- 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:

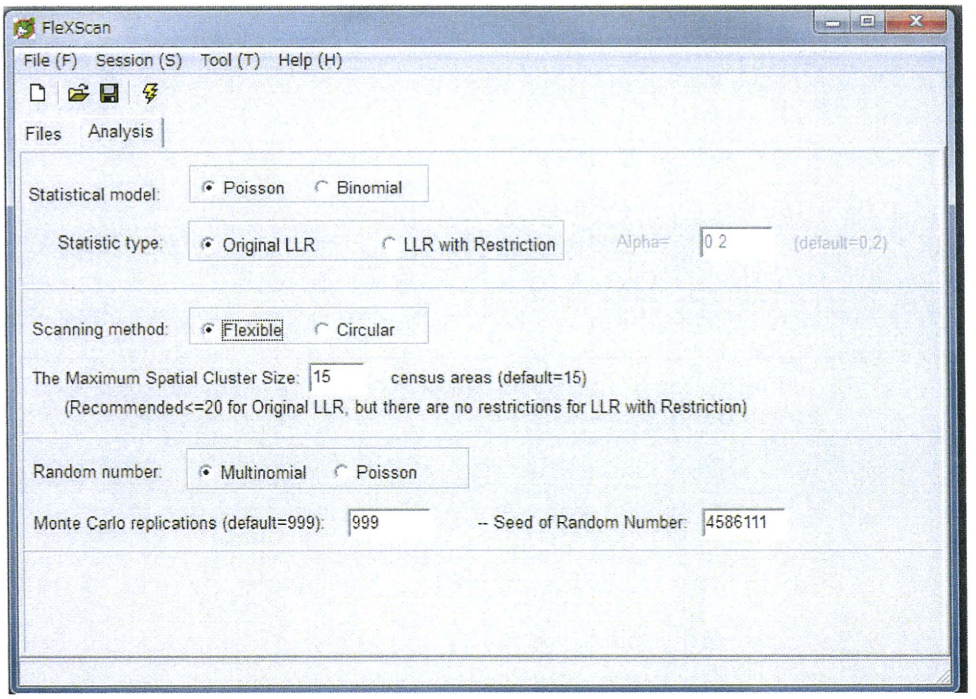
  The type of random number for Monte Carlo simulation
  - ① Multinomial: Total number of cases in whole area is fixed. It can be chosen in either 'Poisson' or 'Binomial' model.
  - ② Poisson: Total number of cases is not fixed, and it can be chosen in 'Poisson model'
  - ③ Binomial: Total number of cases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



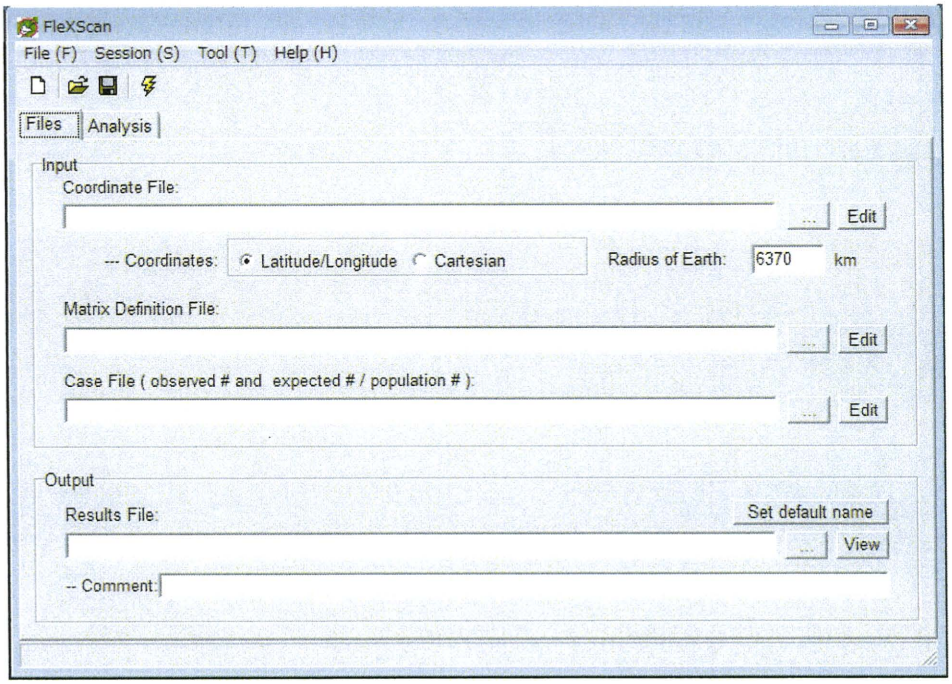
Also, on the 'Files' tab panel,

- 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

©Copyright 2010 Takahashi K, Yokoyama T, Tango T. All rights reserved.

The FleXScan software is copyrighted by Kunihiko Takahashi, Tetsuji Yokoyama and Toshiro Tango, National Institute of Public Health, Japan.

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.

