Sample Data Sets and Data Input Form

To analyze your own data using FleXScan, you need to prepare at least three data files: 1) Coordinate File, 2) Matrix Definition File, and 3) Case File. The detailed structure of each file is explained in the following section showing sample data files for Saitama prefecture in Japan (automatically installed in your 'sample' folder).

① Coordinate File (coo)

The coordinate file provides the geographic coordinates for each area. Coordinates may be specified either using the standard 'Cartesian coordinates' system or in 'latitude and longitude.' 'Cartesian' is the regular planar x,y-coordinate system. Each line of the file represents an area name (or code) and its geographical location.

For the Cartesian coordinates system

- Format : <Area name or code> <X-coordinate> <Y-coordinate>

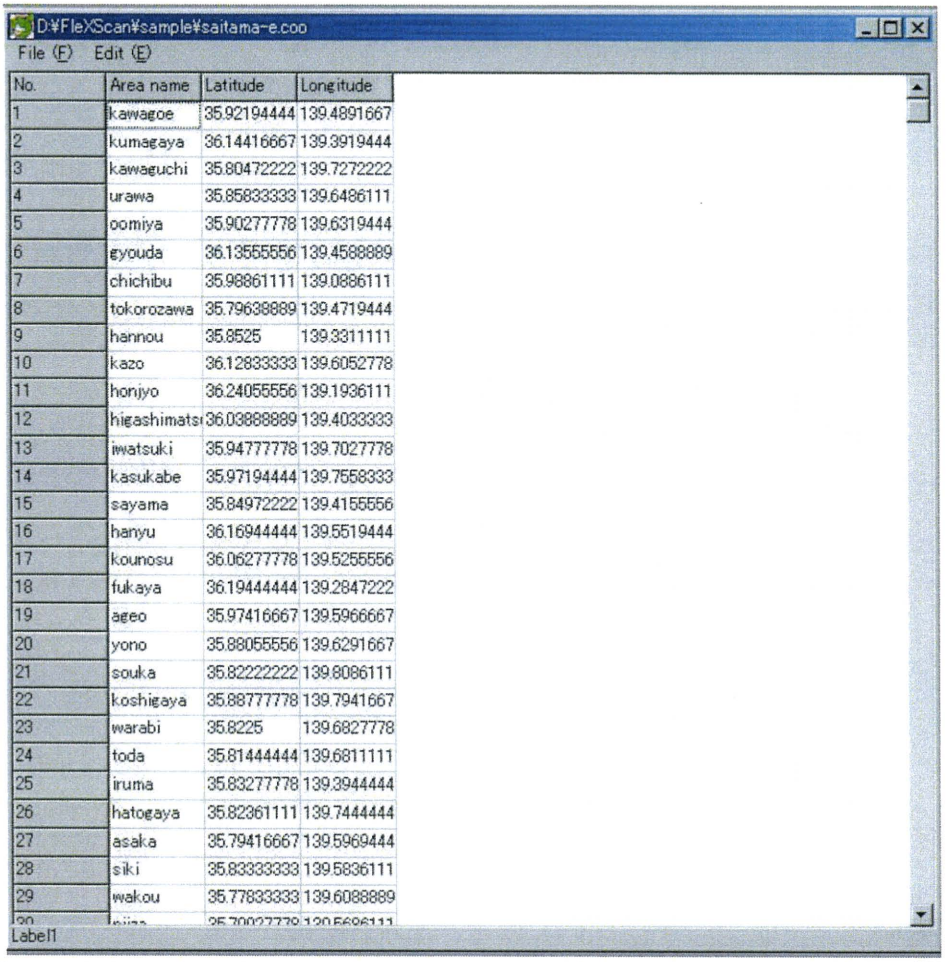
For the Latitude and Longitude coordinates system

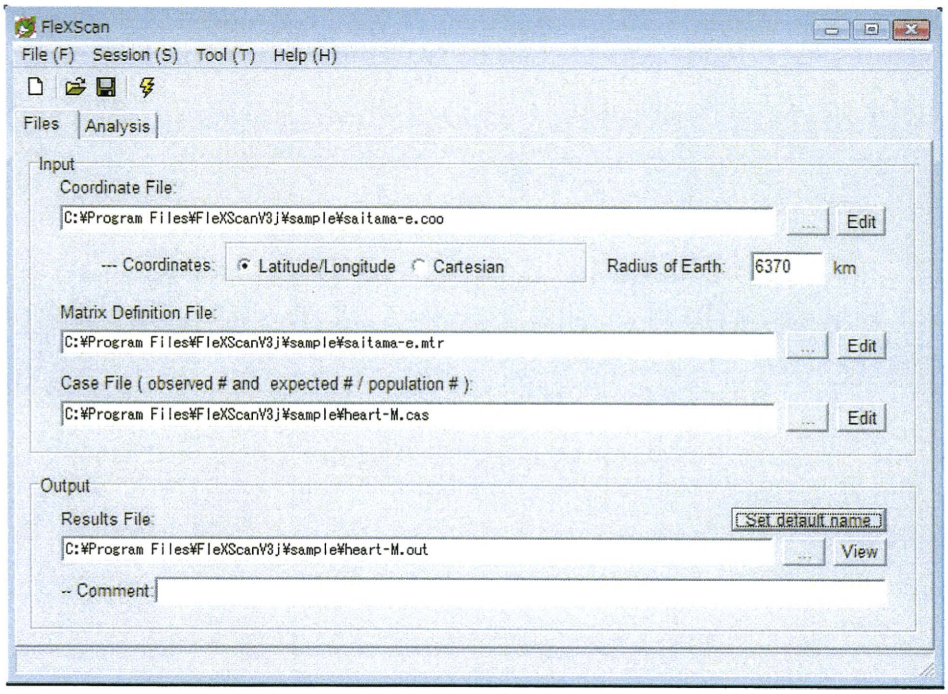
- Format : <Area name or code> <Latitude> <Longitude>

Latitudes and longitudes should be entered as decimal numbers of degrees. You can convert latitudes and longitudes expressed in degrees, minutes, and seconds to decimal number of degrees by the following formula:

xx (degrees) yy (minutes) zz (seconds) → xx + yy/60 + zz/3600 (degrees).

When coordinates are specified in latitudes and longitudes, FleXScan calculates the distance between two points on the surface of the spherical earth with a radius given in the 'Radius of Earth.'


