

解析対象のデータは以下の2つのパターンがあります。

◎ 各市区町村の期待度数を用いたモデルで解析を行う場合

- データファイル：Case File (observed # and expected #) (cas)
  - Format： <市区町村名> <観測度数> <期待度数>

解析対象全域を基準にした期待度数を用います。性・年齢調整など行った期待度数を用いることで性・年齢調整した結果が得られることとなります。

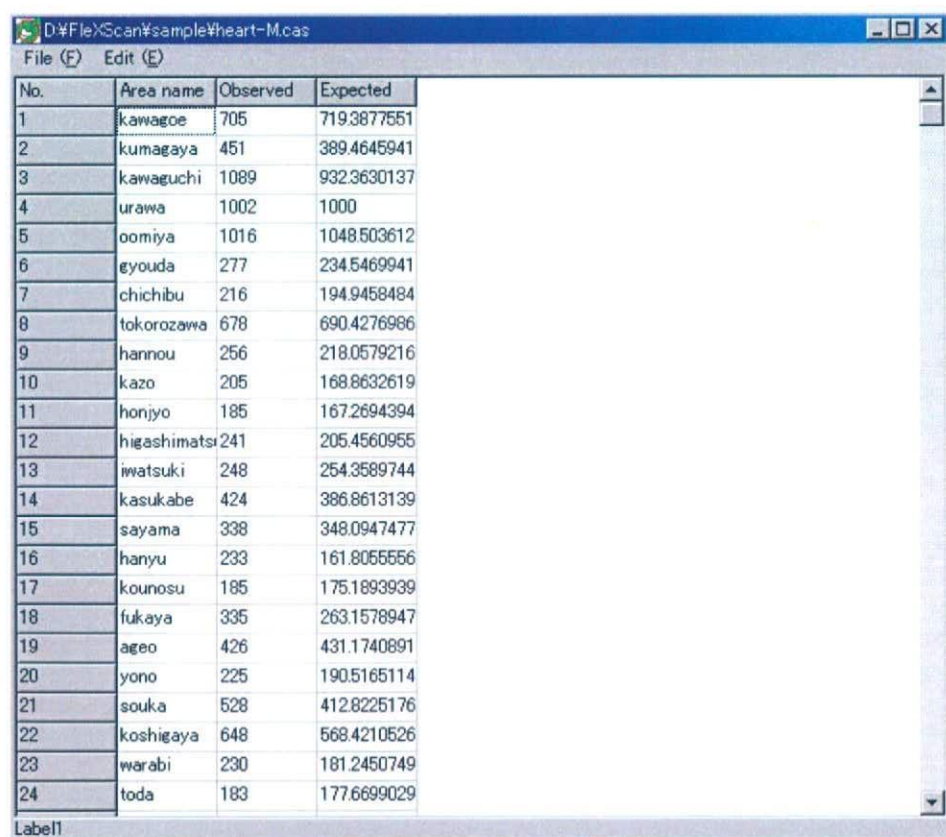
Poisson モデルでの解析を行います。

◎ 各市区町村の人口を用いたモデルで解析を行う場合

- データファイル：Case File (observed # and population #) (cas)
  - Format： <市区町村名> <観測度数> <対象者数>

各地域の対象者数（人口など）とその中での観測数の割合を用います。

二項モデル（Binomial model）での解析を行います。



No.	Area name	Observed	Expected
1	kawagoe	705	719.3877551
2	kumagaya	451	389.4645941
3	kawaguchi	1089	932.3630137
4	urawa	1002	1000
5	oomiya	1016	1048.503612
6	gyouda	277	234.5469941
7	chichibu	216	194.9458484
8	tokorozawa	678	690.4276986
9	hannou	256	218.0579216
10	kazo	205	168.8632619
11	honjyo	185	167.2694394
12	higashimats	241	205.4560955
13	iwatsuki	248	254.3589744
14	kasukabe	424	386.8613139
15	sayama	338	348.0947477
16	hanyu	233	161.8055556
17	kounosu	185	175.1893939
18	fukaya	335	263.1578947
19	ageo	426	431.1740891
20	yono	225	190.5165114
21	souka	528	412.8225176
22	koshigaya	648	568.4210526
23	warabi	230	181.2450749
24	toda	183	177.6699029

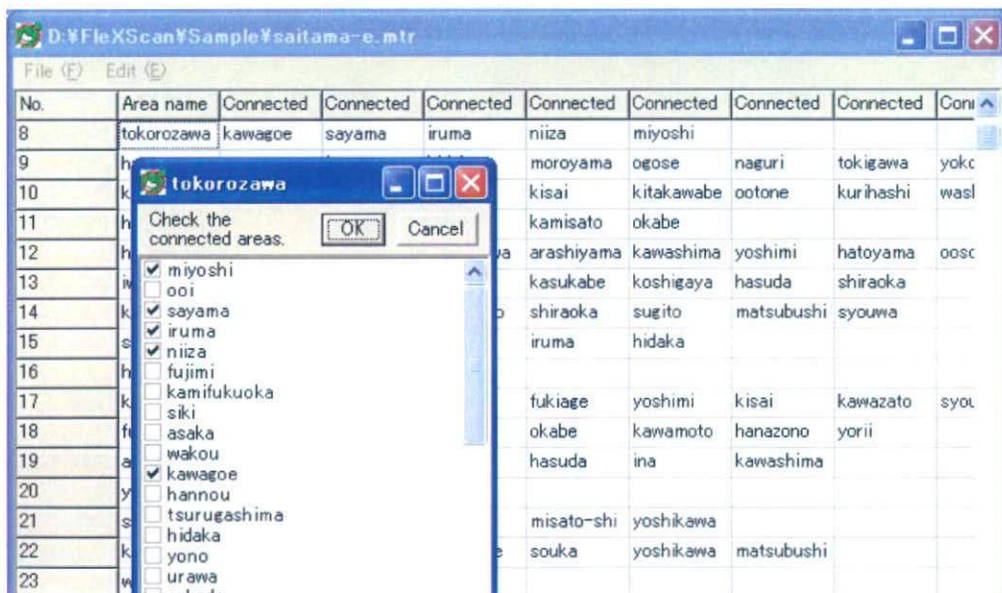
注意：

- Coordinate File、Matrix Definition File、Case File の全ての<市区町村名>はその順番も含めて完全に統一して下さい。統一されていない場合はエラーが出ます。
- データの値は「半角数字」で入力して下さい。
- Coordinate ファイルで、XY 座標を入力した場合には、「Files」タブの  
Coordinates - Cartesian  
をチェックしてください。
- Radius of Earth  
緯度・経度から距離を求める際に用いる地球の半径です。日本付近では  
6370km になります。

