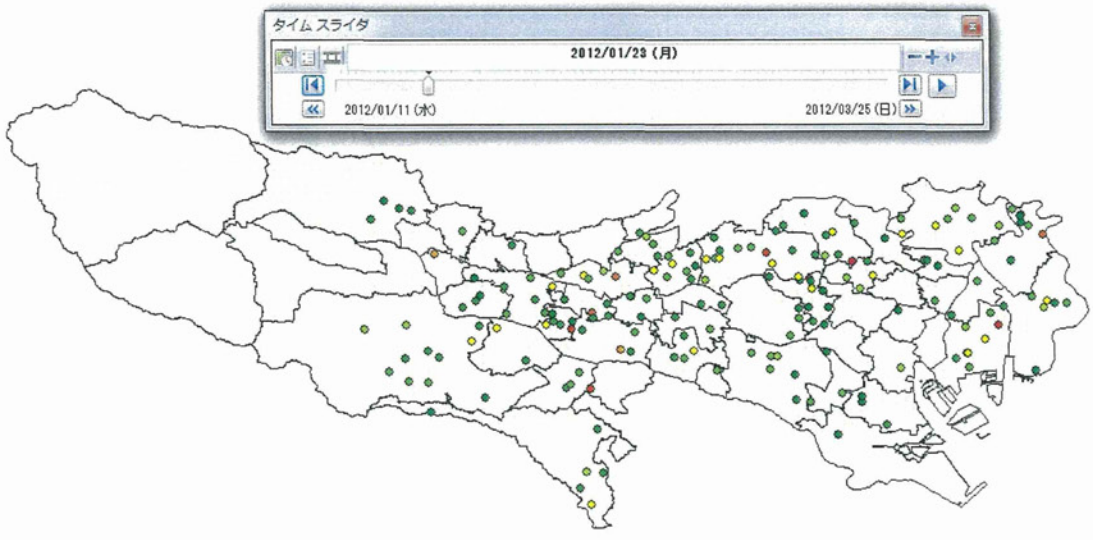
動画によるインフルエンザ流行状況１　2012年1月13日



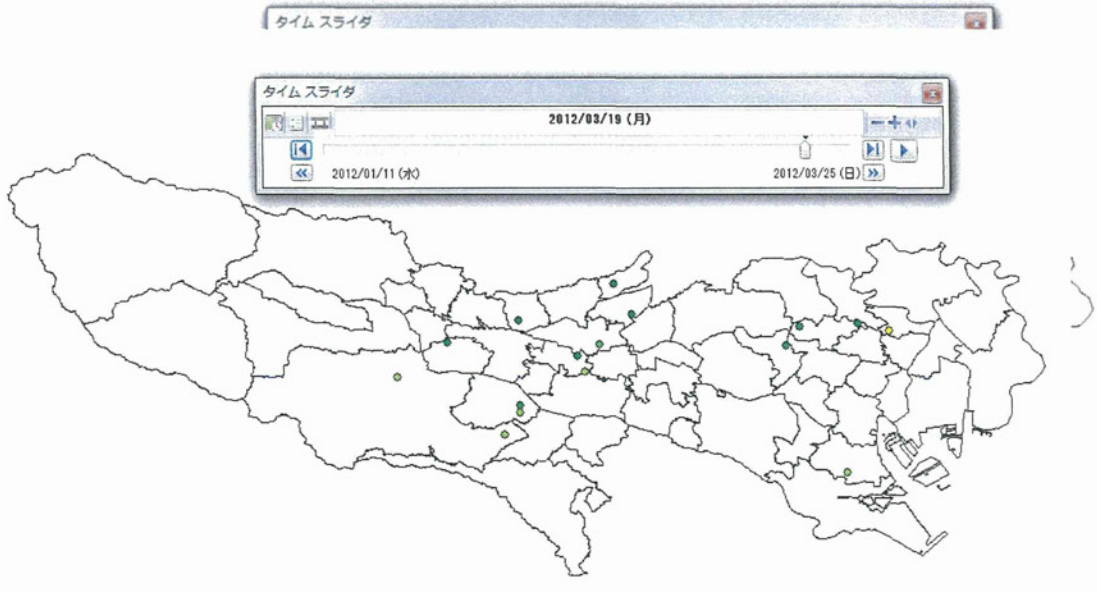
動画によるインフルエンザ流行状況２　2012年1月16日



動画によるインフルエンザ流行状況３　2012年1月23日



動画によるインフルエンザ流行状況４　2012年3月19日



F．研究発表

1．論文発表
なし
2．学会発表
なし
3．著書
なし

G．知的所有権の取得状況
なし

研究成果の刊行に関する一覧表

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Tango T. Takahashi K.	A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters.	Statistics in Medicine	31	4207-4218	2012
高橋邦彦，武村真治，長谷川学，金谷泰宏，齋藤大蔵，阪本敏久	わが国における救急蘇生統計を用いた冬季の心肺機能停止傷病者救急搬送件数の時間的集積性の検出	日本臨床救急医学会雑誌	15	652-661	2012

Special Issue Paper

Statistics in Medicine

Received 24 October 2011, Accepted 1 May 2012 Published online 16 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5478

A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters

Toshiro Tango[a*†] and Kunihiko Takahashi[b]