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

Kunihiko Takahashi*, Toshiro Tango**

*National Institute of Public Health, Japan
**Center for Medical Statistics, Japan

International Biometric Conference 2010

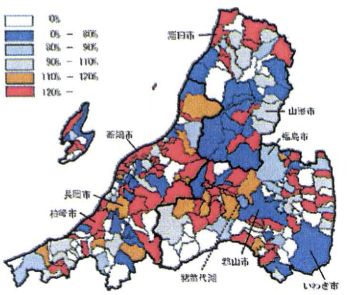
Scan Statistics for Disease Clustering

- 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.
- 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.
  - SA scan statistic by Duczmal & Assunção (2004)
  - Flexible scan statistic by Tango & Takahashi (2005) (with "FleXScan" software)

  and more...

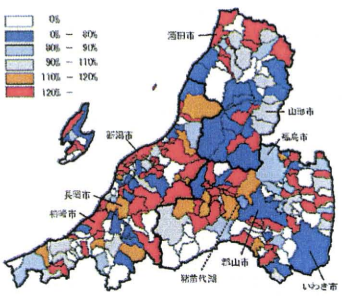
Example 1: cancer of the gallbladder

- Standardized Mortality Ratio (SMR) for cancer of the gallbladder
- Niigata, Fukushima and Yamagata prefectures in Japan.
- Male, 1996 - 2000.
- Total cases : 665 (for 5 years)
- Total population (male): 14,232,598 (person-year)



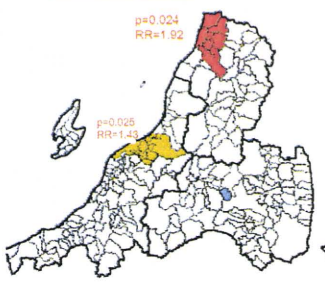
Example 1: cancer of the gallbladder

- We consider the question of whether disease cases are clustered in space.
- "Disease Clustering Tests" can be applied for the analysis.

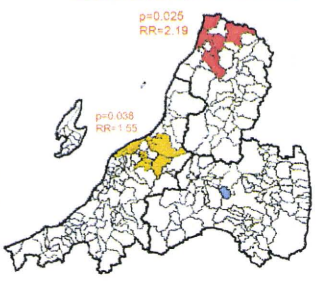


Detected significant cluster

SaTScan (Kulldorff)



FleXScan (Tango & Takahashi)



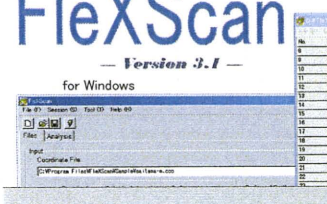
They are similar but the FleXScan detected non-circular clusters.

Software for the Flexible Scan Statistic

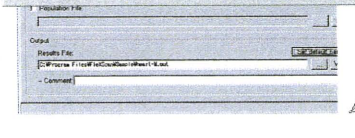
FleXScan

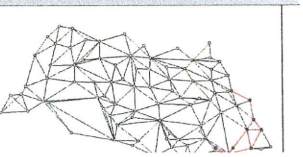
— Version 3.1 —

for Windows



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





Hot-spot model

- Many of scan statistics for disease clustering are based on maximizing the likelihood ratio statistic for each window.
- For Poisson model, the observed cases n_i of the individual region follows $n_i \sim \text{Poisson}(\theta_i \mu_i) \quad i = 1, 2, \ldots, m$ where μ_i denotes the expected number of cases, and θ_i is assumed to be a relative risk parameter.
- Kulldorff (1995) assumes a hot-spot window Z as
$$\theta_i = \theta_a \ \ (i \in Z), \quad \theta_i = \theta_b \ \ (i \notin Z).$$
- The hypothesis whether Z is a cluster or not as
$$H_0 : \theta_a = \theta_b \quad \text{vs.} \quad H_1 : \theta_a > \theta_b$$

Likelihood ratio statistic under hot-spot

- Under the condition of $\sum_{i=1}^{m} n_i = \sum_{i=1}^{m} \mu_i = N$, the likelihood ratio statistic for Z is derived as
$$\lambda(Z) = \left(\frac{n(Z)}{\mu(Z)}\right)^{n(Z)} \left(\frac{N - n(Z)}{N - \mu(Z)}\right)^{N - n(Z)} I\left(n(Z) > \mu(Z)\right),$$
where $n(Z) = \sum_{i \in Z} n_i$, $\mu(Z) = \sum_{i \in Z} \mu_i$.
- The scan statistic is the maximum likelihood ratio over all possible window $Z \in \mathcal{Z}$, $\max_{Z \in \mathcal{Z}} \lambda(Z)$, and the window with the maximum constitutes the most likely cluster.

Likelihood ratio statistic under hot-spot

- Under the condition of $\sum_{i=1}^{m} n_i = \sum_{i=1}^{m} \mu_i = N$, the likelihood ratio statistic for Z is derived as
$$\lambda(Z) = \left(\frac{n(Z)}{\mu(Z)}\right)^{n(Z)} \left(\frac{N - n(Z)}{N - \mu(Z)}\right)^{N - n(Z)} I\left(n(Z) > \mu(Z)\right),$$
where $n(Z) = \sum_{i \in Z} n_i$, $\mu(Z) = \sum_{i \in Z} \mu_i$.
- The scan statistic is the maximum likelihood ratio over all possible window $Z \in \mathcal{Z}$, $\max_{Z \in \mathcal{Z}} \lambda(Z)$, and the window with the maximum constitutes the most likely cluster.