D:¥FleXScan¥sample¥saitama-e.coo
File (F) Edit (E)
No. Area name Latitude Longitude
1 kawagoe 35.92194444 139.4891667
2 kumagaya 36.14416667 139.3919444
3 kawaguchi 35.80472222 139.7272222
4 urawa 35.85833333 139.6486111
5 oomiya 35.90277778 139.6319444
6 gyouda 36.13555556 139.4588889
7 chichibu 35.98861111 139.0886111
8 tokorozawa 35.79638889 139.4719444
9 hannou 35.8525 139.3311111
10 kazo 36.12833333 139.6052778
11 honjyo 36.24055556 139.1936111
12 higashimats 36.03888889 139.4033333
13 iwatsuki 35.94777778 139.7027778
14 kasukabe 35.97194444 139.7558333
15 sayama 35.84972222 139.4155556
16 hanyu 36.16944444 139.5519444
17 kounosu 36.06277778 139.5255556
18 fukaya 36.19444444 139.2847222
19 ageo 35.97416667 139.5966667
20 yono 35.88055556 139.6291667
21 souka 35.82222222 139.8086111
22 koshigaya 35.88777778 139.7941667
23 warabi 35.8225 139.6827778
24 toda 35.81444444 139.6811111
25 iruma 35.83277778 139.3944444
26 hatogaya 35.82361111 139.7444444
27 asaka 35.79416667 139.5969444
28 siki 35.83333333 139.5836111
29 wakou 35.77833333 139.6088889
Label1


FleXScan
File (F) Session (S) Tool (T) Help (H)
Files Analysis
Input
Coordinate File:
C:¥Program Files¥FleXScanV3j¥sample¥saitama-e.coo
Edit
--- Coordinates: Latitude/Longitude Cartesian
Radius of Earth: 6370 km
Matrix Definition File:
C:¥Program Files¥FleXScanV3j¥sample¥saitama-e.mtr
Edit
Case File (observed # and expected # / population #):
C:¥Program Files¥FleXScanV3j¥sample¥heart-M.cas
Edit
Output
Results File:
Set default name
C:¥Program Files¥FleXScanV3j¥sample¥heart-M.out
View
-- Comment:

② Matrix Definition File (mtr)

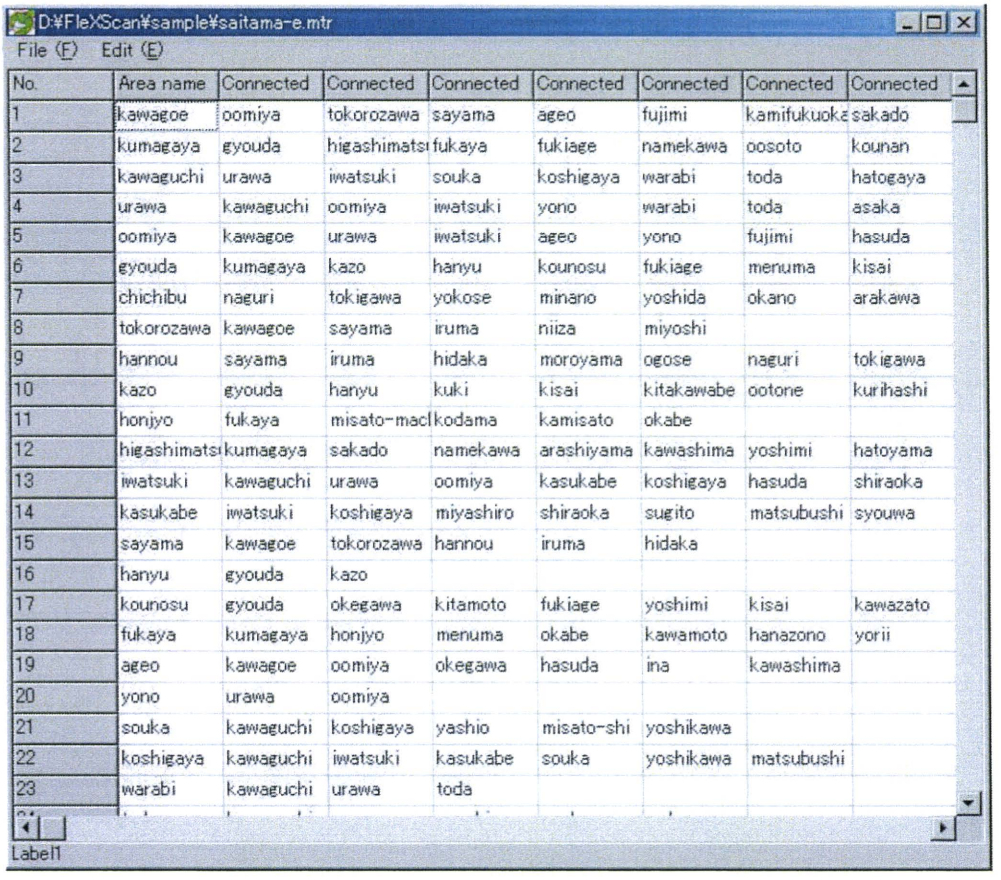
- Format : <Area name or code> <Area 1> <Area 2> …

The first column of each line is the area name, which must be identical to that in Coordinate File. The following columns specify the area name(s) that are adjacent to (i.e., border on) the area described in the first column. For example, kawagoe, sayama, iruma, niiza, and miyoshi areas are adjacent to tokorozawa (see row No.8 of the figure below).

When Area1 is adjacent to Area3 and Area5, the mtr file should be:

Area1 Area3 Area5
Area2 ….
Area3 Area1 …
…

Note that "Area1" also appears in the line of "Area3" in this case (and vice versa), and the matrix must be symmetrical, otherwise an error occurs. The 'Check symmetry' tool in the File menu is available to check the symmetry of the matrix.



No.	Area name	Connected	Connected	Connected	Connected	Connected	Connected	Connected
1	kawagoe	oomiya	tokorozawa	sayama	ageo	fujimi	kamifukuoka	sakado
2	kumagaya	gyouda	higashimatsu	fukaya	fukiage	namekawa	oosoto	kounan
3	kawaguchi	urawa	iwatsuki	souka	koshigaya	warabi	toda	hatogaya
4	urawa	kawaguchi	oomiya	iwatsuki	yono	warabi	toda	asaka
5	oomiya	kawagoe	urawa	iwatsuki	ageo	yono	fujimi	hasuda
6	gyouda	kumagaya	kazo	hanyu	kounosu	fukiage	menuma	kisai
7	chichibu	naguri	tokigawa	yokose	minano	yoshida	okano	arakawa
8	tokorozawa	kawagoe	sayama	iruma	niiza	miyoshi		
9	hannou	sayama	iruma	hidaka	moroyama	ogose	naguri	tokigawa
10	kazo	gyouda	hanyu	kuki	kisai	kitakawabe	ootone	kurihashi
11	honjyo	fukaya	misato-mac	kodama	kamisato	okabe		
12	higashimatsu	kumagaya	sakado	namekawa	arashiyama	kawashima	yoshimi	hatoyama
13	iwatsuki	kawaguchi	urawa	oomiya	kasukabe	koshigaya	hasuda	shiraoka
14	kasukabe	iwatsuki	koshigaya	miyashiro	shiraoka	sugito	matsubushi	syouwa
15	sayama	kawagoe	tokorozawa	hannou	iruma	hidaka		
16	hanyu	gyouda	kazo					
17	kounosu	gyouda	okegawa	kitamoto	fukiage	yoshimi	kisai	kawazato
18	fukaya	kumagaya	honjyo	menuma	okabe	kawamoto	hanazono	yorii
19	ageo	kawagoe	oomiya	okegawa	hasuda	ina	kawashima	
20	yono	urawa	oomiya					
21	souka	kawaguchi	koshigaya	yashio	misato-shi	yoshikawa		
22	koshigaya	kawaguchi	iwatsuki	kasukabe	souka	yoshikawa	matsubushi	
23	warabi	kawaguchi	urawa	toda				

③ Case File (cas)

The frequency of disease in each area is described in Case File. The current version of FleXScan can analyze two types of data.