## ファイルの編集

FleXScan には解析に必要なファイルを編集する際に使用できるツールが用意されています。各ファイルともファイル名を入力して「Edit」ボタンをクリックすることにより編集画面が立ち上がります。Excel 等で入力して、これらの編集画面でコピー・ペーストすることもできます。

- coo ファイル（位置情報ファイル）の編集
  - 各セルに<市区町村名><緯度><経度>を入力して保存します。
- mtr ファイル（接続情報ファイル）の編集
  - まず、先に coo ファイルを完成させて下さい。
  - Area name に（coo ファイルと一致した）市区町村名を入力します。
  - 市区町村を1つ選択して「Edit」－「Area list」を選択すると、他の市区町村が近い順にリストアップされます。ここで接続している市区町村をチェックし「OK」ボタンをクリックすると自動的に Connected セルに追加されます。
  - 全て入力が済んだ後で、「File」－「Check symmetry」を選択すると対称性の確認が出来ます。もし対称でない場合（一方の接続リストに入って他方に入っていない場合）にはエラーの箇所が表示されます。



(補足情報) 実際の計算では、この mtr ファイルの情報をもとに、接続情報行列ファイル (mt0 ファイル) が自動的に作成されます。行列ファイル (mt0) から接続情報ファイル (mtr) への変換も可能です。FlexScan のメニューバー上の「Tool」に変換ツールが入っています。

- cas ファイルの編集

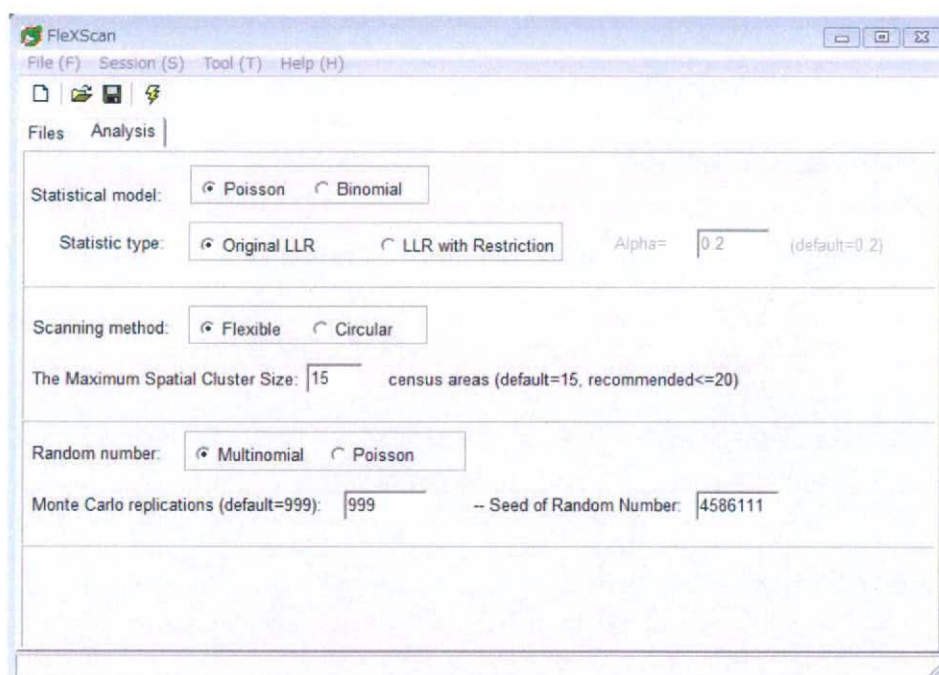
coo ファイルと同様に編集することができます。

## パラメータの値

FleXScan にはいくつかのパラメータが用意されています。それらは「Analysis」タブの項目で設定します。

- 統計モデル : Statistical model
  - ① Poisson : 用いるデータが「観測度数」と「期待度数」の場合に、その比 (O/E 比) に基づいた解析を行う Poisson モデル
  - ② Binomial : データが「観測度数」と「対象者数」(人口) の場合にその割合に基づいた解析を行う二項モデル
  
- 用いる統計量 : Statistic type
  - ① Original LLR : Kulldorff による従来の尤度比統計量  
FleXScan version 2 まで用いられていたものはこの統計量です。
  - ② LLR with Restriction : Tango による制限付尤度比統計量  
この場合制限のパラメータ Alpha を事前に定める (default は 0.2)  
この統計量を用いることで多くの地域を同定してしまうことを防ぎ、また計算時間も大幅に速くなります。詳しくは参考文献を参照下さい。
  
- 検定法の選択 : Scanning method
  - ① Flexible : Tango and Takahashi による flexible scan statistic
  - ② Circular : Kulldorff による scan statistic
  
- The Maximum Spatial Cluster Size :  
検定で用いる統計量の最大連結地域数です。この数を大きくすると広い地域を同定することができるようになりますが、計算時間が長くなります。詳しくは参考文献を参照下さい。
  
- Random number : モンテカルロシミュレーションに用いる乱数
  - ① Multinomial : 多項乱数 (観測数の総数を固定した乱数)
  - ② Poisson : ポアソン乱数 (総数を固定しない乱数。Poisson モデルで選択可)
  - ③ Binomial : 二項乱数 (総数を固定しない乱数。Binomial モデルで選択可)

- Monte Carlo replications  
検定に用いる p 値を計算するためのモンテカルロシミュレーションの回数です。例えば 999 に設定した際は、999 回のシミュレーションの値と実データからの値の  $999+1=1000$  個の統計量から p 値を求めることになります。
- Seed of Random Number  
モンテカルロシミュレーションの乱数を発生させるパラメータです。



## 参考データの入手

FleXScan を使った解析を体験するために、公開されている統計数値などを用いることができます。以下のようなものを参考にするとよいでしょう。

- 「日本の市区町村 位置情報要覧」  
国土地理院作成、(財) 日本地図センター複製発行
- 政府統計の総合窓口 (e-Stat)  
<http://www.e-stat.go.jp/>
- 厚生労働統計  
<http://mhlw.go.jp/toukei/index.html>
- 総務省統計局 統計データ  
<http://www.stat.go.jp/>
- 「統計でみる市区町村のすがた」  
電子媒体 (財) 統計情報研究開発センター
- 「住民基本台帳人口要覧」 (財) 国土地理協会