Spatial scan statistics are widely used tools for detection of disease clusters. Especially, the circular spatial scan statistic proposed by Kulldorff (1997) has been utilized in a wide variety of epidemiological studies and disease surveillance. However, as it cannot detect noncircular, irregularly shaped clusters, many authors have proposed different spatial scan statistics, including the elliptic version of Kulldorff's scan statistic. The flexible spatial scan statistic proposed by Tango and Takahashi (2005) has also been used for detecting irregularly shaped clusters. However, this method sets a feasible limitation of *a maximum of 30 nearest neighbors* for searching candidate clusters because of heavy computational load. In this paper, we show a flexible spatial scan statistic implemented with a restricted likelihood ratio proposed by Tango (2008) to (1) eliminate the limitation of *30 nearest neighbors* and (2) to have surprisingly much less computational time than the original flexible spatial scan statistic. As a side effect, it is shown to be able to detect clusters with any shape reasonably well as the relative risk of the cluster becomes large via Monte Carlo simulation. We illustrate the proposed spatial scan statistic with data on mortality from cerebrovascular disease in the Tokyo Metropolitan area, Japan. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: cluster detection; hotspot cluster; likelihood ratio statistic; Monte Carlo testing; spatial epidemiology

1. Introduction

Many authors have proposed and introduced in several textbooks many different statistical methods for detecting disease clustering and disease clusters [1–4]. Especially, the spatial scan statistic proposed by Kulldorff and Nagarwalla [5] and Kulldorff [6] along with SaTScan software [7] has been applied in a wide variety of epidemiological studies and also in disease surveillance for detection of disease clusters. The spatial scan statistic tries to identify the *most likely cluster* (MLC), which is defined as the set of connected regions that attains the maximum likelihood ratio. However, because it uses a circular window to scan the potential cluster areas, it has difficulty in correctly detecting actual noncircular clusters. To detect arbitrarily shaped clusters that cannot be detected by the circular spatial scan statistic, many authors [8–11] have proposed different spatial scan statistics. Kulldorff *et al.* [12] also proposed an elliptic version of spatial scan statistic. Above all, the flexible spatial scan statistic proposed by Tango and Takahashi [10] along with FleXScan software [13] has also been used for detecting irregularly shaped clusters. However, this method allows only a *maximum of 30 nearest neighbors* for searching candidate clusters because of heavy computational load. Several authors who conducted comparison of tests for disease clusters pointed out this limitation (e.g., [14]).

On the other hand, Tango [15] and Tango and Takahashi [10] have shown that these spatial scan statistics based on maximizing the ordinary likelihood ratio tend to detect an MLC that is much larger than the true cluster by absorbing neighboring regions with nonelevated risk of disease occurrence. To

[a] *Center for Medical Statistics, 2-9-6 Higashi Shimbashi, Minato-ku, Tokyo, 105-0021, Japan*

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

**Correspondence to: Toshiro Tango, Center for Medical Statistics, 2-9-6 Higashi Shimbashi, Minato-ku, Tokyo, 105-0021, Japan.*

[†] *E-mail: tango@medstat.jp*

avoid such an undesirable property, Tango [16] proposed a new spatial scan statistic by restricting the likelihood ratio so that it scans only the regions with elevated risk. Tango [16] compared the performance of detecting *circular clusters* by the two methods, that is, the original Kulldorff's circular spatial scan statistic and the circular spatial scan statistic implemented with the proposed restricted likelihood ratio, via a Monte Carlo simulation study. The results indicated that the latter procedure had the ability to identify the whole, or a part, of the true cluster assumed in the simulation more often than the former.

In this paper, we shall propose a flexible spatial scan statistic implemented with the restricted likelihood ratio. The performance of the proposed spatial scan statistic with the restricted likelihood ratio is compared with those of Tango and Takahashi's original flexible spatial scan statistic and Kulldorff's circular spatial scan statistic via a Monte Carlo simulation study. We illustrate the proposed procedure with data on mortality from cerebrovascular disease in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan.

2. Spatial scan statistics

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

$$H_0 : E(N_i) = \xi_i,\ N_i \sim \text{Poisson}(\xi_i), \quad i = 1, \ldots, m, \tag{1}$$

where Poisson(ξ) denotes the Poisson distribution with mean ξ and the ξ_i are the expected number of cases in region i under the null hypothesis. For calculation of the expected number of cases adjusted for potential confounders such as age, we can use indirect standardization or a Poisson mixed-effects regression model [7]. If we ignore the confounders, we can calculate ξ_i as

$$\xi_i = n \frac{w_i}{\sum_{k=1}^{m} w_k}, \quad i = 1, \ldots, m, \tag{2}$$

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

2.1. Windows to be scanned

In the aforementioned situation, Kulldorff's circular spatial scan statistic imposes a circular window $\mathbf{Z}$ on each centroid of regions. For any of those centroids, the radius of the circle varies continuously from zero up until 50% of the population at risk is covered, which is the standard option in 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 $\mathbf{Z}_{ik}$ $(k = 1, \ldots, K_i)$ denote the window composed of the $(k-1)$-nearest neighbors to region i. Then, all of the windows to be scanned by the circular spatial scan statistic are included in the set

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

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

$$\mathcal{Z} = \mathcal{Z}_2 = \left\{\mathbf{Z}_{ik(j)} \mid 1 \leqslant i \leqslant m,\ 1 \leqslant k \leqslant K,\ 1 \leqslant j \leqslant j_{ik}\right\}.$$

In other words, for any given region i, the circular spatial scan statistic considers K concentric circles, whereas the flexible scan statistic considers K concentric circles plus all of the sets of connected regions (including the single region i) whose centroids are located within the Kth largest concentric circle. Therefore, the size of $\mathcal{Z}_2$ is much larger than that of $\mathcal{Z}_1$, which is at most mK. Because of this type of heavy computational load, that is, combinatorial explosion, the maximum length K should be less than 30; otherwise, the computation may take more than 1 day, or 1 week, depending on the data. Therefore, the value of K is set as 15 as the default in FleXScan.