① 'observed number' and 'expected number,'

➢ Format: <Area name or code> <Observed no.> <Expected no.>

For this data, 'Poisson model' should be selected in the 'Statistical model' for the analysis.

② 'observed number' and 'population,'

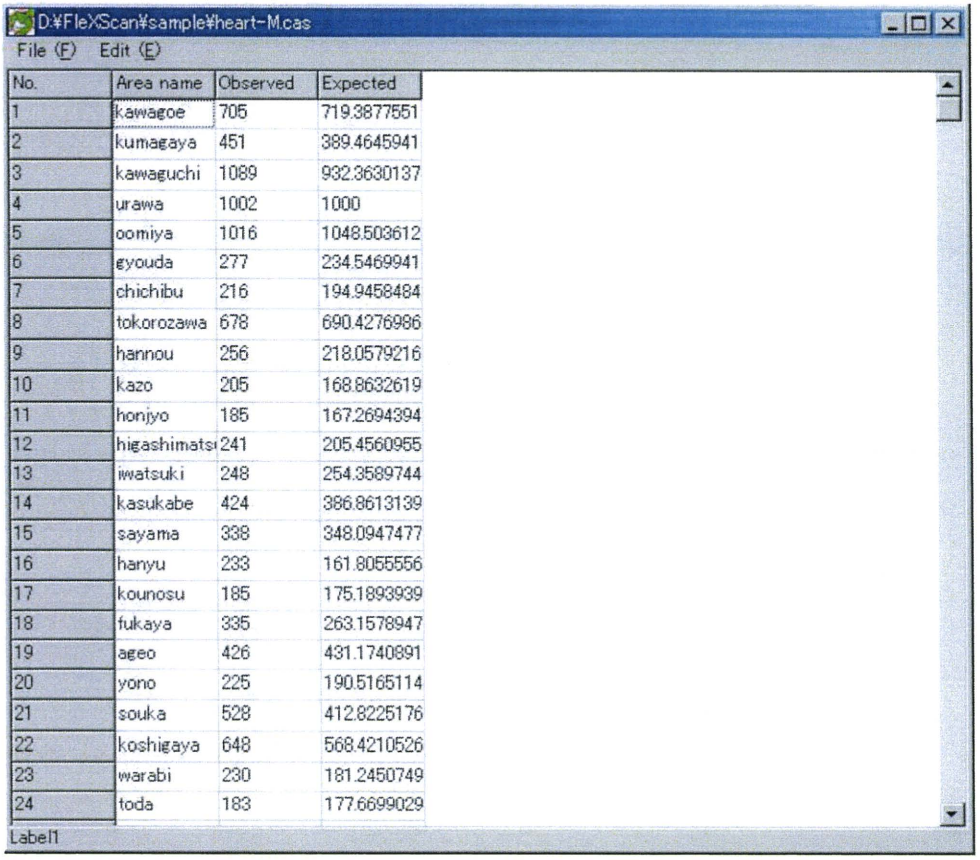
➢ Format: <Area name or code> <Observed no.> <Population >

For this data, 'Binomial model' should be selected in the 'Statistical model' for the analysis.

The first column of each line is the area name, which must be identical to that in the Coordinate File. The second column is the observed number of diseases, and the third column is the expected number of diseases under the null hypothesis, or the background population at risk in each area.

When you use the Poisson model, you need to calculate the expected number by yourself, for example, in the same manner as standardized mortality ratio (SMR). In kawagoe (see No.1 row of the figure) there were 705 deaths and the age-standardized expected number of deaths was 719.4 (i.e., SMR = 719.4 / 705 = 1.02).

But, if you do not need standardization (e.g., you are interested in the crude death rate), you can analyze the case file of 'observed number' and 'population' using the 'Poisson model.'



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

Important Note:

- All area names or codes and their order must be identical among 'Coordinate File,' 'Matrix Definition File,' 'Case File,' and 'Population File' (if necessary).

- The area name or code cannot include a space character. Use an under score or a hyphen in stead of a space character.

  Good ... 10001

  Good ... New_York

  N.G. ... New York

Editing your data set

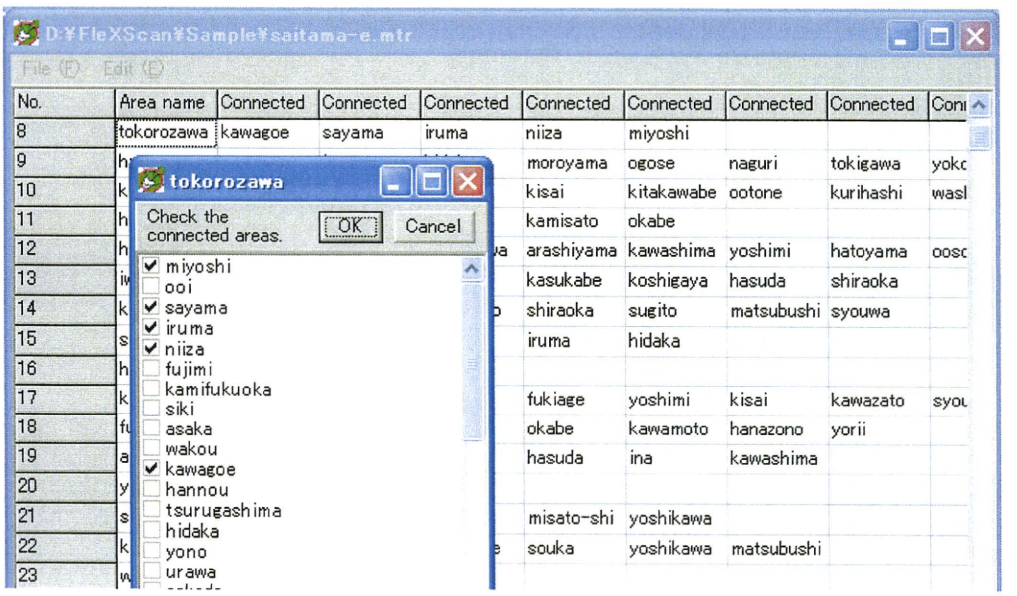
The data files 1) to 3) can be edited using FleXScan data editor. Enter the file name and click the 'Edit' button to execute the data editor. You can copy and paste your data from other software such as MS-Excel. It may be convenient to input your data on MS-Excel and copy and paste it to the FleXScan data editor.

1) Editing Coordinate File (coo).

- Input the area name and its latitude and longitude in each column.
- If x and y-coordinates are used, select 'Cartesian' on the 'Files' tab panel.
- 'Save & return' to finish editing data.

2) Editing Matrix Definition File (mtr).

- Coordinate File must be made before starting to edit Matrix Definition File.
- Input the area name, which must be identical to that of Coordinate File. It will be convenient to Copy & Paste all the area names from Coordinate File to Matrix Definition File.
- Select an area name and execute 'Edit – Area List,' then a list of area names will appear in the order of distance from the selected area. Check the check-box of areas that are adjacent to the selected area. By clicking the 'OK' button, the checked areas will be automatically added to the 'Adjacent' columns (see the figure below).



- The symmetry can be tested by executing 'File – Check symmetry.' If the information is not symmetrical (e.g., Area3 is selected as an adjacent area to Area1, but Area1 is not selected as an adjacent area to Area3), an error message will appear.