## 使用上の注意

- FleXScan の著作権は高橋邦彦、横山徹爾、丹後俊郎（以下、著作者という）が有します。
- FleXScan は非営利目的であれば誰でも自由に利用することができます。ただし FleXScan の二次配布については著作者の承諾が必要です。
- FleXScan を利用して解析を行った場合には、参考資料として FleXScan を明記して下さい。その際  
Takahashi K, Yokoyama T and Tango T. FleXScan v3.0: Software for the Flexible Scan Statistic. National Institute of Public Health, Japan, 2009.  
のように引用して下さい。
- FleXScan は予告なしにバージョンアップを行います。最新版は国立保健医療科学院技術評価部ホームページ ([http://www.niph.go.jp/soshiki/gijutsu/index\\_j.html](http://www.niph.go.jp/soshiki/gijutsu/index_j.html)) から入手できます。最近の User Guide、サンプルデータの配布、最新の情報などは全てホームページ上で行います。詳しくはホームページをご覧ください。



## 研究成果の刊行に関する一覧表

研究成果の刊行に関する一覧表

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書籍名	出版社名	出版地	出版年	ページ
丹後俊郎 横山徹爾 高橋邦彦		丹後俊郎	空間疫学への招待	朝倉書店	東京	2007	全225ページ

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
Takahashi K, Kulldorff M, Tango T, Yih K.	A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring.	International Journal of Health Geographics	7:14		2008
Tango T	A spatial scan statistic with a restricted likelihood ratio.	Japanese Journal of Biometrics	29	75-95	2008
高橋邦彦、横山徹爾、丹後俊郎	疾病地図から疾病集積性へ。	保健医療科学	57:2	86-92	2008
相田潤、森田学、安藤雄一、丹後俊郎、高橋邦彦、青山旬、小坂健	歯科疾患の地域差の検討。	保健医療科学	57:2	92-98	2008
高橋邦彦、丹後俊郎	疾病集積性の検定を用いた症候サーベイランス解析	保健医療科学	57:2	122-129	2008
郡山一明、片岡裕介、竹中ゆかり、浅見泰司、高橋邦彦、丹後俊郎	健康危機管理と小学校欠席サーベイランス	保健医療科学	57:2	130-136	2008

研究成果の刊行物・別刷

## A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring

Kunihiko Takahashi\*<sup>1</sup>, Martin Kulldorff<sup>2</sup>, Toshiro Tango<sup>1</sup> and Katherine Yih<sup>2</sup>

Address: <sup>1</sup>Department of Technology Assessment and Biostatistics, National Institute of Public Health, Japan and <sup>2</sup>Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, Boston, USA

Email: Kunihiko Takahashi\* - [kunihiko@niph.go.jp](mailto:kunihiko@niph.go.jp); Martin Kulldorff - [martin\\_kulldorff@hms.harvard.edu](mailto:martin_kulldorff@hms.harvard.edu); Toshiro Tango - [tango@niph.go.jp](mailto:tango@niph.go.jp); Katherine Yih - [Katherine\\_Yih@harvardpilgrim.org](mailto:Katherine_Yih@harvardpilgrim.org)

\* Corresponding author

Published: 11 April 2008

Received: 28 November 2007

*International Journal of Health Geographics* 2008, **7**:14 doi:10.1186/1476-072X-7-14

Accepted: 11 April 2008

This article is available from: <http://www.ij-healthgeographics.com/content/7/1/14>

© 2008 Takahashi et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Early detection of disease outbreaks enables public health officials to implement disease control and prevention measures at the earliest possible time. A time periodic geographical disease surveillance system based on a cylindrical space-time scan statistic has been used extensively for disease surveillance along with the SaTScan software. In the purely spatial setting, many different methods have been proposed to detect spatial disease clusters. In particular, some spatial scan statistics are aimed at detecting irregularly shaped clusters which may not be detected by the circular spatial scan statistic.

**Results:** Based on the *flexible purely spatial scan statistic*, we propose a flexibly shaped space-time scan statistic for early detection of disease outbreaks. The performance of the proposed space-time scan statistic is compared with that of the cylindrical scan statistic using benchmark data. In order to compare their performances, we have developed a space-time power distribution by extending the purely spatial bivariate power distribution. Daily syndromic surveillance data in Massachusetts, USA, are used to illustrate the proposed test statistic.

**Conclusion:** The flexible space-time scan statistic is well suited for detecting and monitoring disease outbreaks in irregularly shaped areas.

### Background

The anthrax terrorist attacks in 2001, the severe acute respiratory syndrome (SARS) outbreak in 2002, and a concern about pandemic influenza have motivated many public health departments to develop early disease outbreak detection systems. Early detection of disease outbreaks enables public health officials to implement disease control and prevention measures at the earliest possible time. For an infectious disease, improvement in detection time by even one day might enable public health officials to control the disease before it becomes

widespread. In many cities such as New York City [1], Washington, D.C. [2], Boston [3,4], Denver, and Minneapolis, real-time, geographic, early outbreak detection system have been implemented. For a well-defined geographical area, standard disease surveillance uses purely temporal methods that seek anomalies in time series data without using spatial information [5]. The increased need for geographical cluster detection has coincided with an increasing availability of spatial data [6]. Investigators ask whether the geographical cluster is unlikely to have arisen by chance given random variations

from the background incidence, according for the multiple comparisons inherent in the many possible cluster locations and size evaluated. Scan statistics are tools to answer such questions [7,8]. Increasingly, there is interest in the prospective surveillance of new data as it becomes available in order to detect a localized disease outbreak as early as possible. Particularly in light of the perceived threat of bioterrorism and newly emerging infectious diseases, there has been a spate of recent interest in the development of geographic surveillance systems that can detect changes in spatial patterns of disease [9]. Recently, a time periodic geographical disease surveillance system based on a cylindrical space-time scan statistic was proposed by Kulldorff and colleagues [10,11].

Several different approaches to the statistical assessment of potential geographic clustering in either point-or area-based disease data have been developed [12,13]. Almost all of these purely spatial approaches are retrospective, in the sense that they describe statistical tests that are designed to be carried out once, on a set of data that has been collected from the recent past [9]. In particular, the circular spatial scan statistic [8] has been used extensively for the detections and evaluation of purely spatial disease clusters along with the SaTScan software [14]. For example, as part of their cancer surveillance initiative, the New York State Department of Health used the spatial scan statistic to look at the geographical variation of breast, lung, prostate, and colorectal cancer incidence in New York State, finding various statistically significant clusters but no local hotspots with greatly elevated risk [15]. However, as the statistic uses a circular scanning window with variable size to define the potential cluster area, it is difficult to correctly detect some non-circular clusters such as those along a river [16]. Recently, spatial scan statistics for irregular shaped clusters have been proposed, using the same likelihood ratio test formulation as before. The spatial scan statistics proposed by Duczmal and Assunção [17], Patil and Taillie [18], Tango and Takahashi [16], Assunção *et al.* [19] and Kulldorff *et al.* [20] are aimed at detecting irregularly shaped clusters which may not be detected by the circular spatial scan statistic. Due to the unlimited geometric freedom of cluster shapes, some of these statistics run the risk of detecting quite large and very peculiarly shaped clusters. The *flexible spatial scan statistic* [16], which has been used along with the FlexScan software [21], has a parameter  $K$  as the pre-set maximum length of neighbors to be scanned, to avoid detecting clusters with a very peculiar shape.