- In this statistic, a common relative risk within a window is estimated as a SMR by maximum likelihood estimation.
$$\theta_i = \theta_a = \frac{n(Z)}{\mu(Z)} \ \ (i \in Z)$$

Definition of cluster

- The most basic definition of cluster 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.

Definition of cluster

- The most basic definition of cluster 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 $\theta_i > 1.0 \ \ (i \in Z)$, but they are not assumed equal, i.e,
$$\theta_i > 1.0 \ (i \in Z), \quad \theta_i \leq 1.0 \ (i \notin Z)$$
Not common risk

Definition of cluster

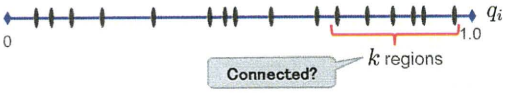
- Now, we consider detecting the most significant area with a fixed length k.
- Considering the hypothesis $K_0 : \theta_i = 1.0 \ \text{ vs. } \ K_1 : \theta_i > 1.0$ we denote by p_i the p-value of the test for each region, and $q_i = 1 - p_i$
- For the simplest situation,

Definition of cluster

- Now, we consider detecting the most significant area with a fixed length k.
- Considering the hypothesis $K_0 : \theta_i = 1.0$ vs. $K_1 : \theta_i > 1.0$ we denote by p_i the p-value of the test for each region, and $q_i = 1 - p_i$
- For the simplest situation,

Definition of cluster

- Now, we consider detecting the most significant area with a fixed length k.
- Considering the hypothesis $K_0 : \theta_i = 1.0$ vs. $K_1 : \theta_i > 1.0$ we denote by p_i the p-value of the test for each region, and $q_i = 1 - p_i$
- For the simplest situation,



- However, the cluster is given as a geographically bounded group, i.e., a set of connected regions.

Definition of cluster

- Also, the length k is unknown.
- Recently, Glaz and Zhang (2006) proposed "Maximum scan score-type statistics"
- They assume that S_k denotes a discrete scan statistic with a fixed window of length k, and the maximum scan score-type statistic is defined by $T = \max_k \left\{ \frac{S_k - E[S_k]}{\sqrt{\mathrm{Var}[S_k]}} \right\}$

Definition of cluster

- Also, the length k is unknown.
- Recently, Glaz and Zhang (2006) proposed "Maximum scan score-type statistics"
- They assume that S_k denotes a discrete scan statistic with a fixed window of length k, and the maximum scan score-type statistic is defined by $T = \max_k \left\{ \frac{S_k - E[S_k]}{\sqrt{\mathrm{Var}[S_k]}} \right\}$

We will try to apply the procedure to disease clustering.

To apply the procedure…

- The Glaz and Zhang's procedure assumes i.i.d. observations under the null hypothesis, but we cannot assume an identical distribution for disease clustering.

$$n_i \sim \mathrm{Poisson}(\theta_i \mu_i)$$

- We examine an approache for applying the procedure to satisfy the i.i.d. condition using the Anscombe's variance stabilization transformation.

Variance stabilization transformation

- We assume that $n_i \sim \mathrm{Poisson}(\theta_i \mu_i)$ $i = 1, 2, \ldots, m$, and denote $U_i = 2\sqrt{N_i + \frac{3}{8}} - 2\sqrt{\mu_i + \frac{1}{8}}$.
- By Anscombe (1948), $U_i \approx N(\xi_i, 1)$, asymptotically, where $\xi_i = E(U_i) = 2\sqrt{\theta_i \mu_i + \frac{1}{8}} - 2\sqrt{\mu_i + \frac{1}{8}}$

Variance stabilization transformation

- We assume that $n_i \sim \text{Poisson}(\theta_i \mu_i)$ $i = 1, 2, \ldots, m$, and denote $U_i = 2\sqrt{N_i + \frac{3}{8}} - 2\sqrt{\mu_i + \frac{1}{8}}$.
- By Anscombe (1948), $U_i \approx N(\xi_i, 1)$, asymptotically, where $\xi_i = E(U_i) = 2\sqrt{\theta_i \mu_i + \frac{1}{8}} - 2\sqrt{\mu_i + \frac{1}{8}}$ and $\xi_i = 0 \ (\theta_i = 1); \ \xi_i > 0 \ (\theta_i > 1); \ \xi_i < 0 \ (\theta_i < 1)$

Variance stabilization transformation

- We assume that $n_i \sim \text{Poisson}(\theta_i \mu_i)$ $i = 1, 2, \ldots, m$, and denote $U_i = 2\sqrt{N_i + \frac{3}{8}} - 2\sqrt{\mu_i + \frac{1}{8}}$.
- By Anscombe (1948), $U_i \approx N(\xi_i, 1)$, asymptotically, where $\xi_i = E(U_i) = 2\sqrt{\theta_i \mu_i + \frac{1}{8}} - 2\sqrt{\mu_i + \frac{1}{8}}$ and $\xi_i = 0 \ (\theta_i = 1); \ \xi_i > 0 \ (\theta_i > 1); \ \xi_i < 0 \ (\theta_i < 1)$
- For a specific window $Z_k = \{i_1, i_2, \ldots, i_k\}$ under $\sum_{i=1}^m n_i = \sum_{i=1}^m \mu_i$