Supplementary technical information:

The Matrix Definition File is designed in a user-friendly format. FleXScan will automatically convert it to a Connection Information Matrix File (mt0 file) and utilize it for the calculation. You can see the format of the mt0 file using a text editor. Conversely, if you already have the area connection data in the format of an mt0 file, it can be converted to an mtr file by executing 'Tool – Matrix-file converter.'

3) Editing Case File (cas).

- Coordinate File must be made before starting to edit Case File.
- Case File can be edited in the same manner as Coordinate File.

Parameters

You can change several parameters for the analysis on the 'Analysis' tab panel.

- Statistical model
  - ① Poisson: for the data of the 'observed number' and the 'expected number.'
  - ② Binomial: for the data of the 'observed number' and the 'population.'

- Statistic type
  - ① Original LLR:

    The likelihood ratio statistic by Kulldorff, which has been used in the previous FleXScan version 1 and version 2.
  - ② LLR with Restriction:

    The restricted likelihood ratio statistic by Tango, with a preset parameter for restriction 'Alpha' (default is 0.2). This statistic avoids detecting undesirably large clusters, and improves calculation time. See the reference article for details.

- Scanning method
  - ① Flexible : flexible scan statistic by Tango and Takahashi
  - ② Circular : circular scan statistic by Kulldorff

- The Maximum Spatial Cluster Size

  The number of maximum spatial cluster size to scan. See the reference article for details. If Original LLR is selected, recommended<=20. If LLR with Restriction is selected, there are no restrictions.

- Random number:

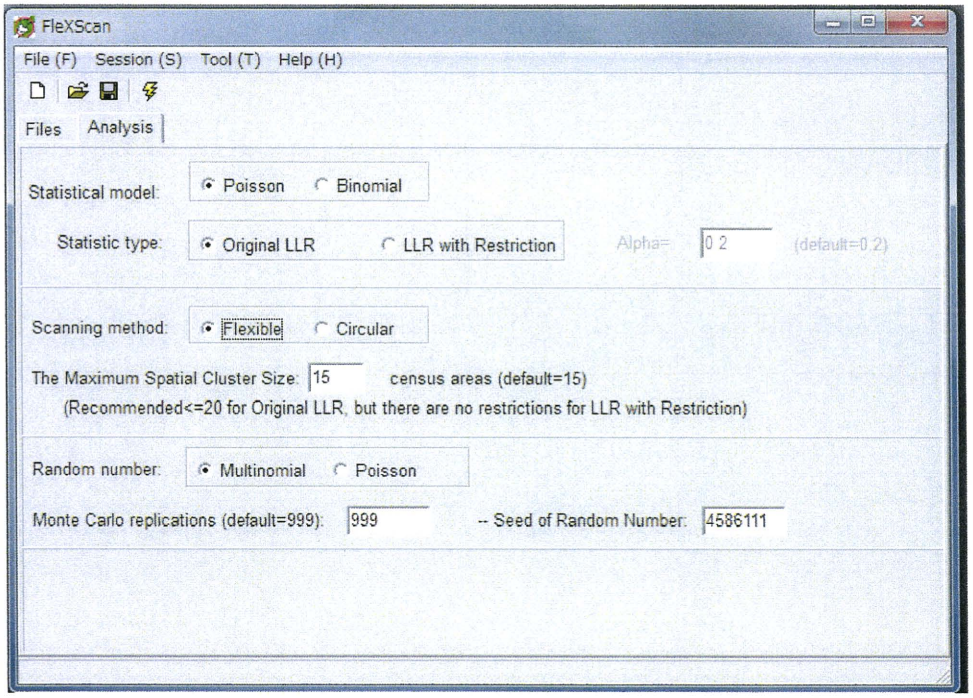
  The type of random number for Monte Carlo simulation
  - ① Multinomial: Total number of cases in whole area is fixed. It can be chosen in either 'Poisson' or 'Binomial' model.
  - ② Poisson: Total number of cases is not fixed, and it can be chosen in 'Poisson model'
  - ③ Binomial: Total number of cases is not fixed, and it can be chosen in 'Binomial model'

- Monte Carlo replications

  The number of Monte Carlo replications to calculate a p-value for statistical test. For example, if this number is set to 999, the p-value is calculated from the simulated 999 plus 1 observed log-likelihood ratio values (999+1=1000 in total).

- Seed of Random Number

  The seed for generating random numbers in the Monte Carlo simulation.



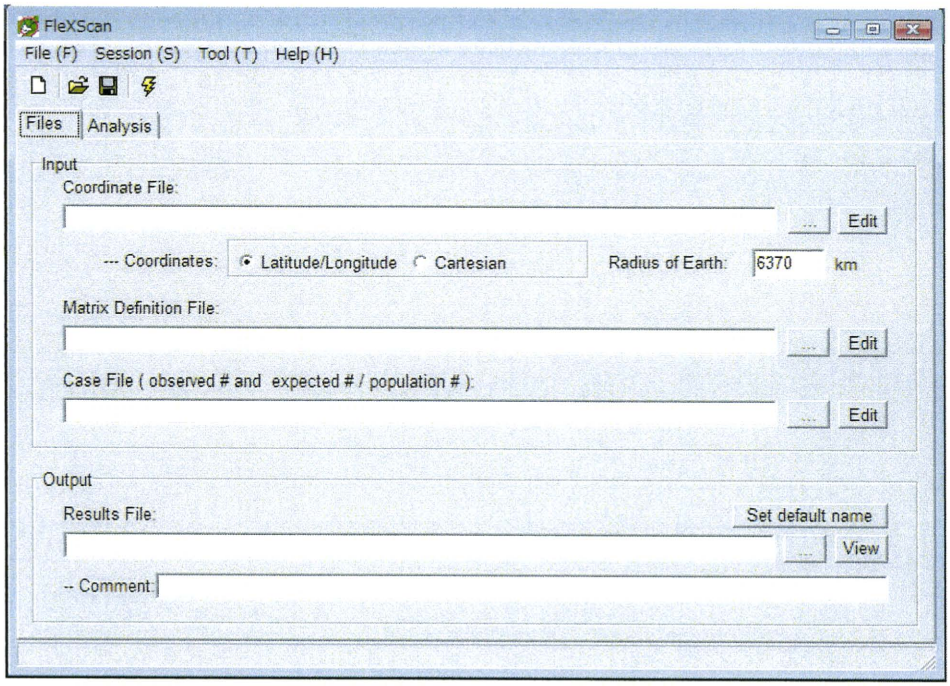
Also, on the 'Files' tab panel,

- Coordinates

  The type of coordinates used by the coordinates file.

- Radius of Earth

  Radius of Earth to calculate a distance between two sets of latitude and longitude. It is approximately 6370 km in Japan.


FleXScan
File (F) Session (S) Tool (T) Help (H)
Files
Analysis
Input
Coordinate File:
...
Edit
--- Coordinates:
Latitude/Longitude
Cartesian
Radius of Earth:
6370
km
Matrix Definition File:
...
Edit
Case File (observed # and expected # / population #):
...
Edit
Output
Results File:
Set default name
...
View
-- Comment:

Software Licenses

- The FleXScan software may be used freely, with proper references to both the software and the statistical methods papers. The suggested citations are:

  - Tango T. and Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. International Journal of Health Geographics 2005, **4**:11.