2.2. *Likelihood ratio statistic*

With the use of the notation of window $\mathbf{Z} \in \mathcal{Z}$, the null hypothesis (1) is expressed as

$$H_0 : E(N(\mathbf{Z})) = \xi(\mathbf{Z}), \quad \text{for all } \mathbf{Z} \in \mathcal{Z}, \tag{3}$$

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. Under the alternative hypothesis H_1, 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, that is,

$$H_1 : E(N(\mathbf{Z})) > \xi(\mathbf{Z}), \quad \text{for some } \mathbf{Z} \in \mathcal{Z}. \tag{4}$$

For each window $\mathbf{Z}$, it is possible to compute a likelihood of observing the observed number of cases within and outside the window, respectively. Under the assumption of a Poisson distribution (1), the well-known likelihood ratio statistic proposed by Kulldorff [6] is given by

$$\lambda_K = \max_{\mathbf{Z} \in \mathcal{Z}} \lambda_K(\mathbf{Z}) = \max_{\mathbf{Z} \in \mathcal{Z}} \left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{n(\mathbf{Z})} \left(\frac{n - n(\mathbf{Z})}{n - \xi(\mathbf{Z})}\right)^{n - n(\mathbf{Z})} I\left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} > \frac{n - n(\mathbf{Z})}{n - \xi(\mathbf{Z})}\right), \tag{5}$$

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 *MLC*. However, it does not seem to be well recognized that the spatial scan statistics using the likelihood ratio defined earlier tend to detect an MLC that is much larger than the true cluster by swallowing neighboring regions with nonelevated risk [10, 15]. Tango [16] proposed the following:

Proposition ([16])
In the process of scanning the window based on $\lambda_K(\mathbf{Z})$, there is a possibility that there exists two disjoint windows $\mathbf{Z}_1$ and $\mathbf{Z}_2$ and several regions $\{i_1\}, \ldots, \{i_r\}$ such that

$$\lambda_K(\{\mathbf{Z}_1, \mathbf{Z}_2, \{i_1\}, \ldots, \{i_r\}\}) > \max\{\lambda_K(\mathbf{Z}_1),\ \lambda_K(\mathbf{Z}_2)\}, \tag{6}$$

where

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

The aforementioned proposition means 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 nonsignificantly elevated risk due to random fluctuation or with nonelevated risk.

2.3. *Restricted likelihood ratio statistic*

To avoid or scale back such undesirable phenomena, Tango [16] proposed the following restricted likelihood ratio by taking each individual region's risk into account:

$$\lambda_T(\mathbf{Z}) = \left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{n(\mathbf{Z})} \left(\frac{n - n(\mathbf{Z})}{n - \xi(\mathbf{Z})}\right)^{n - n(\mathbf{Z})} I\left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} > \frac{n - n(\mathbf{Z})}{n - \xi(\mathbf{Z})}\right) \prod_{i \in \mathbf{Z}} I\left(p_i < \alpha_1\right), \tag{7}$$

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

$$p_i = \Pr\{N_i \geqslant n_i\ + 1 \mid N_i \sim \text{ Poisson}(\xi_i)\} + \frac{1}{2}\Pr\{N_i = n_i \mid N_i \sim \text{ Poisson}(\xi_i)\}, \tag{8}$$

and α_1 is the prespecified significance level for the individual region. The reason why the middle p-value was used is to adjust for conservatism of the ordinary definition of p-value for small ξ_i and count outcomes. In this formulation, we devised $I(p_i < \alpha_1)$ as a screening criterion. Therefore, as in the case of the original flexible spatial scan statistic, the p-value of the flexible scan statistic based on the restricted likelihood ratio is obtained through Monte Carlo hypothesis testing [17].

Tango [16] investigated properties of the circular spatial scan statistic with the restricted likelihood ratio via a Monte Carlo simulation study. It was shown to have better ability to identify the true circular cluster compared with Kulldorff's original one in all of the circular cluster models considered as described in the succeeding discussion. The results of the Monte Carlo simulation study also suggest the following guidance regarding the choice of α_1 for a restricted likelihood ratio statistic of the nominal α level of 0.05 ($= \alpha_0$.): (1) $\alpha_1 = 0.10$–0.20 to detect small clusters with a sharp increase in risk; (2) $\alpha_1 = 0.20$–0.30 to detect small to middle-sized clusters with a moderate increase in risk; and (3) $\alpha_1 = 0.30$–0.40 to detect larger clusters with a slight increase in risk. Tango [16] further recommends $\alpha_1 = 0.20$ as a default.

In what follows, we shall examine the properties of the proposed flexible spatial scan statistic implemented with the restricted likelihood ratio statistic $\lambda_T = \max_{\mathbf{Z}\in\mathcal{Z}} \lambda_T(\mathbf{Z})$ with $\alpha_1 = 0.20$ by comparing its performance with that of Kulldorff's circular spatial scan statistic and that of the original flexible spatial scan statistic via illustration with mortality data and Monte Carlo simulations at a significance level of $\alpha_0 = 0.05$.