$i \in Z_k$	$i \notin Z_k$
$H_0 : \theta_i = 1.0$ vs. $H_1 : \theta_i > 1.0$	$K_0 : \theta_i = 1.0$ vs. $K_1 : \theta_i < 1.0$
$H_0' : \xi_i = 0$ vs. $H_1' : \xi_i > 0$	$K_0' : \xi_i = 0$ vs. $K_1' : \xi_i < 0$

Variance stabi

$U_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ under H_0

- We assume that $n_i \sim \text{Poisson}(\theta_i \mu_i)$ $i = 1, 2, \ldots, m$, and denote $U_i = 2\sqrt{N_i + \frac{3}{8}} - 2\sqrt{\mu_i + \frac{1}{8}}$.
- By Anscombe (1948), $U_i \approx N(\xi_i, 1)$, asymptotically, where $\xi_i = E(U_i) = 2\sqrt{\theta_i \mu_i + \frac{1}{8}} - 2\sqrt{\mu_i + \frac{1}{8}}$ and $\xi_i = 0 \ (\theta_i = 1); \ \xi_i > 0 \ (\theta_i > 1); \ \xi_i < 0 \ (\theta_i < 1)$
- For a specific window $Z_k = \{i_1, i_2, \ldots, i_k\}$ under $\sum_{i=1}^m n_i = \sum_{i=1}^m \mu_i$

$i \in Z_k$	$i \notin Z_k$
$H_0 : \theta_i = 1.0$ vs. $H_1 : \theta_i > 1.0$	$K_0 : \theta_i = 1.0$ vs. $K_1 : \theta_i < 1.0$
$H_0' : \xi_i = 0$ vs. $H_1' : \xi_i > 0$	$K_0' : \xi_i = 0$ vs. $K_1' : \xi_i < 0$

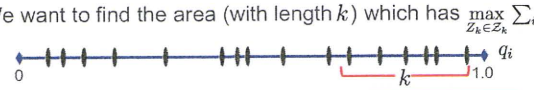
P-value

- Hence $q_i = \Pr\{U_i \le u_i\} = \Phi(u_i)$ is the "1-(p-value)" for $i \in Z_k$ of $H_0' : \xi_i = 0$ vs. $H_1' : \xi_i > 0$, while the "p-value" for $i \notin Z_k$ of $K_0' : \xi_i = 0$ vs. $K_1' : \xi_i < 0$.
- If $\sum_{i \in Z_k} q_i$ takes large value, $\sum_{i \notin Z_k} q_i = Q - \sum_{i \in Z_k} q_i$ becomes small, where $Q = \sum_{i=1}^m q_i$.
- We want to find the area (with length k) which has $\max_{Z_k \in \mathcal{Z}_k} \sum_{i \in Z_k} q_i$

- This is equivalent to finding " $U(Z_k) = \max_{Z_k \in \mathcal{Z}_k} \sum_{i \in Z_k} U_i$ "

P-value

- Hence $q_i = \Pr\{U_i \le u_i\} = \Phi(u_i)$ is the "1-(p-value)" for $i \in Z_k$ of $H_0' : \xi_i = 0$ vs. $H_1' : \xi_i > 0$, while the "p-value" for $i \notin Z_k$ of $K_0' : \xi_i = 0$ vs. $K_1' : \xi_i < 0$.
- If $\sum_{i \in Z_k} q_i$ takes large value, $\sum_{i \notin Z_k} q_i = Q - \sum_{i \in Z_k} q_i$ becomes small, where $Q = \sum_{i=1}^m q_i$.
- We want to find the area (with length k) which has $\max_{Z_k \in \mathcal{Z}_k} \sum_{i \in Z_k} q_i$



- This is equivalent to finding " $U(Z_k) = \max_{Z_k \in \mathcal{Z}_k} \sum_{i \in Z_k} U_i$ "
- However, the length k is unknown.

Standardized score-type statistic

- Under the null hypothesis

$$E(U(Z_k)) = 0, \quad V(U(Z_k)) = k \qquad (U(Z_k) = \max_{Z_k \in \mathcal{Z}_k} \textstyle\sum_{i \in Z_k} U_i)$$

- In a similar manner to Glaz and Zhang (2006), a maximum scan score-type statistic is defined as

$$U = \max_k \frac{U(Z_k) - E(U(Z_k))}{\sqrt{V(U(Z_k))}} = \max_k \frac{U(Z_k)}{\sqrt{k}} = \max_{Z \in \mathcal{Z}} \frac{\sum_{i \in Z} u_i}{\sqrt{\#Z}}$$

where $\#Z$ denotes the length of the window Z.

- The window with the maximum constitutes the most likely cluster, and the p-values are obtained through Monte Carlo hypothesis testing.

Properties

- Denote $\boldsymbol{u}(Z_k) = \{u_{i_1}, \ldots, u_{i_k}\}$ for window $Z_k = \{i_1, \ldots, i_k\}$
- The proposed scan statistic

$$U = \max_k \frac{U(Z_k) - E(U(Z_k))}{\sqrt{V(U(Z_k))}} = \max_k \frac{U(Z_k)}{\sqrt{k}} = \max_{Z \in \mathcal{Z}} \frac{\sum_{i \in Z} u_i}{\sqrt{\#Z}}$$

detects ...