In this paper, we propose a flexibly shaped space-time scan statistic ("flexible space-time scan statistic" hereafter) for the early detection of disease outbreaks. It is based on the flexible purely spatial scan statistic [16] and the prospective space-time scan statistic [10]. The performance of

our proposed space-time scan statistic is compared with that of the cylindrical scan statistic, using the benchmark data provided by Kulldorff *et al.* [22]. In order to evaluate its performance we propose a space-time power distribution by extending the purely spatial bivariate power distribution [16]. Daily syndromic surveillance data in Massachusetts, USA, are used to illustrate the proposed method with real data.

#### The flexible space-time scan statistic

Consider the situation where an entire study area is divided into  $m$  regions (for example, counties, ZIP codes, enumeration districts, etcetera), and each region is periodically reporting the number of cases of a disease or syndrome under study. We assume that, under the null hypothesis of no clustering, the number of cases  $N_{id}$  is a Poisson random variable with the observed value  $n_{id}$  and the expected values  $\mu_{id}$  in each region  $i$  ( $i = 1, \dots, m$ ) at time  $d$ , where  $\mu_{id}$  is proportional to its population size, or a covariate-adjusted population at risk. Since we are only interested in detecting clusters that are alive (active) at the current time  $t_p$ , we only consider 'alive' clusters that are present in the following  $T$  time intervals:

$$[t_p - T + 1, t_p], [t_p - T + 2, t_p], \dots, [t_p - 1, t_p], [t_p, t_p]$$

where  $T$  is a pre-specified maximum temporal length of the cluster.

A time periodic geographical disease surveillance system based on a *cylindrical space-time scan statistic* has already been proposed by Kulldorff [10]. The cylindrical space-time scan statistic uses a cylindrical window in three dimensions where the base of the cylinder represents space and the height represents time. As with the purely spatial scan statistic, the cylindrical space-time scan statistic imposes a circular base  $Z$  on each centroid of regions for each of  $T$  time intervals. For each of centroids, the radius of the circle is varied from zero up to a pre-set maximum radius, for example, so that the window never includes more than 50% of the total population at risk [8]. In this paper, we use a pre-set maximum number of regions  $K$  to be included in the cluster as an upperbound of the radius. If the base contains the centroid of a region, then that whole region is included in the base. In total, a very large number of different but overlapping circular bases are created, each with a different set of neighboring regions and each being a possible candidate area containing a disease outbreak. Let  $Z_{ik}$ ,  $k = 1, \dots, K$  denote the base composed by the region  $i$  and the  $(k - 1)$ -nearest neighbors to  $i$ . Then, all the cylindrical windows to be scanned by the cylindrical scan statistic are the cylinders with the base in the set

$$Z_1 = \{Z_{ik} \mid 1 \leq i \leq m, 1 \leq k \leq K\} \quad (1)$$

and the heights in the set

$$\mathcal{Y} = \{[t_p - t + 1, t_p] \mid 1 \leq t \leq T\}, \tag{2}$$

On the other hand, a *flexible space-time scan statistic* which we propose in this paper imposes a three dimensional prismatic window with an arbitrarily shaped base  $Z$ . For any given region  $i$ , we create the set of arbitrarily shaped bases consisting of  $k$  connected regions ( $1 \leq k \leq K$ ) including  $i$ . To avoid detecting a cluster of unlikely peculiar shape, the connected regions are restricted as the subset of the  $K$ -nearest neighbors to the region  $i$ , where  $K=1$  implies the region  $i$  itself. Let  $Z_{ik(j)}$ ,  $j=1, \dots, j_{ik}$  denote the  $j$ -th window which is a set of  $k$  regions connected starting from the region  $i$ , where  $j_{ik}$  is the number of  $j$  satisfying  $Z_{ik(j)} \subseteq Z_{ik}$  for  $k=1, \dots, K$ . Then, all the windows to be scanned are the prisms whose base is included in the set

$$Z_2 = \{Z_{ik(j)} \mid 1 \leq i \leq m, 1 \leq k \leq K, 1 \leq j \leq j_{ik}\} \tag{3}$$

with height in the set  $\mathcal{Y}$ . In other words, for any given region  $i$ , the cylindrical scan statistic consider  $K$  concentric circles for the base, whereas the flexible scan statistic consider  $K$  concentric circles plus all the sets of connected regions including the single region  $i$ , whose centroids are located within the  $K$ -th largest concentric circle.

Define  $L(W)$  as the likelihood under the alternative hypothesis that there is a cluster in the space-time window  $W \in \mathcal{W}$ , where  $\mathcal{W} = Z_1 \times \mathcal{Y}$  (or  $Z_2 \times \mathcal{Y}$ ) and  $L_0$  the likelihood under the null hypothesis. Then, conditioning on the observed total number of cases,  $N$ , the definition of the space-time scan statistic  $S$  is the maximum likelihood ratio over all possible windows  $W$ ,

$$S = \frac{\max_{W \in \mathcal{W}} \{L(W)\}}{L_0} = \max_{W \in \mathcal{W}} \left\{ \frac{L(W)}{L_0} \right\}. \tag{4}$$

Let  $n_W$  be the number of cases in window  $W$ . For the Poisson model, let  $\mu_W$  be the expected number in window  $W$  under the null hypothesis, so that  $\mu_G = N$  for  $G$ , the entire study space in three dimensions. It can then be shown that

$$\frac{L(W)}{L_0} = \left( \frac{n_W}{\mu_W} \right)^{n_W} \left( \frac{N - n_W}{N - \mu_W} \right)^{N - n_W} \tag{5}$$

if  $n_W > \mu_W$  and  $L(W)/L_0 = 1$  otherwise. The window for which the likelihood ratio is maximized identifies the most likely cluster (MLC) [8]. To find the distribution of the log likelihood ratio (LLR) under the null hypothesis, Monte Carlo hypothesis testing [23] is required.  $p$ -value of the test is based upon the null distribution of LLR with

large number  $B$  of Monte Carlo replications of data sets generated under the null hypothesis, i.e.,

$$\hat{p} = \frac{1 + \sum_{v=1}^B I(LLR_v \geq LLR^*)}{B+1}$$

where  $LLR_v$  and  $LLR^*$  is the value of the test statistic for the  $v$ -th Monte Carlo replicate and that for the observed data, respectively, and  $I(\cdot)$  is the indicator function.