3. Illustration

As an illustration, we shall apply the three spatial scan statistics, that is, the proposed flexible spatial scan statistic with $\alpha_1 = 0.2$, the original flexible spatial scan statistic, and Kulldorff's circular spatial scan statistic, to data on mortality from cerebrovascular disease (1993–1997) in women in the areas

Table I. The most likely clusters detected at $\alpha_0 = 0.05$ by three methods: the original flexible spatial scan statistic with $K = 20$, the proposed flexible spatial scan statistic with $K = 50$ (about half the number of regions and essentially no restrictions) and $\alpha_1 = 0.02$, and Kulldorff's circular spatial scan statistic with $K = 50$ in their application to the data on mortality from cerebrovascular disease (female, 1993–1997) in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan.

No.	Region no.	Observed no. of cases	Expected no. of 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}
7	1	164	144.7	1.13	0.057
8	17	1110	999.3	1.11	0.00029
9	21	1530	1335.1	1.15	9.2×10^{-8}
10	16	798	743.4	1.07	0.024
		The proposed flexible scan's MLC= {23, 22, 18, 7, 8, 6, 1, 17, 21, 16}, which is the same result as that of the original flexible scan $\log \lambda_T = 151.7$, $\hat{\theta} = 1.18$, p-value = 0.001 Running time: the original = 338 s, the proposed = less than 1 s			
11	5	548	566.3	0.97	0.778
12	2	267	251.3	1.06	0.161
		Kulldorff's circular scan's MLC= {23, 22, 18, 7, 8, 6, 1, 17, 21, 16, 5, 2} $\log \lambda_K = 140.6$, $\hat{\theta} = 1.17$, p-value = 0.001 Running time: 2 s			

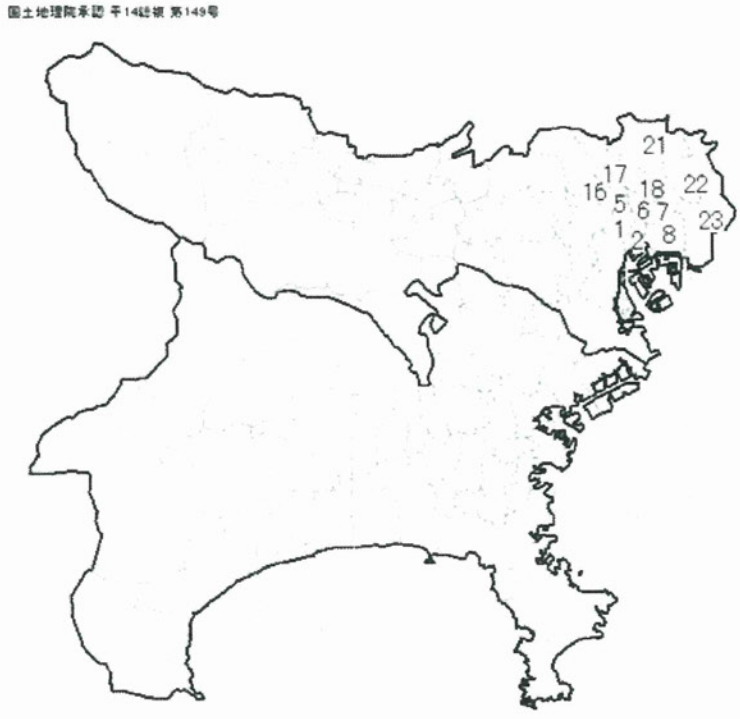


Figure 1. The 113 regions that comprise wards, cities, and villages in the Tokyo Metropolis and Kanagawa prefecture in Japan. The region numbers used in Section 3 are indicated.

of Tokyo Metropolis and Kanagawa prefecture in Japan. The total number of observed deaths from cerebrovascular disease in women over the 5-year period was 45,700 in this area. Regarding the maximum length for K-nearest neighbors, we shall select $K = 20$ for the original flexible spatial scan statistics and $K = 50$ for the proposed flexible and Kulldorff's circular spatial scan statistics. The value of $K = 50$ in this example corresponds to about half the number of regions. We calculate the p-value of spatial scan statistics with the use of 999 replications for the Monte Carlo hypothesis testing. We provide the results in Table I and summarize these as follows: the proposed flexible spatial scan statistic detected an MLC consisting of 10 regions $\{1, 6, 7, 8, 16, 17, 18, 21, 22, 23\}$ (Figure 1 shows these region numbers) with $\log \lambda_T (= \log \lambda_K) = 151.7$, $\hat{\theta} = 1.18$, and $p = 1/(999 + 1) = 0.001$. Most of these 10 regions within the MLC are shown to have significantly elevated risk. The original flexible spatial scan statistic detected the same MLC in this example. However, the running time of the proposed flexible spatial scan statistic was less than 1 s, which is much less than the running time of 338 s (using PC Window 7, Intel(R), Core(TM)2 Duo CPU P8800) of the original flexible spatial scan statistic. On the other hand, Kulldorff's circular spatial scan statistic added two more regions $\{2, 5\}$ to the MLC detected by the proposed flexible spatial scan statistic, with $\log \lambda_K = 140.6$, $\hat{\theta} = 1.17$, and $p = 0.001$. However, the two regions $\{2, 5\}$ did not have significantly elevated risk. Especially, region $\{5\}$, called 'Bunkyo-ku', is well known as a 'healthy district' in this area, and it had a nonelevated relative risk of 0.97 and $p_5 = 0.778$ as expected. Therefore, it seems unacceptable that 'Bunkyo-ku' is included in the MLC.

4. Simulation study

In this section, we shall carry out an extensive Monte Carlo simulation study to compare the performance of three spatial scan statistics: Kulldorff's circular spatial scan statistic, the original flexible spatial scan statistic, and the proposed flexible spatial scan statistic with restricted likelihood ratio. We will sometimes refer to these statistics as CS, FS, and FSR, respectively.