- $\boldsymbol{u}(Z_3) = \{u_1, u_1, u_1\}$, $\boldsymbol{u}(Z_2) = \{u_1, u_1\}$, $\boldsymbol{u}(Z_1) = \{u_1\}$

Properties

- Denote $\boldsymbol{u}(Z_k) = \{u_{i_1}, \ldots, u_{i_k}\}$ for window $Z_k = \{i_1, \ldots, i_k\}$
- The proposed scan statistic

$$U = \max_k \frac{U(Z_k) - E(U(Z_k))}{\sqrt{V(U(Z_k))}} = \max_k \frac{U(Z_k)}{\sqrt{k}} = \max_{Z \in \mathcal{Z}} \frac{\sum_{i \in Z} u_i}{\sqrt{\#Z}}$$

detects ...

- $\boldsymbol{u}(Z_3) = \{u_1, u_1, u_1\}$, $\boldsymbol{u}(Z_2) = \{u_1, u_1\}$, $\boldsymbol{u}(Z_1) = \{u_1\}$

$\frac{3u_1}{\sqrt{3}} = \sqrt{3}u_1$ $\quad$ $\frac{2u_1}{\sqrt{2}} = \sqrt{2}u_1$ $\quad$ u_1

Properties

- Denote $\boldsymbol{u}(Z_k) = \{u_{i_1}, \ldots, u_{i_k}\}$ for window $Z_k = \{i_1, \ldots, i_k\}$
- The proposed scan statistic

$$U = \max_k \frac{U(Z_k) - E(U(Z_k))}{\sqrt{V(U(Z_k))}} = \max_k \frac{U(Z_k)}{\sqrt{k}} = \max_{Z \in \mathcal{Z}} \frac{\sum_{i \in Z} u_i}{\sqrt{\#Z}}$$

detects ...

- for $u_1 > u_2$,

$\boldsymbol{u}(Z_2) = \{u_1, u_2\}$, $\boldsymbol{u}(Z_1) = \{u_1\}$

Properties

- Denote $\boldsymbol{u}(Z_k) = \{u_{i_1}, \ldots, u_{i_k}\}$ for window $Z_k = \{i_1, \ldots, i_k\}$
- The proposed scan statistic

$$U = \max_k \frac{U(Z_k) - E(U(Z_k))}{\sqrt{V(U(Z_k))}} = \max_k \frac{U(Z_k)}{\sqrt{k}} = \max_{Z \in \mathcal{Z}} \frac{\sum_{i \in Z} u_i}{\sqrt{\#Z}}$$

detects ...

- for $u_1 > u_2$,

$\boldsymbol{u}(Z_2) = \{u_1, u_2\}$, $\boldsymbol{u}(Z_1) = \{u_1\}$

if $u_2 > (\sqrt{2} - 1)u_1$

$u_1 = 1.64$ $(p_1 = 0.05) \Rightarrow u_2 > 0.67$ $(p_2 < 0.25)$
$u_1 = 2.33$ $(p_1 = 0.01) \Rightarrow u_2 > 0.96$ $(p_2 < 0.17)$
$u_1 = 3.09$ $(p_1 = 0.001) \Rightarrow u_2 > 1.28$ $(p_2 < 0.10)$

Properties

- Denote $\boldsymbol{u}(Z_k) = \{u_{i_1}, \ldots, u_{i_k}\}$ for window $Z_k = \{i_1, \ldots, i_k\}$
- The proposed scan statistic

$$U = \max_k \frac{U(Z_k) - E(U(Z_k))}{\sqrt{V(U(Z_k))}} = \max_k \frac{U(Z_k)}{\sqrt{k}} = \max_{Z \in \mathcal{Z}} \frac{\sum_{i \in Z} u_i}{\sqrt{\#Z}}$$

detects ...

- In general,

$\boldsymbol{u}(Z_{k+1}) = \{u_1, \ldots, u_k, u_{k+1}\}$, $\boldsymbol{u}(Z_k) = \{u_1, \ldots, u_k\}$

Properties

- Denote $\boldsymbol{u}(Z_k) = \{u_{i_1}, \ldots, u_{i_k}\}$ for window $Z_k = \{i_1, \ldots, i_k\}$
- The proposed scan statistic

$$U = \max_k \frac{U(Z_k) - E(U(Z_k))}{\sqrt{V(U(Z_k))}} = \max_k \frac{U(Z_k)}{\sqrt{k}} = \max_{Z \in \mathcal{Z}} \frac{\sum_{i \in Z} u_i}{\sqrt{\#Z}}$$

detects ...