**Syndromic surveillance in Massachusetts**

We applied the prospective flexible space-time scan statistic to daily syndromic surveillance data in eastern Massachusetts mimicking a real time surveillance system. The data came from an electronic medical record system used by Harvard Vanguard Medical Associates [3,24]. We used the rash and respiratory data during August 1–30, 2005. The data are geographically aggregated to ZIP codes. The number of ZIP codes used were different for each syndrome, for example cases of the rash were analyzed in 252 ZIP codes and respiratory in 385. Note that for the flexible space-time scan statistic, the ZIP code whose data does not exist, was treated like a ravine. For example, assume that ZIP codes  $i_1$  and  $i_2$ ,  $i_2$  and  $i_3$  are adjacent each other, respectively, but  $i_1$  and  $i_3$  are not adjacent. If the data of  $i_2$  does not exist under the situation, then it is assumed that  $i_1$  and  $i_3$  are not directly connected.

Based on the prior daily data for over a year in MA, the expected number of cases were calculated as the predicted means from a generalized linear mixed model (GLMM) as developed by Kleinman *et al*, adjusted for seasonal effect, day of week, etc, these are the same expectations used in the actual real time surveillance system [25]. We set  $K=20$  as the maximum length of the geographical window, and the maximum temporal length to be  $T=7$  days. The number of replications for the Monte Carlo procedure was set to  $B=999$ . In disease outbreak detection, the recurrence interval (RI) is often used as an alternative to the  $p$ -value [14]. The measure reflects how often a cluster will be observed by chance, assuming that analyzes are repeated on a regular basis with a periodicity equal to the period of the study. For daily surveillance such as this analysis, the  $p$ -value of 0.001 corresponds to the RI of 1,000 days, i.e., 2.7 years, and an alpha level of 0.0027 corresponds to one expected false alarm every year.

The results of analysis during August 1–30 by the flexible and the cylindrical space-time scan statistics are given in Tables 1, 2 and Figure 1. The tables show results for the days with  $p < 0.0054$ , which corresponds to the RI of at least 6 months. When looking at rash outbreaks (Table 1), both tests detected the same cluster with a single ZIP code 01951 on August 7, with the same temporal length (6 days) and the same RI (2.7 years). Note that the clusters

**Table 1: Detected outbreaks of Rash based on daily syndromic surveillance data in eastern Massachusetts during August 1–30, 2005.**

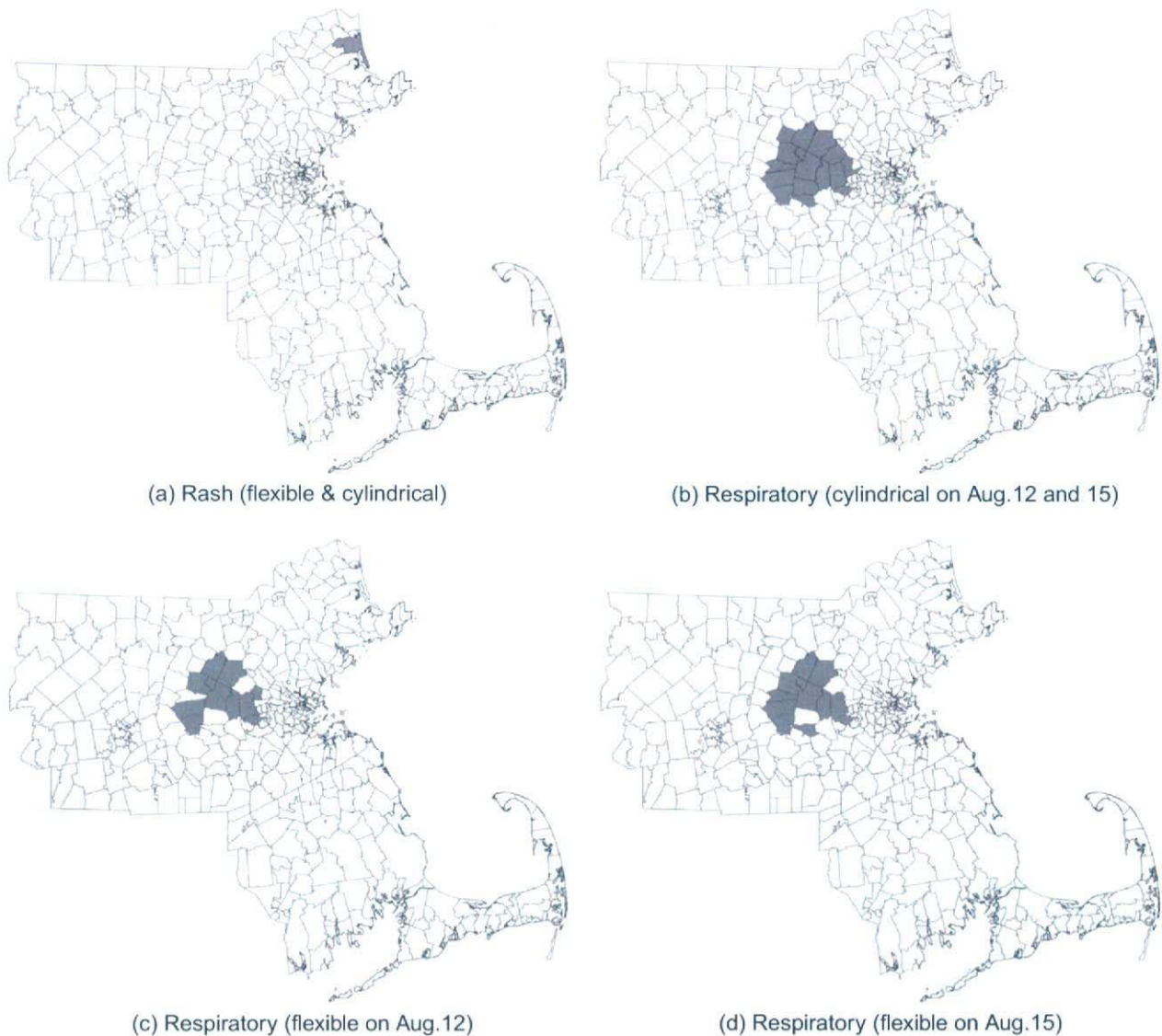
Day	zip codes	cluster period	cases	expected	llr	R.I.(p-value)
<b>Rash:</b>						
<b>- flexible</b>						
Aug.07	01951	Aug.02–07	7	0.0427	27.949	2.7 years(0.001)
Aug.08	01951	Aug.02–08	7	0.0545	26.259	2.7 years(0.001)
Aug.09	01951	Aug.03–09	6	0.0545	21.562	2.7 years(0.001)
Aug.10	01951	Aug.04–10	5	0.0545	17.315	2.7 years(0.001)
<b>- cylindrical</b>						
Aug.07	01951	Aug.02–07	7	0.0427	27.949	2.7 years(0.001)
Aug.08	01951	Aug.02–08	7	0.0545	26.259	2.7 years(0.001)
Aug.09	01951	Aug.03–09	6	0.0545	21.562	2.7 years(0.001)
Aug.10	01951	Aug.04–10	5	0.0545	17.315	2.7 years(0.001)