4.1. Simulated data

As the entire study population, we will use the area of Tokyo Metropolis and Kanagawa prefecture, which consist of 113 regions (wards, cities, and villages), in Japan (Figure 1), described in Section 3. The maximum length of *K-nearest neighbors* is set as $K = 20$ for FS and $K = 50$ (indicating virtually no restrictions) for FSR. The maximum cluster size for CS is also set as $K = 50$. We shall consider the following two different clustered areas with different types of shapes:

1. a crossed-line-shaped cluster $\boldsymbol{Z} = \{25, 29, 30, 31, 32, 33, 38, 39, 40, 41, 47, 49, 53, 72, 73, 74, 75, 76, 86, 93, 92, 97\}(s^* = 22)$ and
2. a ring-shaped cluster $\boldsymbol{Z} = \{32, 56, 57, 58, 59, 62, 63, 73, 74, 75, 76, 78, 81, 82, 86, 93, 92\}$ $(s^* = 17)$

where s^* denotes the length of the cluster. In one simulation study, we assume that there is only one true cluster out of the aforementioned two clusters. Figures 2 and 3 respectively show these two clusters. The reason why we chose these two clustered areas is that neither CS nor FS (because of the restrictions on the number of regions) can detect these two clusters exactly.

Data to be simulated underwent the following three steps:

1. In each region $i(= 1, \ldots, m = 113)$, we have a random (nonclustered) observed number of cases n_{0i}, which have a Poisson distribution with the null expected number of cases ξ_i, which was defined as

$$\xi_i = n_0 \frac{w_i}{\sum_{k=1}^{m} w_k}, \quad i = 1, \ldots, m,$$

where n_0 denotes the prespecified total expected number of cases under the null hypothesis of no clustering, and we consider here three values of n_0, $n_0 = 100, 200$, and 500.

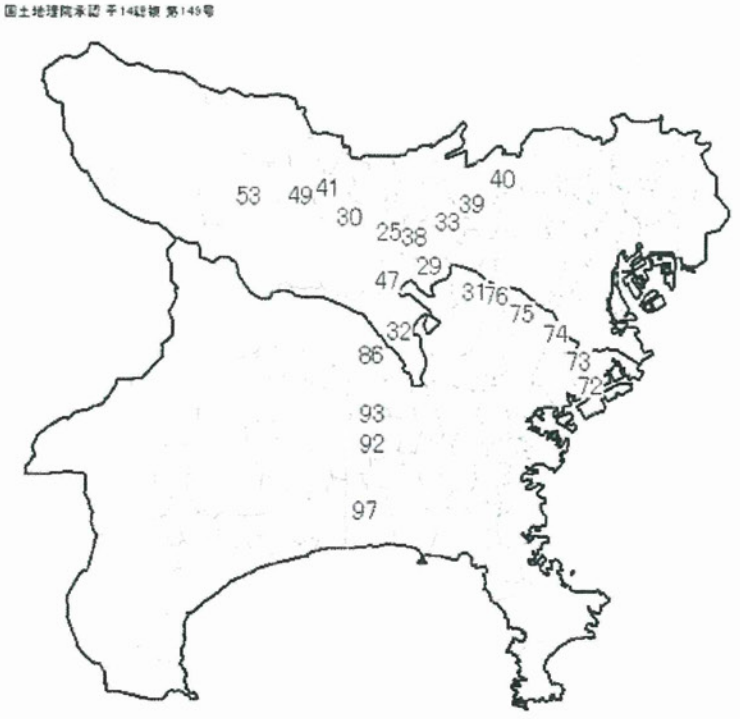


Figure 2. The 113 regions that comprise wards, cities, and villages in the Tokyo Metropolis and Kanagawa prefecture in Japan, which were used as the entire study population for the Monte Carlo simulation study in Section 4. The 22 regions whose region number is shown here constitute the crossed-line-shaped cluster assumed in the simulation study.

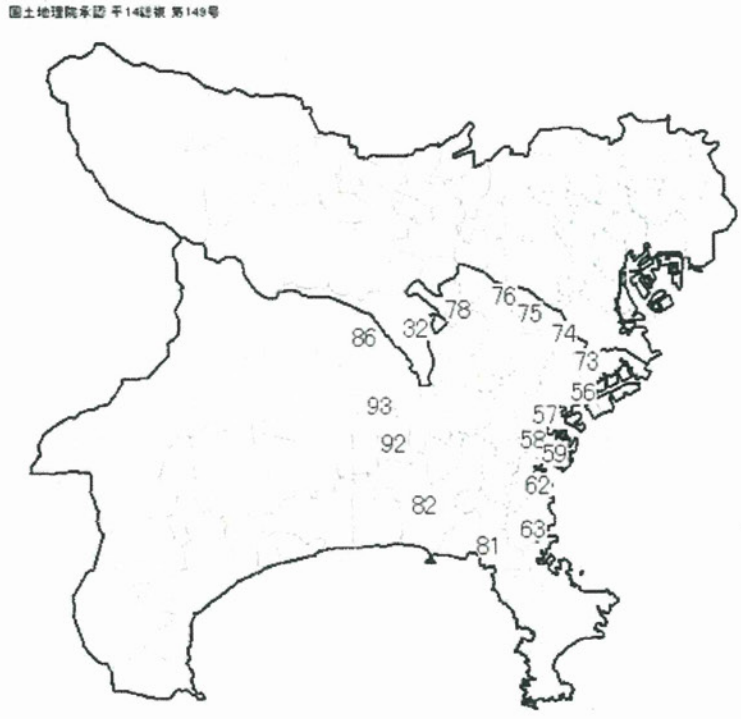


Figure 3. The 113 regions that comprise wards, cities, and villages in the Tokyo Metropolis and Kanagawa prefecture in Japan, which were used as the entire study population for the Monte Carlo simulation study in Section 4. The 17 regions whose region number is shown here constitute the ring-shaped cluster assumed in the simulation study.