- In general,

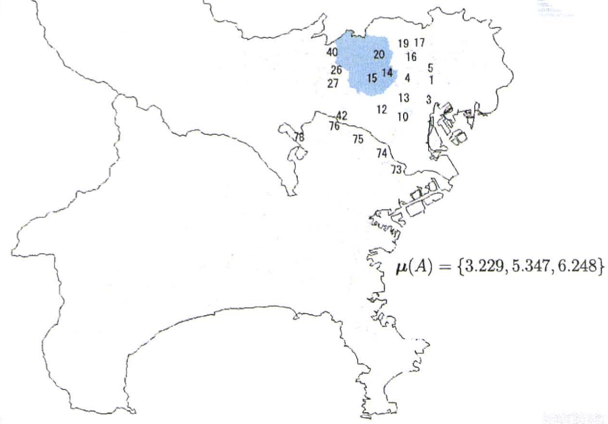
$\boldsymbol{u}(Z_{k+1}) = \{u_1, \ldots, u_k, u_{k+1}\}$, $\boldsymbol{u}(Z_k) = \{u_1, \ldots, u_k\}$

if $u_{k+1} > \left(\sqrt{\frac{k+1}{k}} - 1\right) \sum_{i=1}^{k} u_i = \left(\sqrt{k(k+1)} - k\right) \frac{\sum_{i=1}^{k} u_i}{k}$

Simulations

- m=113 regions; total pop.=19,803,618; total cases=200.
- Comparison of powers of
  - the conventional likelihood ratio statistic,
  - U statistic using Anscombe's transformation,
- Assumed true cluster
  - A={14, 15, 20} (circular cluster) $\mu(A) = \{3.229, 5.347, 6.248\}$
  - C={14, 15, 26, 27} (non-circular cluster) $\mu(C) = \{3.229, 5.347, 1.405, 1.672\}$
- Several relative risks are considered for each regions independently.
- alpha=0.05, 1000 trials are carried out.
- To compare the performance of the spatial scan statistic, we shall use the "bivariate power distribution" proposed by Tango and Takahashi (2005).

Cluster A = {14, 15, 20}
(circular cluster – circular scan)



likelihood ratio statistic λ / score-type statistic U

(i) $\theta_{14} = \theta_{15} = \theta_{20} = 3.0$

Likelihood ratio statistic λ:

$l \backslash s$	0	1	2	3	Total
1:	1	12			13
2:	0	0	64		64
3:	0	0	0	*575	575
4:	0	0	0	159	159
5:	0	0	0	37	37
6:	1	0	2	20	23
7:	0	0	0	9	9
8:	0	0	0	8	8
9:	0	0	5	2	7
10:	0	0	2	1	3
11:	0	0	1	2	3
12:	0	0	1	15	16
13:	0	0	1	12	13
14:	0	0	0	9	9
15:	0	0	0	12	12
Total	2	12	76	861	951

$P(p = 0.001) = 0.803$, $P(p \leqslant 0.01) = 0.887$

$P^{-}_{FP} = 0.651$

Score-type statistic U:

$l \backslash s$	0	1	2	3	Total
1:	0	36			36
2:	0	0	67		67
3:	0	0	0	*659	659
4:	0	0	0	138	138
5:	0	0	0	20	20
6:	1	0	1	8	10
7:	0	0	0	3	3
8:	1	0	0	3	4
9:	0	0	2	1	3
10:	0	0	0	1	1
11:	0	0	1	0	1
12:	0	0	0	1	1
13:	0	0	0	2	2
14:	0	0	0	2	2
Total	2	36	71	838	947

$P(p = 0.001) = 0.781$, $P(p \leqslant 0.01) = 0.894$

$P^{-}_{FP} = 0.762$

likelihood ratio statistic λ / score-type statistic U

(i) $\theta_{14} = \theta_{15} = \theta_{20} = 3.0$

$l \backslash s$	0	1	2	3	Total
1:	1	12			13
2:	0	0	64		64
3:	0	0	0	*575	575
4:	0	0	0	159	159
5:	0	0	0	37	37
6:	1	0	2	20	23
7:	0	0	0	9	9
8:	0	0	0	8	8
9:	0	0	5	2	7
10:	0	0	2	1	3
11:	0	0	1	2	3
12:	0	0	1	15	16
13:	0	0	1	12	13
14:	0	0	0	9	9
15:	0	0	0	12	12
Total	2	12	76	861	951

$P(p = 0.001) = 0.803$, $P(p \leqslant 0.01) = 0.887$

$P^{-}_{FP} = 0.651$

$l \backslash s$	0	1	2	3	Total
1:	0	36			36
2:	0	0	67		67
3:	0	0	0	*659	659
4:	0	0	0	138	138
5:	0	0	0	20	20
6:	1	0	1	8	10
7:	0	0	0	3	3
8:	1	0	0	3	4
9:	0	0	2	1	3
10:	0	0	0	1	1
11:	0	0	1	0	1
12:	0	0	0	1	1
13:	0	0	0	2	2
14:	0	0	0	2	2
Total	2	36	71	838	947

$P(p = 0.001) = 0.781$, $P(p \leqslant 0.01) = 0.894$

$P^{-}_{FP} = 0.762$

likelihood ratio statistic λ / score-type statistic U

(i) $\theta_{14} = \theta_{15} = \theta_{20} = 3.0$

$l \backslash s$	0	1	2	3	Total
1:	1	12			13
2:	0	0	64		64
3:	0	0	0	*575	575
4:	0	0	0	159	159
5:	0	0	0	37	37
6:	1	0	2	20	23
7:	0	0	0	9	9
8:	0	0	0	8	8
9:	0	0	5	2	7
10:	0	0	2	1	3
11:	0	0	1	2	3
12:	0	0	1	15	16
13:	0	0	1	12	13
14:	0	0	0	9	9
15:	0	0	0	12	12
Total	2	12	76	861	951