  - Takahashi K, Yokoyama T and Tango T. FleXScan v3.1: Software for the Flexible Scan Statistic. National Institute of Public Health, Japan, 2010.

  Also for the spatial scan statistic with a restricted likelihood ratio,

  - Tango T. A spatial scan statistic with a restricted likelihood ratio. Japanese Journal of Biometrics 2008, **29**:75-95.

Copyrights

http://www.niph.go.jp/soshiki/gijutsu/index_e.html

A maximum scan score-type statistic based on Anscombe's variance stabilization transformation for disease clustering

Kunihiko Takahashi[1] and Toshiro Tango[2]

[1] Department of Technology Assessment and Biostatistics, National Institute of Public Health, Japan
[2] Center for Medical Statistics, Japan

In spatial epidemiological studies, the circular spatial scan statistic proposed by Kulldorff and Nagarwalla (1995) has been widely used along with SaTScan software for cluster detection testing. Also, several different scan statistics for the test have been proposed to detect arbitrarily shaped clusters which cannot be detected by the circular scan statistic. Many of these statistics are based on maximizing the likelihood ratio statistic for each window Z. This likelihood ratio statistic is derived from assuming a "hot-spot" model for the cluster, and a common relative risk $\theta_i = \theta(Z)$ for $i \in Z$ is estimated as a Standardized Mortality/Morbidity Ratio (SMR) by maximum likelihood estimation

On the other hand, one definition of clusters is given as "a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance." Without any assumptions about the shape or form of the cluster, the most basic definition would be "any area within the study region of significant elevated risk" (Lawson, 2006). This definition allows us to consider not only the hot-spot models for clusters but also others. As a model, we can define a cluster more generally having elevated risk $\theta_i > 1.0$ for each i within the window Z, but they are not assumed equal. Recently, a discrete maximum scan score-type statistic has been proposed by Glaz and Zhang (2006) for testing the null hypothesis that the observations are independent and identically distributed according to a specified distribution, against an alternative that the observations cluster within a window of unknown length on a one-dimensional sequence. However, we cannot assume an identical distribution for the disease clustering in this situation. The variance of observed cases in the individual region is influenced by its expected counts μ_i, and they are considerably small for a rare disease. Also the typical estimates of relative risk such as SMR have variance which is proportional to $1/\mu_i$, and thus can yield large changes in the estimate with relatively small changes in expected value.

In this work, we propose a new scan statistic to detect clusters such as $\theta_i > 1.0$ for each region within the cluster, which is based on a score test in a similar manner to Glaz and Zhang's discrete score-type statistic. To circumvent the influence of various μ_i for each region i within the window, we will apply Anscombe(1948)'s variance stabilization transformation to the data in our procedure. Monte Carlo studies show that the proposed scan statistic can detect clusters more accurately than those based on the likelihood ratio.

References

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35**, 246–254.

Glaz, J. and Zhang, Z. (2006). Maximum scan score-type statistics. *Statistics & Probability Letters* **76**, 1316–1322.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine* **14**, 799–810.

Lawson, A. B. (2006). *Statistical Methods in Spatial Epidemiology* (2nd ed.). Chichester: John Wiley & Sons, Ltd.

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

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, Kohriyama K.	A space-time scan statistic for detecting emerging outbreaks	Biometrics	67	106-115	2011
高橋邦彦	国立保健医療科学院職員の活動：サーベイランス解析の視点から	保健医療科学	58(3)	265-266	2009

研究成果の刊行物・別刷

特集：新型インフルエンザ流行対策―国立保健医療科学院の取り組みと今後の活動に向けて―

国立保健医療科学院職員の活動

高橋邦彦，富塚太郎，藤原武男，橘とも子，秋葉道宏，田中吉之，江藤亜紀子，武村真治，
鈴木晃，大澤元毅，鍵直樹，阪東美智子

国立保健医療科学院

Professional Activities of NIPH Experts against Pandemic Influenza

K. Takahashi, T. Tomizuka, T. Fujiwara, T. Tachibana, M. Akiba, Y. Tanaka, A. Eto, S. Takemura,
A. Suzuki, H. Osawa, N. Kagi, M. Bando

National Institute of Public Health

〈国立保健医療科学院職員の活動〉

サーベイランス解析の視点から

高橋邦彦

国立保健医療科学院技術評価部

今回，厚生労働省・新型インフルエンザ対策推進本部事務局の業務に携わる機会を得た．具体的には日々報告される患者情報および関連サーベイランスデータの収集・集計作業の一部を担当しながら本部で行われる会議やミーティングにも参加し，その時々の問題点や方針の検討の様子を垣間見ることができた．さらに実際のデータに触れる中で，患者から保健所，地方衛生研究所，自治体，国へと情報がどのように報告され日々の状況把握が行われているかについて理解することができた．具体的な作業の情報や問題点については他の先生方に委ね，本稿では私自身が係っているサーベイランスにおける統計解析の面から，いくつかの点について今後の検討課題を含め報告したい．

当初，新型インフルエンザ発生について各国とも正確な発生状況の把握に努め，日本においても国立感染症研究所，地方衛生研究所における検体検査によって症例を確定してきた．しかし感染者数の増加に伴い各国とも全数把握は困難となり，徐々に擬似症例等をもってWHOへの報告数とされてきた．一方で今回のインフルエンザが強毒性でなかったこともあり，当初から各国で報告される感染者数は過小評価であるとの研究，指摘がされており，例えば米国で3万人弱（感染の疑いが濃厚な人を含む）の報告があった時点で，非公式ながらCDCが米国内において推計感染者数は延べ100万人になったと発表した．実際の患者数と報告数が大きく異なる場合，致死率をはじめ感染症の様子を把握する他の指標にも大きな影響を与えてしまう．今回の対策本部においてもその推計について議論する機会ももてたが，この推計には様々な情報を組み込んだ統計モデルが有用であるものの各国それぞれ状況が大きく異なるため，世界共通にひとつのモデルでの推計はかなり難しいと考えられる．しかし適切な統計モデルが確立できれば，その推計に必要なデータも明らかになるため，今後，各国で取得可能なデータと適切な統計モデルの構築，およびそのデータ収集のシステム作りについて有機的に検討・改良が必要であると考える．

一方，集団発生（アウトブレイク）を早期発見するサーベイランスは重要であり，特に新型インフルエンザの発生についてはインフルエンザ様症状患者の発生状況を日々監視することも有用であろう．日本においては全国で約5,000の定点医療機関からインフルエンザ様症状の発生動向が週単位で毎週報告されている．紙面の都合上ここで詳細を述べることはできないが，わが国のインフルエンザ発生動向の把握においてこの定点観測はひとつの有用なツールであるものの，同時にかなりの限界があることも指摘されている．そこで国内外のいくつかの地域では，さらに広く毎日のデータを解析できるシステム作りが検討されている．実際米国においてはいくつかのシステムが稼動してお