detected by both tests from August 8 to 10 are not signals of an outbreak because the number of cases on August 8 must be 0, and on August 9 and 10, the number of cases of the cluster was decreasing. For respiratory syndrome (Table 2), each test detected a different cluster with the same RI of 2.7 years on August 12. The cluster detected by the flexible scan statistic contained 12 ZIP codes, while that from the cylindrical scan statistic contained 18 ZIP codes, with 11 ZIP codes detected in common. On August 13 and 14, the flexible scan statistic detected significant clusters with larger RIs, 333 days and 250 days respectively, while the cylindrical scan statistic detected clusters with short RIs, 91 days and 30 days respectively. The flexible scan statistic also detected a cluster on August 15 (RI = 1.4 years) with a temporal length of 6 days, while the

cylindrical scan statistic detected a cluster with a temporal length of 5 days (RI = 200 days). For the 6 days from August 12 to 17 (results on August 16 and 17 are not shown in Table 2 because of shorter RIs), the cylindrical scan statistic kept detecting the same cluster, while the flexible scan statistic detected a similar but slightly different cluster each day. However, we should acknowledge the similar lack of evidence in Table 2 for a continued outbreak on August 13 to 14, because the number of additional cases on those days is very close to the expected number of additional cases. On the other hand, there is some evidence for an excess of cases on August 15 (23 additional cases), although the estimated relative risk is substantially reduced.

**Table 2: Detected outbreaks of Respiratory based on daily syndromic surveillance data in eastern Massachusetts during August 1–30, 2005.**

Day	zip codes	cluster period	cases	expected	llr	R.I.(p-value)
<b>Respiratory:</b>						
<b>- flexible</b>						
Aug.12	01720, 01742, 01752, 01754, 01772, 01775, 01776, 01778, 02451, 02462, 02481, 02493	Aug.11–12	42	12.452	17.635	2.7 years (0.001)
Aug.13	01720, 01742, 01749, 01752, 01754, 01772, 01775, 01776, 01778, 02451, 02462, 02481, 02493	Aug.11–13	46	14.950	16.634	333 days (0.003)
Aug.14	01720, 01742, 01749, 01752, 01754, 01772, 01775, 01776, 01778, 02451, 02462, 02481, 02493	Aug.11–14	49	16.957	15.927	250 days (0.004)
Aug.15	01702, 01720, 01742, 01749, 01752, 01754, 01772, 01775, 01776, 01778, 02481, 02493	Aug.10–15	72	29.975	16.726	1.4 years (0.002)
<b>- cylindrical</b>						
Aug.12	01701, 01702, 01718, 01719, 01720, 01742, 01749, 01752, 01754, 01772, 01773, 01775, 01776, 01778, 02451, 02453, 02481, 02493	Aug.11–12	51	20.036	12.688	2.7 years (0.001)
Aug.13	01701, 01702, 01718, 01719, 01720, 01742, 01749, 01752, 01754, 01772, 01773, 01775, 01776, 01778, 02451, 02453, 02481, 02493	Aug.11–13	55	23.768	10.945	91 days (0.011)
Aug.14	01701, 01702, 01718, 01719, 01720, 01742, 01749, 01752, 01754, 01772, 01773, 01775, 01776, 01778, 02451, 02453, 02481, 02493	Aug.11–14	59	26.959	10.221	30 days (0.033)
Aug.15	01701, 01702, 01718, 01719, 01720, 01742, 01749, 01752, 01754, 01772, 01773, 01775, 01776, 01778, 02451, 02453, 02481, 02493	Aug.11–15	82	40.981	11.662	200 days (0.005)



**Figure 1**  
**Detected outbreaks of Rash and Reepiratory in eastern Massachusetts during August 1–30, 2005, by the cylindrical scan statistic ((a) and (b)) and the flexible scan statistic ((a), (c) and (d)).**

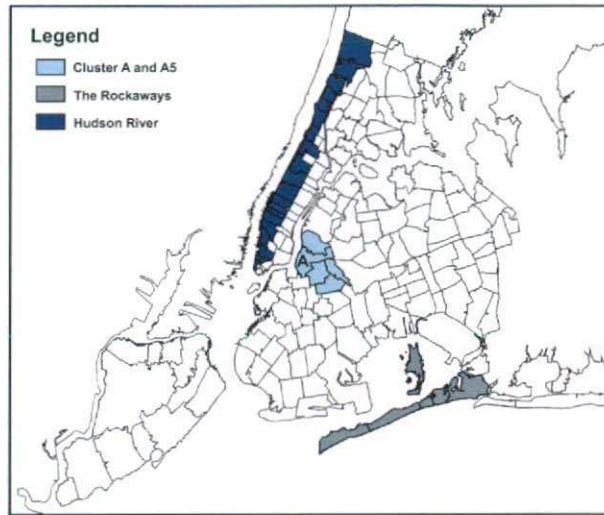
**Statistical power, sensitivity and positive predictive value**

In this section, we compare the flexible and cylindrical space-time scan statistics, using benchmark data from 176 New York City ZIP codes ([14,22]). This benchmark data has been described in detail elsewhere [22], and here we only give a brief overview. Based on 2002 numbers, the total population is 8,003,510. The benchmark data sets contain a number of randomly located cases of a hypothetical disease or syndrome, generated either under the null model with no outbreaks or under one of eight differ-

ent alternative models with an outbreak in one of four different locations and with either a high or modest excess risk. For each of the null and alternative models, three different sets of data sets were generated, with 31, 32, and 33 days, respectively. For each of the null models, 9,999 random data sets were generated. For each of the alternative models, 1,000 random data sets were generated.