2. In each region $i \in \boldsymbol{Z}$, we have an additional number of cases n_{1i} due to clustering such that

$$\Pr\{X \geqslant n_{1i} \mid X \sim \text{Poisson}(\xi_i)\} = Q \tag{9}$$

in which the number n_{1i} denotes the upper $100Q$ percentile of the Poisson distribution. We call the value of Q *strength of cluster* throughout the paper in that the smaller the value of Q, the stronger the cluster. We consider here three values of Q, $Q = 0.05, 0.01$, and 0.001.

3. Finally, we have

$$n_i = \begin{cases} n_{0i} + n_{1i}, & \text{if } i \in \boldsymbol{Z} \\ n_{0i}, & \text{otherwise.} \end{cases} \tag{10}$$

Thus, we evaluate a total of 18 ($2 \times 3 \times 3$) clustering scenarios.

4.2. *Bivariate power distribution*

We carried out Monte Carlo simulations with the use of 1000 replications at significance levels of $\alpha_0 = 0.05$ and $\alpha_1 = 0.20$. To compare the performance of the cluster detection tests, the usual power has been used by many authors. However, it should be noted that the usual power estimates reflected the *power to reject the null hypothesis for whatever reason*, whereas the probability of both rejecting the null hypothesis and accurately identifying the true cluster is a different matter altogether. To compare the performance of the spatial cluster detection tests, Tango and Takahashi [10] proposed a bivariate power distribution $P(l, s|s^*)$ based on Monte Carlo simulation where l is the *length* of the significant MLC and s is the number of regions identified out of the assumed true cluster with s^* regions.

$$\begin{aligned} P(l, s) &= \Pr\{L = l, S = s \mid s^*\} \\ &= \frac{\#\{\text{significant MLC has } \textit{length } l \text{ and includes } s \text{ true regions}\}}{\#\{\text{trials for each simulation}\}}, \end{aligned} \tag{11}$$

where L and S denote the random variable of l and s under the specified model, respectively, and $1 \leqslant l$ and $0 \leqslant s \leqslant \min\{l, s^*\}$. We are especially interested in the power around the point $(l = s^*, s = s^*)$ and $P(s^*, s^*)$, the probability of exact detection. The usual power, on the other hand, 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 does not detect any clusters.

4.3. *Results*

We provide the results in Tables II–V. Tables II–IV show the estimated bivariate power distributions $P(l, s) \times 1000$ of CS, FS, and FSR, respectively, for the crossed-line-shaped cluster $\boldsymbol{Z}$ shown in Figure 2, where the strength of the cluster is set as $Q = 0.01$ and the total expected number of cases under the null hypothesis as $n_0 = 500$. The three spatial scan statistics have perfect usual power, that is, 100% power. However, Table II indicates that the support of the power distribution of CS tends to be scattered over the range $\{(l, s) : 1 \leqslant l \leqslant l_{\max}, 1 \leqslant s \leqslant s^*\}$ and has zero probability of detecting the cluster exactly, indicating that the detected clusters never contain the whole true cluster and tend to be much larger than the true cluster by swallowing up neighboring regions with nonelevated risk. Table III also

Table II. Estimated bivariate power distributions $P(l, s) \times 1000$ of Kulldorff's circular spatial scan statistic with $K = 50$ for the crossed-line-shaped cluster $\boldsymbol{Z} = \{25, 29, 30, 31, 32, 33, 38, 39, 40, 41, 47, 49, 53, 72, 73, 74, 75, 76, 86, 93, 92, 97\}(s^* = 22)$ with $Q = 0.01$.

Length l	Included s hotspot regions												
	0	1–2	3–4	5–6	7–8	9–10	11–12	13–14	15–16	17–18	19–20	21	22
1–2	0	0											
3–4	0	0	0										
5–6	0	0	0	0									
7–8	0	0	0	0	0								
9–10	0	0	0	0	0	0							
11–12	0	0	0	0	9	0	0						
13–14	0	0	0	0	0	9	0	0					
15–16	0	0	0	0	0	103	0	0	0				
17–18	0	0	0	0	0	18	0	0	0	0			
19–20	0	0	0	0	0	0	21	0	0	0	0		
21	0	0	0	0	0	0	0	0	0	0	0	0	
22	0	0	0	0	0	0	0	141	0	0	0	0	0[#]
23	0	0	0	0	0	0	0	3	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	1	108	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	2	1	0	0	0	0
30	0	0	0	0	0	0	0	0	3	0	0	0	0
31–35	0	0	0	0	0	0	0	0	14	0	0	0	0
36–40	0	0	0	0	0	0	0	1	3	308	0	0	0
41–45	0	0	0	0	0	0	0	0	1	195	0	0	0
46–50	0	0	0	0	0	0	0	0	1	58	0	0	0
Total*	0	0	0	0	9	130	21	148	131	561	0	0	0

Significance level was $\alpha_0 = 0.05$. We carried out $n_0 = 500$ and 1000 trials. Running time per trial was 1 s on average.
*Usual power is 1000/1000.
[#]The proportion of exact detection is 0/1000.