り，アウトブレイクの早期検出を目指している．例えばニューヨーク市保健精神衛生局（DOHMH）では市内救急医療機関61機関中51機関から救急患者のデータが毎日集められて，それをもとに兆候（シグナル）の検出を目的とした解析がリアルタイムで行われている．具体的には，市全体での発生動向の解析と同時に，患者の居住地情報をもとにzip-code単位での発生を検出できるサーベイランスを行っている．シグナルが検出された場合には，関係機関，関係担当者らにその情報が提供され，詳細調査などが検討されることになる．そのシステムでは疾病集積性の検定法であるKulldorff（1997）のcircular scan statistic（SaTScan）が一つの統計解析手法として組み込まれシグナルの有意性の判定と同時にその地域の同定を行って報告がされている．最近このSaTScanを改良したTango and Takahashi（2005）によるflexible scan statistic（FleXScan）も注目・利用されてきており，我々はニューヨーク市の担当者らとの共同研究として，ニューヨーク市の実際のサーベイランスデータを用いて従来のSaTScanと我々の提案するFleXScanの比較検討を行っている．今回の新型インフルエンザに係るインフルエンザ様症状患者のデータの解析では，SaTScanでは検出できなかったシグナルがFleXScanでは検出され，この結果は実際の担当者の感覚に合致するものであった．今後，学校の欠席数のデータなどでも更なる検討を続け，ニューヨーク市においてFleXScanを用いたサーベイランス解析を行える環境が整うよう実証研究を進めている．

わが国においても毎日のデータを自動的に収集するサーベイランスシステムの研究は行われているものの，残念ながら，現時点では公式なシステムとしては稼動していない．今後，日本においても毎日のデータに基づくサーベイランスについて，実現可能なシステムから，その解析，結果のレポートまで包括的な検討が必要となってくるであろう．特に実際の医療現場や自治体，保健所，地方衛生研究所などの実情を反映しながら，サーベイランスに関する研究を積極的に進めると同時に，自治体機関等と国との橋渡しとして国立保健医療科学院が重要な役割を果たすことが期待される．今回の経験や得られた情報を今後の研究に十分に生かしていきたい．

本部業務の遂行にあたり，国立感染症研究所感染症情報センターの先生方，特に現在のサーベイランスシステムのデータに関する様々な情報をお教えいただいた大日康史先生に感謝します．

〈国立保健医療科学院職員の活動〉

厚生労働省新型インフルエンザ対策推進本部における活動

富塚太郎

国立保健医療科学院政策科学部

私は6月上旬より，厚生労働省新型インフルエンザ対策推進本部技術班への併任を頂き，主にフロントラインの自治体からの報告を集計し分析する仕事に従事させて頂いた．政策研究者としては，対策推進室での仕事や地域の医師とのコミュニケーションの中から，新型インフルエンザ対策に関する政策分析として，地域で新型インフルエンザ対策を行っている発熱外来の設置や運営に関する政策過程に問題意識を持ち，調査している．

発熱外来に関連する政策文書としては，平成17年12月に鳥インフルエンザ等に関する関係省庁対策会議による「新型インフルエンザ対策行動計画」の中で，フェーズ3A（ヒトへの新しい亜型のインフルエンザ感染が確認されているが，ヒトからヒトへの感染は基本的にない）において，厚生労働省が都道府県等に対して発熱外来等を行う医療機関の準備を要請する旨が初めて示されている．その後，平成19年3月26日の新型インフルエンザ専門家会議による「新型インフルエンザ対策ガイドライン（フェーズ4以降）」で発熱外来の設置を含めた医療体制に関するガイドラインが示され，都道府県等が主体となって発熱外来設置可能医療機関のリスト作成や住民への受診経路の周知を行う旨が示され，直近では，平成21年2月11日に厚生労働省から「新型インフルエンザ対策指針」と「新型インフルエンザ対策ガイドライン」が出された中で，各都道府県と保健所設置市・特別区に対して，診療体制の整備の一つとして発熱外来を担当する医療機関のリスト作成をはじめとするpre-pandemic preparationの必要性等を提示している．

今回の流行で，発熱外来は実際にはどのように運用され，どのように機能した・しなかったのだろうか．発熱外来の運用自体は，設置する自治体・医療機関に大きな緊張と負担をもたらすが，その内容はどうであっただろうか．6月2日までの調査によると，実際に都道府県や保健所設置市・特別区によりリストされた発熱外来設置可能医療機関は980を超え，7月3日までの調査によると実際に患者を診療した発熱外来は750程度．患者が集団発生した地域での発熱外来を通した対応には，多くの混乱が報告され，県知事から厚生労働大臣への支援要請等も行われていた．調

BIOMETRICS 67, 106 115
March 2011

DOI: 10.1111/j.1541-0420.2010.01412.x

A Space–Time Scan Statistic for Detecting Emerging Outbreaks

Toshiro Tango,[1,*] Kunihiko Takahashi,[1] and Kazuaki Kohriyama[2]

[1]Department of Technology Assessment and Biostatistics, National Institute of Public Health, 3-6 Minami 2 chome Wako Saitama-ken 351-0197, Japan
[2]Emergency Life-Saving Technique Academy of KYUSHU, Kitakyushu, Japan
*email: tango@niph.go.jp

SUMMARY. As a major analytical method for outbreak detection, Kulldorff's space–time scan statistic (2001, *Journal of the Royal Statistical Society, Series A* **164**, 61 72) has been implemented in many syndromic surveillance systems. Since, however, it is based on circular windows in space, it has difficulty correctly detecting actual noncircular clusters. Takahashi et al. (2008, *International Journal of Health Geographics* **7**, 14) proposed a flexible space time scan statistic with the capability of detecting noncircular areas. It seems to us, however, that the detection of the most likely cluster defined in these space time scan statistics is not the same as the detection of localized emerging disease outbreaks because the former compares the observed number of cases with the *conditional* expected number of cases. In this article, we propose a new space time scan statistic which compares the observed number of cases with the *unconditional* expected number of cases, takes a time-to-time variation of Poisson mean into account, and implements an outbreak model to capture localized emerging disease outbreaks more timely and correctly. The proposed models are illustrated with data from weekly surveillance of the number of absentees in primary schools in Kitakyushu-shi, Japan, 2006.

KEY WORDS: Efficient score test; Likelihood ratio test; Negative binomial distribution; Poisson distribution; Surveillance.

1. Introduction

Since the World Trade Center attacks of September 11, 2001, the anthrax-laden letters that followed in October 2001, and the severe acute respiratory syndrome outbreak in 2002, there has been considerable interest in developing syndromic surveillance systems that would be used for early detection of disease outbreak and prevention of widespread morbidity and mortality (for example, Lazarus et al., 2002; Lombardo et al., 2003; Mostashari et al., 2003; Platt et al., 2003; Heffernan et al., 2004). Early detection of disease outbreaks enables public health officials to implement disease control and prevention measures at the earliest possible time. Over the last decade, many statistical methods have been directed at detecting changes or aberrations in public health surveillance time-series data (Sonesson and Bock, 2003). However, 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 localized changes in spatial patterns of disease (Lawson and Kleinman, 2005). Above all, Kulldorff (2001)'s space–time scan statistic and Kulldorff et al. (2005)'s space–time permutation scan statistic have been implemented in many syndromic surveillance systems along with the SaTScan software (Kulldorff and Information Management Services, Inc., 2009). Because this approach is based on circular windows in space, it has difficulty correctly detecting actual noncircular clusters. To detect noncircular spatial clusters, Patil and Taillie (2004), Duczmal and Assunção (2004), Tango and Takahashi (2005), Assunção et al. (2006) and Kulldorff et al. (2006) have proposed different spatial scan statistics. Regarding extension to a space time scan statistic, Takahashi et al. (2008) proposed a space time scan statistic with the capability of detecting noncircular areas using a flexible spatial window devised by Tango and Takahashi (2005), which is implemented in the `FleXScan` software (Takahashi, Yokoyama, and Tango, 2009).