$P(p = 0.001) = 0.803$, $P(p \leqslant 0.01) = 0.887$

$P^{-}_{FP} = 0.651$

$l \backslash s$	0	1	2	3	Total
1:	0	36			36
2:	0	0	67		67
3:	0	0	0	*659	659
4:	0	0	0	138	138
5:	0	0	0	20	20
6:	1	0	1	8	10
7:	0	0	0	3	3
8:	1	0	0	3	4
9:	0	0	2	1	3
10:	0	0	0	1	1
11:	0	0	1	0	1
12:	0	0	0	1	1
13:	0	0	0	2	2
14:	0	0	0	2	2
Total	2	36	71	838	947

$P(p = 0.001) = 0.781$, $P(p \leqslant 0.01) = 0.894$

$P^{-}_{FP} = 0.762$

likelihood ratio statistic λ / score-type statistic U

(i) $\theta_{14} = \theta_{15} = \theta_{20} = 3.0$

$l \backslash s$	0	1	2	3	Total
1:	1	12			13
2:	0	0	64		64
3:	0	0	0	*575	575
4:	0	0	0	159	159
5:	0	0	0	37	37
6:	1	0	2	20	23
7:	0	0	0	9	9
8:	0	0	0	8	8
9:	0	0	5	2	7
10:	0	0	2	1	3
11:	0	0	1	2	3
12:	0	0	1	15	16
13:	0	0	1	12	13
14:	0	0	0	9	9
15:	0	0	0	12	12
Total	2	12	76	861	951

$P(p = 0.001) = 0.803$, $P(p \leqslant 0.01) = 0.887$

$P^{-}_{FP} = 0.651$

$l \backslash s$	0	1	2	3	Total
1:	0	36			36
2:	0	0	67		67
3:	0	0	0	*659	659
4:	0	0	0	138	138
5:	0	0	0	20	20
6:	1	0	1	8	10
7:	0	0	0	3	3
8:	1	0	0	3	4
9:	0	0	2	1	3
10:	0	0	0	1	1
11:	0	0	1	0	1
12:	0	0	0	1	1
13:	0	0	0	2	2
14:	0	0	0	2	2
Total	2	36	71	838	947

$P(p = 0.001) = 0.781$, $P(p \leqslant 0.01) = 0.894$

$P^{-}_{FP} = 0.762$

likelihood ratio statistic λ / score-type statistic U

(i) $\theta_{14} = \theta_{15} = \theta_{20} = 3.0$

$l \backslash s$	0	1	2	3	Total
1:	1	12			13
2:	0	0	64		64
3:	0	0	0	*575	575
4:	0	0	0	159	159
5:	0	0	0	37	37
6:	1	0	2	20	23
7:	0	0	0	9	9
8:	0	0	0	8	8
9:	0	0	5	2	7
10:	0	0	2	1	3
11:	0	0	1	2	3
12:	0	0	1	15	16
13:	0	0	1	12	13
14:	0	0	0	9	9
15:	0	0	0	12	12
Total	2	12	76	861	951

$P(p = 0.001) = 0.803,\ P(p \leqslant 0.01) = 0.887$

$P^{-}_{FP} = 0.651$

$l \backslash s$	0	1	2	3	Total
1:	0	36			36
2:	0	0	67		67
3:	0	0	0	*659	659
4:	0	0	0	138	138
5:	0	0	0	20	20
6:	1	0	1	8	10
7:	0	0	0	3	3
8:	1	0	0	3	4
9:	0	0	2	1	3
10:	0	0	0	1	1
11:	0	0	1	0	1
12:	0	0	0	1	1
13:	0	0	0	2	2
14:	0	0	0	2	2
Total	2	36	71	838	947

$P(p = 0.001) = 0.781,\ P(p \leqslant 0.01) = 0.894$

$P^{-}_{FP} = 0.762$

likelihood ratio statistic λ / score-type statistic U

(i) $\theta_{14} = \theta_{15} = \theta_{20} = 3.0$

$l \backslash s$	0	1	2	3	Total
1:	1	12			13
2:	0	0	64		64
3:	0	0	0	*575	575
4:	0	0	0	159	159
5:	0	0	0	37	37
6:	1	0	2	20	23
7:	0	0	0	9	9
8:	0	0	0	8	8
9:	0	0	5	2	7
10:	0	0	2	1	3
11:	0	0	1	2	3
12:	0	0	1	15	16
13:	0	0	1	12	13
14:	0	0	0	9	9
15:	0	0	0	12	12
Total	2	12	76	861	951

$P(p = 0.001) = 0.803,\ P(p \leqslant 0.01) = 0.887$

$P^{-}_{FP} = 0.651$

$l \backslash s$	0	1	2	3	Total
1:	0	36			36
2:	0	0	67		67
3:	0	0	0	*659	659
4:	0	0	0	138	138
5:	0	0	0	20	20
6:	1	0	1	8	10
7:	0	0	0	3	3
8:	1	0	0	3	4
9:	0	0	2	1	3
10:	0	0	0	1	1
11:	0	0	1	0	1
12:	0	0	0	1	1
13:	0	0	0	2	2
14:	0	0	0	2	2
Total	2	36	71	838	947