Table III. Estimated bivariate power distributions $P(l, s) \times 1000$ of Tango and Takahashi's original flexible spatial scan statistic with $K = 20$ for the crossed-line-shaped cluster $Z = \{25, 29, 30, 31, 32, 33, 38, 39, 40, 41, 47, 49, 53, 72, 73, 74, 75, 76, 86, 93, 92, 97\}(s^* = 22)$ with $Q = 0.01$.

Length l	Included s hotspot regions												
	0	1–2	3–4	5–6	7–8	9–10	11–12	13–14	15–16	17–18	19–20	21	22
1–2	0	0											
3–4	0	0	0										
5–6	0	0	0	0									
7–8	0	0	0	0	0								
9–10	0	0	0	0	1	19							
11–12	0	0	0	0	0	28	781						
13–14	0	0	0	0	0	5	155	0					
15–16	0	0	0	0	0	0	9	0	0				
17–18	0	0	0	0	0	0	2	0	0	0			
19–20	0	0	0	0	0	0	0	0	0	0	0		
21	0	0	0	0	0	0	0	0	0	0	0	0	
22	0	0	0	0	0	0	0	0	0	0	0	0	0[#]
Total*	0	0	0	0	1	52	947	0	0	0	0	0	0

Significance level was $\alpha_0 = 0.05$. We carried out $n_0 = 500$ and 1000 trials. Running time per trial was 120 s on average.

*Usual power is 1000/1000.

[#]The proportion of exact detection is 0/1000.

Table IV. Estimated bivariate power distributions $P(l, s) \times 1000$ of the proposed flexible spatial scan statistic with restricted likelihood ratio with $K = 50$ for the crossed-line-shaped cluster $Z = \{25, 29, 30, 31, 32, 33, 38, 39, 40, 41, 47, 49, 53, 72, 73, 74, 75, 76, 86, 93, 92, 97\}(s^* = 22)$ with $Q = 0.01$.

Length l	Included s hotspot regions												
	0	1–2	3–4	5–6	7–8	9–10	11–12	13–14	15–16	17–18	19–20	21	22
1–2	0	0											
3–4	0	0	0										
5–6	0	0	0	0									
7–8	0	0	0	0	0								
9–10	0	0	0	0	0	0							
11–12	0	0	0	0	0	0	0						
13–14	0	0	0	0	0	0	0	0					
15–16	0	0	0	0	0	0	0	0	0				
17–18	0	0	0	0	0	0	0	0	0	0			
19–20	0	0	0	0	0	0	0	0	0	0	0		
21	0	0	0	0	0	0	0	0	0	0	0	0	
22	0	0	0	0	0	0	0	0	0	0	0	0	107[#]
23	0	0	0	0	0	0	0	0	0	0	0	0	260
24	0	0	0	0	0	0	0	0	0	0	0	0	271
25	0	0	0	0	0	0	0	0	0	0	0	0	173
26	0	0	0	0	0	0	0	0	0	0	0	0	100
27	0	0	0	0	0	0	0	0	0	0	0	0	54
28	0	0	0	0	0	0	0	0	0	0	0	0	23
29	0	0	0	0	0	0	0	0	0	0	0	0	9
30	0	0	0	0	0	0	0	0	0	0	0	0	2
31	0	0	0	0	0	0	0	0	0	0	0	0	1
Total*	0	0	0	0	0	0	0	0	0	0	0	0	1000

Significance levels were $\alpha_0 = 0.05$ and $\alpha_1 = 0.2$. We carried out $n_0 = 500$ and 1000 trials. Running time per trial was less than 1 s on average.

*Usual power is 1000/1000.

[#]The proportion of exact detection is 107/1000.

Table V. Estimated bivariate power distributions $P(l, s^*) \times 1000$, $l = s^*, \ldots, l_{\max}$ of the proposed flexible spatial scan statistic with $K = 50$ for the two true clusters assumed in our Monte Carlo simulation study.

		$n_0 = 100$			$n_0 = 200$			$n_0 = 500$		
Cluster type	Strength Q of cluster l	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
Cross	22#	0	88	188	53	108	211	59	107	245
	23	2	233	335	148	241	340	178	260	344
	24	3	260	263	185	279	255	258	271	227
	25	3	224	158	183	195	130	211	173	111
	26	2	124	43	126	111	45	133	100	46
	27	2	51	11	83	42	13	87	54	19
	28	0	13	1	45	13	3	35	23	5
	29	0	7	1	18	7	1	19	9	3
	30	1	0	0	6	3	2	12	2	0
	31	0	0	0	4	0	0	3	1	0
	32	0	0	0	1	0	0	1	0	0
	Total	13	1000	1000	852	999	1000	996	1000	1000
Loop	17#	37	124	288	70	180	233	138	278	385
	18	90	256	372	181	347	372	304	356	347
	19	134	261	215	255	258	255	267	190	176
	20	93	189	84	195	134	94	146	116	63
	21	70	110	34	102	53	36	89	40	19
	22	40	41	5	54	19	8	35	14	7
	23	13	16	2	31	6	2	13	4	1
	24	2	2	0	10	0	0	3	0	1
	25	0	1	0	2	0	0	1	1	1
	26	0	0	0	4	0	0	1	1	0
	27	0	0	0	0	0	0	1	0	0
	Total	479	1000	1000	905	997	1000	998	1000	1000

Significance levels were $\alpha_0 = 0.05$ and $\alpha_1 = 0.2$. We carried out 1000 trials.
#$l = s^*$ for the exact detection.

shows undesirable property of FS in that FS always detects clusters smaller than the true cluster, which is obviously due to the small value of K to avoid the heavy computational load of FS. Table IV, on the other hand, shows a good property of FSR. Namely, the power distribution appears to be concentrated in a relatively narrow range of the length l on the line $s = s^* = 22$, thereby indicating that the observed significant MLC always contains the true cluster in this case. Regarding the average *running time* per trial, the average running time is less than 1 s for FSR, 1 s for CS, and about 120 s for FS, indicating that the running time for FSR with virtually no restrictions on K is surprisingly less than that for FS with $K = 20$.