When using these space time scan statistics for surveillance, Kleinman et al. (2005) showed the importance of adjusting for naturally occurring temporal trends and geographical patterns. To accomplish this, they applied Kleinman, Lazarus, and Platt (2004)'s approach based on a logistic regression model with adjustment for not only fixed temporal effects such as a month, a day-of-week, or a linear trend over past years but also random regional effects. The model is fitted using data from a predefined baseline period such as over the past 1 year. Kleinman (2005) extended his idea to a Poisson regression model with regional random-effects. These model-based approaches have been used to calculate the expected number of cases for a given surveillance day in a given region when applying the space time scan statistics. It seems to us, however, that the most likely cluster, and any secondary clusters detected by these space time scan statistics, are not always appropriate for the purpose of detection of localized emerging disease outbreaks. Zhou and Lawson (2008), on the other hand, considered the application of Bayesian spatial modeling with a vector exponentially weighted moving average method.

In this article, we propose a new space time scan statistic which (i) compares the observed number of cases with the *unconditional* expected number of cases, (ii) takes a time-to-time

variation of Poisson mean into account and (iii) implements an outbreak model to capture localized emerging disease outbreaks more timely and correctly. The proposed space time scan statistics are illustrated with data from weekly surveillance of the number of absentees in primary schools in Kitakyushu-shi, Japan, 2006.

2. Methods

Consider the situation where an entire study area is divided into m regions (for example, counties, zip-codes, enumeration districts) with each region periodically reporting the number of cases n_{it} (for region i at time t) of a disease or syndrome under study. Because we are only interested in detecting outbreaks that are alive (active) at the current time t_P, we only consider outbreaks that are present in the following T time intervals or *temporal windows:*

$$I_u = [t_P - u + 1, t_P],\ u = 1, \ldots, T, \tag{1}$$

where T is a prespecified *maximum temporal length* of the cluster or outbreak. In the next subsection, we will briefly review the existing space time scan statistics to summarize the basic ideas and underlying assumptions behind these scan statistics.

2.1 *Existing Space–Time Scan Statistics*

As practically available space time scan statistics, we will consider Kulldorff (2001)'s scan statistic and Takahashi et al. (2008)'s scan statistic. In this article, we will consider only the Poisson model in which the observed number of cases in region i at time t is independently distributed according to

$$N_{it} \sim \text{Poisson}(\theta_{it}\mu_{it}), \tag{2}$$

where N_{it} denote the random variable for n_{it}, θ_{it} denote the unknown relative risk in region i at time t, and μ_{it} denote the *conditional* expected number of cases such that

$$\sum_{i=1}^{m}\sum_{t=1}^{T}\mu_{it} = \sum_{i=1}^{m}\sum_{t=1}^{T}n_{it} = n. \tag{3}$$

If we can apply Kleinman et al. (2004)'s approach to adjust for naturally occurring temporal trends and geographical variations, we can obtain the adjusted expected number of cases $e_{it}^{(R)}$. In this case, the *conditional* expected value is calculated as

$$\mu_{it} = \frac{ne_{it}^{(R)}}{\sum_{i=1}^{m}\sum_{t=1}^{T}e_{it}^{(R)}}. \tag{4}$$

Both space time scan statistics assume that the relative risk has the following *hot-spot model*:

$$\theta_{it} = \begin{cases} \theta(W) = a_{\text{in}}, & \text{if } (i,t) \in W = Z \times I_u \\ \theta(W^c) = a_{\text{out}}, & \text{otherwise,} \end{cases} \tag{5}$$

where W^c denotes all the domains except for W and Z is a spatial window. Kulldorff's space–time scan statistic uses a *cylindrical* domain with a circular window while Takahashi et al.'s space time scan statistic uses a *prismatic* domain with an arbitrarily shaped spatial window Z. We will omit here the details of the scanning methods. Under the model (5), both space time scan statistics consider the following hypotheses for each of all possible sets of domains $W = Z \times I_u$:

$$H_0 : a_{\text{in}} = a_{\text{out}}, \quad H_1 : a_{\text{in}} > a_{\text{out}}. \tag{6}$$

Then, the space time scan statistic λ is the *conditional* maximum likelihood ratio over all possible domains W

$$\lambda = \sup_{W \in \mathcal{W}} \left(\frac{n(W)}{\mu(W)}\right)^{n(W)} \left(\frac{n - n(W)}{n - \mu(W)}\right)^{n - n(W)} \times I\left(\frac{n(W)}{\mu(W)} > \frac{n - n(W)}{n - \mu(W)}\right), \tag{7}$$

where $I()$ is the indicator function, $n()$ and $\mu()$ denotes the observed number of cases and the expected number of cases within the specified domain, respectively, i.e.,

$$n(W) = \sum_{(i,t)\in W} n_{it}, \quad \mu(W) = \sum_{(i,t)\in W} \mu_{it}. \tag{8}$$

The domain $W^* = Z^* \times I_{u^*}$ for which the *conditional* likelihood ratio is maximized identifies the *Most Likely Cluster* (MLC). Monte Carlo hypothesis testing (Dwass, 1957) is required to obtain the null distribution of λ and the Monte Carlo simulated p-value. There also may be secondary clusters that do not overlap with MLC, which may be of great interest. The p-value of the secondary clusters is obtained by comparing the likelihood of secondary clusters with that of MLC in Monte Calro simulated data.

However, there seems to be at least three reasons why Kulldorff's and Takahashi et al.'s space–time scan statistics are not appropriate for syndromic surveillance. (i) "Why can't the conditional expected number of cases μ (3) or (4) be used for detecting *emerging disease outbreaks* ?" To see how inadequate it is, let us consider the study area comprising three regions with the observed number of cases $\boldsymbol{n} = (5,5,5)'$ and the expected number of cases $\boldsymbol{e} = (1,1,1)'$ unconditionally calculated using data from the prespecified baseline period. By emerging disease outbreaks we usually mean a significant increasing trend in the observed number of cases, starting from an unknown time point. In this example, we can see a clear increase in counts in all regions. However, if we apply the conditional expectation (4) then we have $\boldsymbol{\mu} = (5,5,5)'$, indicating no increase at all. (ii) Kulldorff's and Takahashi et al.'s space time scan statistics consider the *hot-spot* model (6) for emerging disease outbreaks. The temporal pattern of emerging disease outbreaks, however, usually has a *gradual or steep increase* in the number of cases under study in the initial stage. The hot-spot pattern is a special case of temporal patterns of emerging disease outbreaks. Therefore, they are expected to be less powerful in detecting such outbreak patterns. (iii) When we were analyzing data from weekly surveillance on the number of absentees for each primary school in Kitakyushu, Japan (to be illustrated later), we often observed nonnegligible random week-to-week variation of the Poisson mean or temporal overdispersion. It is well known that, without taking this type of overdispersion, if any, into account, the false alarm rate increases in the field of statistical process control (Montgomery, 1991). These observations have motivated us to develop a new space time scan statistic to be described in the next subsection.