For each data set, the total number of randomly allocated cases was 100 times the number of days (i.e., 3,100 cases





**Figure 2**  
**NYC 176 ZIP codes area and assumed clusters (i) Cluster A, (ii) Cluster A5, (iii) The Rockaways, and (iv) Hudson River.**

in the data sets containing 31 days). The number 100 was chosen to reflect the occurrence rate of certain syndromes common to the NYC emergency department(ED)-based syndromic surveillance system. Under the null model, each person living in NYC is equally likely to contract the disease, and the time of each case is assigned with equal probability to any given day. Thus, each case was randomly assigned to ZIP code  $i$  and day  $d$  with probability proportional to  $\mu_{id} = pop_i$ , where  $pop_i$  is the population of ZIP code  $i$ . For the alternative models, one or more ZIP codes were assigned an increased risk on Day 31 and, when applicable, on Days 32 and 33 as well. For these ZIP code and day combinations,  $\mu_{id}$  was multiplied by an assigned relative risk. For all other ZIP code and day combinations,  $\mu_{id}$  did not change. Each case was then randomly assigned with probability proportional to the new set of  $\mu_{id}$  to generate data under the alternative models.

Eight alternative models were evaluated, based on four different outbreak areas of length  $s^*$  and total population  $pop^*$  therein, with either high or medium relative risk (RR) [22] (Figure 2).

1. Cluster A: a single ZIP code area in Brooklyn (circular area)

$s^* = 1$ ,  $pop^* = 85,089$ , RR: high = 9.91, medium = 5.66

2. Cluster A5: the same ZIP code with 4 neighboring ZIP codes (non-circular area)

$s^* = 5$ ,  $pop^* = 318,754$ , RR: high = 4.47, medium = 3.06

3. The Rockaways, 5 ZIP codes area (non-circular area)

$s^* = 5$ ,  $pop^* = 106,738$ , RR: high = 8.48, medium = 5.01

4. Hudson River: 20 ZIP codes areas along the shore of the Hudson River (non-circular area)

$s^* = 20$ ,  $pop^* = 827,382$ , RR: high = 2.97, medium = 2.24

A maximum length of the geographic window  $K = 20$  was used for the flexible scan statistic, while the cylindrical scan statistic used a maximum of either  $K = 20$  or a 50 % of the population at risk. A period of  $T = 3$  days was used as the maximum temporal length of the cluster. We did not use the options to include purely temporal clusters (see details in [14]).

#### Standard statistical power

First of all, we estimated the standard statistical power, which is the probability that the null hypothesis is rejected at the  $\alpha = 0.05$  significance level, without considering the overlap between the detected and real clusters. The random data sets generated under the null model were used to get the critical values of the scan statistics. For  $\alpha = 0.05$ , this is defined as the 500th highest log likelihood ratio when ranking those value from all the 9,999 simulated data sets. The estimated power was then calculated as the proportion of the 1,000 random data sets that had a higher log likelihood ratio than the critical value obtained from the null data sets. The results are shown in Table 3. In general, the cylindrical space-time scan statistic has higher power for the three more compact clusters, while the flexible space-time scan statistic have higher power for the long and narrow the Hudson River cluster. On Day 33 of the high excess risk outbreaks, both methods have very high power.

#### Space-time power distribution

In order to compare the performance of the cluster detection tests, the standard power has been derived in the same manner as for usual hypothesis tests. However, it should be noted that standard statistical power reflect 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 purely spatial cluster detection tests, Tango and Takahashi [16] proposed a spatial bivariate power distribution  $P_0(I, s | s^*)$  based on Monte Carlo simulation where  $I$  is the length of the significant MLC, while  $s$  is the number of regions identified out of the true cluster with  $s^*$  regions.

**Table 3: Standard power of the prospective space-time scan statistics – flexible and cylindrical – at different days of the outbreak**

Outbreak areas	No. of zip codes $s^*$	excess risk	Power on Day 31			Power on Day 32			Power on Day 33		
			flex. $K = 20$	cylind. $K = 20$	cylind. 50% pop	flex. $K = 20$	cylind. $K = 20$	cylind. 50% pop	flex. $K = 20$	cylind. $K = 20$	cylind. 50% pop
Cluster A	1	high	0.764	0.860	0.862	0.988	0.996	0.996	0.999	0.999	0.999
Cluster A5	5	high	0.797	0.850	0.847	0.994	0.996	0.996	1.000	1.000	1.000
The Rockaways	5	high	0.769	0.855	0.840	0.992	0.997	0.997	1.000	1.000	1.000
Hudson River	20	high	0.656	0.597	0.632	0.964	0.933	0.949	0.998	0.994	0.995
Cluster A	1	med.	0.272	0.357	0.357	0.651	0.733	0.737	0.844	0.915	0.916
Cluster A5	5	med.	0.382	0.435	0.428	0.752	0.801	0.795	0.914	0.940	0.941
The Rockaways	5	med.	0.261	0.373	0.348	0.648	0.768	0.759	0.848	0.924	0.917
Hudson River	20	med.	0.290	0.257	0.297	0.631	0.582	0.610	0.845	0.782	0.803

$$P_0(l, s, | s^*) = \frac{\Pr\{L = l, S = s | s^*\}}{\#\{\text{significant MLC has length } l \text{ and includes } s \text{ true regions}\} / \#\{\text{trials for each simulation}\}} \tag{6}$$

where  $L$  and  $S$  denote the random variable of  $l$  and  $s$  under the specified model, respectively, and  $l \geq 1$  and  $0 \leq s \leq s^*$ . In a similar manner, we propose a space-time  $tti$ -variate power distribution for a space-time cluster detection test based on Monte Carlo simulation where the temporal length of the true cluster is denoted  $t^*$ :

$$P_3(l, s, t | s^*, t^*) = \frac{\Pr\{L = l, S = s, U = t | s^*, t^*\}}{\#\{\text{significant MLC has geographical length } l \text{ and includes } s \text{ true regions with temporal length } t\} / \#\{\text{trials for each simulation}\}} \tag{7}$$

where  $U$  denotes the random variable of  $t$  and  $1 \leq t \leq T$ .

In Tables 4, 5 and 6, we show the estimated tri-variate power distribution  $P(l, s, t | s^*, t^*) \times 1,000$  for (a) Cluster A ( $s^* = 1$ ) on Day 31 ( $t^* = 1$ ) (b) Cluster A5 ( $s^* = 5$ ) on Day 33 ( $t^* = 3$ ) and (c) the Rockaways cluster ( $s^* = 5$ ) on Day 33 ( $t^* = 3$ ), in all cases with high excess risk.