Although we omit here details of the estimated bivariate power distributions for other parameter values such as $Q = 0.05, 0.001$ and $n_0 = 100, 200$, and also for the case of the ring-shaped cluster shown in Figure 3, the relative performances of CS, FS, and FSR were essentially unchanged. Table V shows the estimated bivariate power distributions on the line $s = s^*(l = s^*, \ldots, l_{\max})$ of FSR for each of the two types of the true cluster, indicating clearly that FSR can detect an arbitrarily shaped cluster reasonably well as the strength or the relative risk of the cluster becomes large. Both CS and FS, on the other hand, have no such property. However, Table V also suggests that FSR still has an undesirable property to a certain extent in that the probability of detecting the true cluster together with a few of the neighboring regions with nonelevated risk is not small even when Q is quite small.

5. Discussion

In this paper, we examined the performance of the flexible spatial scan statistic implemented with the restricted likelihood ratio via a Monte Carlo simulation study and also compared the performance of the three spatial scan statistics, Kulldorff's circular spatial scan, Tango and Takahashi's flexible spatial scan,

and the proposed flexible spatial scan. As a result, the proposed flexible spatial scan statistic was shown to eliminate the constraint of the maximum of *30 nearest neighbors* for searching the cluster candidates and to have surprisingly much less running time than the original flexible spatial scan statistic of Tango and Takahashi. Furthermore, it tended to detect clusters of any shape reasonably well as the relative risk of the cluster becomes large as shown in Table V. Other spatial scan statistics do not seem to have this property. For example, Kulldorff's circular spatial scan statistic was shown to have relatively high usual power but tended to detect an MLC that was much larger than expected from the data. When applied to a noncircular cluster, the supports of the bivariate power distribution $P(l, s)$ of Kulldorff's circular spatial scan statistic were scattered over a wide range of points (l, s). In comparison, we did not include the elliptical version of spatial scan statistic [12], which introduces an *eccentricity penalty* that discourages eccentric clusters. The reason is that an elliptical window is obviously not able to detect clusters of any shape and is plagued with a large dose of subjectivity and noninterpretability in the penalty parameters.

The idea of the restricted likelihood ratio $\lambda_T(\mathbf{Z})$, on the other hand, seems to be quite natural and interpretable in that the regions with nonelevated risk should not be included in the MLC. However, this restriction may be too strict in some cases depending on the choice of α_1. For example, for a small to moderate increase in risk in a local area, random variation might result in nonelevated observed count $(p_i > \alpha_1)$ in regions with truly elevated risk. In this case, the proposed procedure may only identify the elevated portion of the true cluster. To remedy this situation to a certain extent, we recommend adopting a larger value for α_1 as discussed in Section 2.3.

Our simulation study adopted the value of $\alpha_1 = 0.20$, which was recommended as a default option by Tango [16]. We think this problem is quite similar to *selection of covariates* in any regression model. A well-known *stepwise regression* combines the two procedure of *forward selection* and *backward elimination* where p-values of selection criterion and elimination criterion are usually set at around 0.15–0.20. Needless to say, we can vary the value depending on the situation and/or the user's specific consideration. However, we do not think that the basic property of the proposed flexible spatial scan statistic observed in our simulation study would change drastically for other values of $\alpha_1 (< 0.5)$.

Finally, the proposed flexible spatial scan statistic has a better property in that the running time is quite fast and it can detect the true cluster with any shape reasonably well when the value of Q is large. However, it still suffers from the undesirable property that the probability of perfect detection of the true cluster cannot be 1.00 even when Q is quite small. We would like to examine this issue in our future work.

We have conducted all of the computations and simulations on a PC with Windows 7.

Acknowledgements

The authors would like to thank the referee for his insightful comments and suggestions. This research was partly funded by 2011 Grant-in-Aid for Scientific Research (grant no. 23300107) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Lawson AB, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R (eds). *Disease Mapping and Risk Assessment for Public Health*. John Wiley & Sons: New York, 1999.
2. Lawson AB, Denison D. *Spatial Cluster Modelling*. CRC Press: Boca Raton, Florida, 2002.
3. Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons: New York, 2004.
4. Tango T. *Statistical Methods for Disease Clustering*. Springer: New York, 2010.
5. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995; **14**:799–810. DOI: 10.1002/sim.4780140809.
6. Kulldorff M. A spatial scan statistic. *Communications in Statistics: Theory and Methods* 1997; **26**:1481–1496.
7. Kulldorff M. Information Management Services Inc. SaTScan v9.1.1: Software for the spatial and space-time scan statistics, 2011. http://www.satscan.org/.
8. Duczmal L, Assunção R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* 2004; **45**:269–286.
9. Patil GP, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 2004; **11**:183–197.
10. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 2005; **4**:11. DOI: 10.1186/1476-072X-4-11.
11. Assunção R, Costa M, Tavares A, Ferreira S. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine* 2006; **25**:723–742. DOI: 10.1002/sim.2411.