2.2 A New Space–Time Scan Statistic

2.2.1 *Temporal overdispersion.* We will assume that, under the null hypothesis of no outbreaks, the number of cases N_{it} in region i $(i = 1, \ldots, m)$ at some surveillance time t follows an independent negative binomial distribution $NB(\mu_{it}, \phi_{it})$ by taking the possibility of nonnegligible time-to-time variation of Poisson mean or *temporal overdispersion*, into account

$$H_0 : N_{it} \sim NB(\mu_{it}, \phi_{it}), \tag{9}$$

where the negative binomial distribution used here is given by

$$\Pr\{N_{it} = n_{it} \mid \mu_{it}, \phi_{it}\} = \frac{\Gamma(\phi_{it} + n_{it})}{\Gamma(\phi_{it})\, n_{it}!} \left(\frac{\phi_{it}}{\phi_{it} + \mu_{it}}\right)^{\phi_{it}} \left(\frac{\mu_{it}}{\phi_{it} + \mu_{it}}\right)^{n_{it}} \tag{10}$$

and

$$E(N_{it}) = \mu_{it}, \quad \mathrm{Var}(N_{it}) = \mu_{it} + \mu_{it}^2/\phi_{it} = \mu_{it} w_{it}. \tag{11}$$

The temporal overdispersion is given by

$$w_{it} = 1 + \mu_{it}/\phi_{it} \tag{12}$$

and ϕ_{it} denote the parameter regulating overdispersion. Needless to say, if we do not observe any overdispersion in region i, then we have only to set $\phi_{it} = \infty$ or $w_{it} = 1$.

2.2.2 *How to estimate* (μ_{it}, ϕ_{it}). The expected number of cases μ_{it} and the parameter ϕ_{it} should be estimated using data from a predefined *baseline period.* If we can use the surveillance data reported for several years in the past, then we can apply an appropriate *negative binomial regression model* within the framework of generalized linear mixed models (for example, see Hardin and Hilbe, 2007; Hilbe, 2007)

$$\log E(Y_{it}) = \sum_{j=1} x_{itj}\beta_j + b_i, \quad Y_{it} \sim NB\big(\mu_{it}^{(Y)}, \phi_i^{(Y)}\big), \tag{13}$$

where Y_{it} denotes a random variable for the observed counts data y_{it} in the predefined baseline period, x_{itj} denotes the value of covariate j such as a month, a day-of-week, and so on; b_i denotes a random regional effects independently normally distributed with mean 0; and ϕ_{it} is usually set constant over time, i.e., $\phi_{it} = \phi_i$. This model is an extension of generalized linear mixed models used by Kleinman et al. (2004) and Kleinman (2005). The independence assumption among the b_i is made for ease of application, the same reason used by Kleinman et al. (2004). Using this regression model, we can set $\mu_{it} = \hat{\mu}_{it}^{(Y)}$ and $\phi_{it} = \hat{\phi}_i^{(Y)}$.

However, if the surveillance system starts in the recent past as in the case of our example illustrated later, we cannot adopt the above-stated regression approach. In this case, we can use a so-called *moving average* method. This says that the null expected number of cases μ_{it} is assumed to be constant during the *baseline period* defined such as $\{t-1, t-2, \ldots t-B\}$ with baseline length B. Then these two parameters can be simply estimated using baseline mean $\bar{y}_{i\cdot}$ and baseline variance s_i^2 by the moment method, i.e.,

$$\mu_{it} = \hat{\mu}_{it}^{(Y)} = \bar{y}_{i\cdot} \tag{14}$$

$$\phi_{it} = \hat{\phi}_i^{(Y)} = \begin{cases} \dfrac{\bar{y}_{i\cdot}^2}{s_i^2 - \bar{y}_{i\cdot}}, & \text{if } s_i^2 > \bar{y}_i \\ \infty, & \text{otherwise.} \end{cases} \tag{15}$$

2.2.3 *An outbreak model.* Under the above-stated situation, the alternative hypothesis is given by

$$H_1 : N_{it} \sim NB(\theta_{it}\mu_{it}, \phi_{it}), \tag{16}$$

where (μ_{it}, ϕ_{it}) are known and θ_{it} denote the unknown relative risk in region i at surveillance time t. We propose the following *outbreak model* to capture localized emerging disease outbreaks:

$$\theta_{it} = \begin{cases} h(\tau + \beta_W (t - t_p + u)), & \text{if } (i,t) \in W = Z \times I_u \\ 1, & \text{otherwise,} \end{cases} \tag{17}$$

where $h(\tau)$ denote the relative risk at $t = t_p - u$ and the *initial slope* of emerging disease outbreak which starts just after the time point $t_p - u$ within the domain W is

$$\left[\frac{\partial \theta_{it}}{\partial t}\right]_{t=t_p-u} = \beta_W h'(\tau) \tag{18}$$

and $h(.)$ denotes any monotonically increasing function with $h(\tau) = 1$ and the first and second differentials $h'(.)$ and $h''(.)$ are assumed to be finite. Then, the above-stated hypothesis testing is simply reduced to the following hypothesis testing over all possible sets of domains $W = Z \times I_u$:

$$H_0 : \beta_W = 0, \quad H_1 : \beta_W > 0. \tag{19}$$

Under the outbreak model (17), the likelihood is

$$L(\beta \mid \mu_{it}, \phi_{it}, t_p, u, t) = \prod_{i \in Z} \prod_{t \in I_u} \frac{\Gamma(\phi_{it} + n_{it})}{\Gamma(\phi_{it})\, n_{it}!} \left(\frac{\phi_{it}}{\phi_{it} + \theta_{it}\mu_{it}}\right)^{\phi_{it}} \left(\frac{\theta_{it}\mu_{it}}{\phi_{it} + \theta_{it}\mu_{it}}\right)^{n_{it}} \times \prod_{(i,t) \in W^c} \frac{\Gamma(\phi_{it} + n_{it})}{\Gamma(\phi_{it})\, n_{it}!} \left(\frac{\phi_{it}}{\phi_{it} + \mu_{it}}\right)^{\phi_{it}} \left(\frac{\mu_{it}}{\phi_{it} + \mu_{it}}\right)^{n_{it}}. \tag{20}$$

Then, as with Kulldorff's space–time scan statistics, we can construct a likelihood ratio test for testing the null hypothesis $H_0 : \beta_W = 0$ against $H_1 : \beta_W > 0$. However, the likelihood ratio test statistic requires the functional form $h(.)$ and the maximum likelihood estimator for β_W. Therefore, we will derive here an efficient score test that does not depend on the functional form $h(.)$ and does not require the maximum likelihood estimator for β_W, which is asymptotically equivalent to the likelihood ratio test. The efficient score test statistic for the null hypothesis $H_0 : \beta_W = 0$ is

$$S = \sup_{Z \in Z,\ 1 \le u \le T} \frac{\sum_{i \in Z} \sum_{t \in I_u} (n_{it} - \mu_{it})(t - t_p + u)/w_{it}}{\sqrt{\sum_{i \in Z} \sum_{t \in I_u} \mu_{it}(t - t_p + u)^2/w_{it}}} \sim N(0,1). \tag{21}$$

The derivation of the efficient score test is given in the Appendix. Needless to say, if the Poisson distribution can be