This tri-variate power distribution provides us with a detailed description of the space-time cluster detection tests performance. For the outbreak in cluster A with a single ZIP code, the cylindrical scan statistic has higher power to detect the cluster with complete accuracy, with  $P_1(l = 1, s = 1, t = 1 | s^*, t^*) = 697/1000$ , compared to 315/1000 for the flexible. Moreover, the flexible scan statistic has a heavier tail in the  $(s, t) = (1, 3)$  column than the cylindrical one. However the cylindrical scan detected some large clusters including several with  $l \geq 15$ . For outbreaks in the non-circular shaped A5 and Rockaway clusters, the flexible scan statistic has higher power for complete accurate detection. Indeed, the cylindrical scan statistic cannot detect these clusters with complete accuracy since they are

not circular, so that the power for complete accuracy is zero. Moreover, note that for cluster A5, the flexible scan statistic is more likely to include all the five areas in the true cluster ( $797 + 12 = 809/1000$  versus  $601 + 12 = 613/1000$ ), and it is also more likely to avoid including any of the ZIP codes outside the true cluster ( $12 + 74 + 2 + 287 + 3 = 378/1000$  versus  $37 + 1 + 301 + 7 = 346/1000$ ). For the Rockaway cluster, the flexible scan statistic is again more likely to include all the five areas in the true cluster ( $667 + 4 + 1 = 672$  versus  $1 + 0 + 1 = 1$ ), but the cylindrical scan statistic avoids the ZIP codes outside the cluster more often ( $2 + 8 + 52 + 1 + 876 + 6 + 1 + 0 + 0 + 0 = 946/1000$  versus  $0 + 0 + 6 + 0 + 181 + 1 + 0 + 571 + 2 + 0 = 761/1000$ ). Tables 5 and 6 show that the temporal accuracy of the detected cluster is very good for both methods. For example, for cluster A5, the flexible scan has  $P_1(+, +, 3 | s^*, t^*) = \sum_l \sum_s P_1(l, s, 3 | s^*, t^*) = (15 + 171 + 797)/1000 = 0.983$  while the cylindrical scan has  $P_1(+, +, 3 | s^*, t^*) = (41 + 338 + 601)/1000 = 0.980$ .

The complexity of the three-dimensional tri-variate power distributions suggests that we need some summary measure. Since the temporal accuracy is very similar, we focus on the geographical accuracy. We will compute the extended power of spatial cluster detection tests, as developed by Takahashi and Tango [26]. We will also define and compute geographical sensitivity and false positive rates.

*The extended power*

We can consider two types of spatial misclassifications when applying the cluster detection test (CDT). One is a *false negative test result* (FN) in which the CDT misses a region included in the true cluster. Sensitivity is  $1 - \text{FN rate}$ . 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. The numbers of FNs and FPs for geographical detection are  $s^* - s$  and  $l - s$ , respectively.

**Table 4: Space-time power distribution  $P_1(l, s, t | s^*, t^*)$  for the Cluster A ( $s^* = 1$ ) on Day 31 ( $t^* = 1$ ) with high risk (RR= 9. 91), where  $t$  is a temporal length of detected cluster. The mark "\*" is the powers of accurate detection.**

(A) flexible (K = 20)							
length $l$ of areas	includes $s$ assumed areas						total
	0			1			
	29- $t = 3$	30- 2	31- 1	29- $t = 3$	30- 2	31- 1	
1	0	0	0	0	6	*315	321
2	0	0	0	0	1	50	51
3	0	0	0	1	2	34	37
4	0	0	0	1	6	34	41
5	0	0	0	2	5	48	55
6	1	0	0	2	2	51	56
7	1	0	0	4	13	35	53
8	0	1	0	1	12	28	42
9	1	1	2	5	6	28	43
10	0	0	0	2	7	22	31
11	1	0	0	1	4	8	14
12	1	1	2	1	0	10	15
13	0	0	0	1	1	2	4
14	0	0	0	0	0	1	1
15	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0
total	5	3	4	21	65	666	764

(B) cylindrical (K = 20)							
length $l$ of areas	includes $s$ assumed areas						total
	0			1			
	29- $t = 3$	30- 2	31- 1	29- $t = 3$	30- 2	31- 1	
1	0	0	0	5	18	*697	720
2	0	0	0	5	4	63	72
3	0	0	0	1	2	18	21
4	0	0	0	0	4	10	14
5	0	0	0	0	0	2	2
6	1	0	0	1	0	3	5
7	1	0	0	1	1	1	4
8	2	0	0	0	1	1	4
9	0	1	1	0	0	2	4
10	0	0	0	0	0	0	0
11	0	0	0	0	0	1	1
12	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0
14	0	0	0	0	0	2	2
15	0	0	0	0	1	3	4
16	0	0	0	0	0	0	0
17	0	0	0	2	0	1	3
18	1	0	0	0	0	1	2
19	0	0	0	0	0	1	1
20	0	0	0	0	0	1	1
total	5	1	1	15	31	807	860

**Table 5: Space-time power distribution  $P_1(l, s, t | s^*, t^*)$  for the Cluster A5 ( $s^* = 5$ ) on Day 33 ( $t^* = 3$ ) with high risk ( $RR = 4.47$ ), where  $t$  is a temporal length of detected cluster, and the raw all cells of which have zero powers of both tests is not shown. The mark "\*" is the powers of accurate detection.**

(A) flexible (K = 20)							
length $l$ of areas	includes $s$ assumed areas						total
	3		4		5		
	31- $t = 3$	32- 2	31- $t = 3$	32- 2	31- $t = 3$	32- 2	
1							0
2							0
3	12	0					12
4	2	0	74	2			78
5	0	0	37	2	*287	3	329
6	1	0	26	1	158	2	188
7	0	0	16	0	118	2	136
8	0	0	5	0	105	2	112
9	0	0	4	0	67	2	73
10	0	0	6	0	39	1	46
11	0	0	2	0	11	0	13
12	0	0	0	0	10	0	10
13	0	0	1	0	1	0	2
14	0	0	0	0	1	0	1
15	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0
total	15	0	171	5	797	12	1000

(B) cylindrical (K = 20)							
length $l$ of areas	includes $s$ assumed areas						total
	3		4		5		
	31- $t = 3$	32- 2	31- $t = 3$	32- 2	31- $t = 3$	32- 2	
1							0
2							0
3	37	1					38
4	2	0	301	7			310
5	2	0	32	0	*0	0	34
6	0	0	5	0	516	10	521
7	0	0	0	0	64	1	65
8	0	0	0	0	5	0	5
9	0	0	0	0	3	0	3
10	0	0	0	0	3	0	3
11	0	0	0	0	4	0	4
12	0	0	0	0	2	0	2
13	0	0	0	0	3	1	4
14	0	0	0	0	1	0	1
15	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0
total	41	1	338	7	601	12	1000