$P(p = 0.001) = 0.781,\ P(p \leqslant 0.01) = 0.894$

$P^{-}_{FP} = 0.762$

likelihood ratio statistic λ / score-type statistic U

(i) $\theta_{14} = \theta_{15} = \theta_{20} = 3.0$

$l \backslash s$	0	1	2	3	Total
1:	1	12			13
2:	0	0	64		64
3:	0	0	0	*575	575
4:	0	0	0	159	159
5:	0	0	0	37	37
6:	1	0	2	20	23
7:	0	0	0	9	9
8:	0	0	0	8	8
9:	0	0	5	2	7
10:	0	0	2	1	3
11:	0	0	1	2	3
12:	0	0	1	15	16
13:	0	0	1	12	13
14:	0	0	0	9	9
15:	0	0	0	12	12
Total	2	12	76	861	951

$P(p = 0.001) = 0.803,\ P(p \leqslant 0.01) = 0.887$

$P^{-}_{FP} = 0.651$

$l \backslash s$	0	1	2	3	Total
1:	0	36			36
2:	0	0	67		67
3:	0	0	0	*659	659
4:	0	0	0	138	138
5:	0	0	0	20	20
6:	1	0	1	8	10
7:	0	0	0	3	3
8:	1	0	0	3	4
9:	0	0	2	1	3
10:	0	0	0	1	1
11:	0	0	1	0	1
12:	0	0	0	1	1
13:	0	0	0	2	2
14:	0	0	0	2	2
Total	2	36	71	838	947

$P(p = 0.001) = 0.781,\ P(p \leqslant 0.01) = 0.894$

$P^{-}_{FP} = 0.762$

likelihood ratio statistic λ / score-type statistic U

(ii) $\theta_{14} = 2.0, \theta_{15} = 3.0, \theta_{20} = 4.0$

$l \backslash s$	0	1	2	3	Total
1:	0	99			99
2:	0	0	23		23
3:	1	0	0	*632	633
4:	0	0	0	115	115
5:	0	0	0	29	29
6:	0	0	0	14	14
7:	0	0	0	5	5
8:	0	0	0	3	3
9:	0	0	13	1	14
10:	0	0	3	2	5
11:	0	0	12	2	14
12:	0	0	1	5	6
13:	0	0	0	12	12
14:	0	0	0	4	4
15:	0	0	0	5	5
Total	1	99	52	829	981

$P(p = 0.001) = 0.906,\ P(p \leqslant 0.01) = 0.957$

$P^{-}_{FP} = 0.754$

$l \backslash s$	0	1	2	3	Total
1:	0	271			271
2:	0	0	12		12
3:	0	0	0	*573	573
4:	0	0	0	88	88
5:	0	0	0	16	16
6:	0	0	0	5	5
7:	0	0	0	1	1
8:	0	0	0	1	1
9:	0	0	1	0	1
10:	0	0	1	1	2
11:	0	0	1	1	2
12:	0	0	0	0	0
13:	0	0	0	0	0
14:	0	0	0	1	1
15:	0	0	0	1	1
Total	0	271	15	688	974

$P(p = 0.001) = 0.853,\ P(p \leqslant 0.01) = 0.947$

$P^{-}_{FP} = 0.856$

likelihood ratio statistic λ / score-type statistic U

(ii) $\theta_{14} = 2.0, \theta_{15} = 3.0, \theta_{20} = 4.0$

$l \backslash s$	0	1	2	3	Total
1:	0	99			99
2:	0	0	23		23
3:	1	0	0	*632	633
4:	0	0	0	115	115
5:	0	0	0	29	29
6:	0	0	0	14	14
7:	0	0	0	5	5
8:	0	0	0	3	3
9:	0	0	13	1	14
10:	0	0	3	2	5
11:	0	0	12	2	14
12:	0	0	1	5	6
13:	0	0	0	12	12
14:	0	0	0	4	4
15:	0	0	0	5	5
Total	1	99	52	829	981

$P(p = 0.001) = 0.906,\ P(p \leqslant 0.01) = 0.957$

$P^{-}_{FP} = 0.754$

$l \backslash s$	0	1	2	3	Total
1:	0	271			271
2:	0	0	12		12
3:	0	0	0	*573	573
4:	0	0	0	88	88
5:	0	0	0	16	16
6:	0	0	0	5	5
7:	0	0	0	1	1
8:	0	0	0	1	1
9:	0	0	1	0	1
10:	0	0	1	1	2
11:	0	0	1	1	2
12:	0	0	0	0	0
13:	0	0	0	0	0
14:	0	0	0	1	1
15:	0	0	0	1	1
Total	0	271	15	688	974

$P(p = 0.001) = 0.853,\ P(p \leqslant 0.01) = 0.947$

$P^{-}_{FP} = 0.856$

likelihood ratio statistic λ / score-type statistic U

(ii) $\theta_{14} = 2.0, \theta_{15} = 3.0, \theta_{20} = 4.0$

$l \backslash s$	0	1	2	3	Total
1:	0	99			99
2:	0	0	23		23
3:	1	0	0	*632	633
4:	0	0	0	115	115
5:	0	0	0	29	29
6:	0	0	0	14	14
7:	0	0	0	5	5
8:	0	0	0	3	3
9:	0	0	13	1	14
10:	0	0	3	2	5
11:	0	0	12	2	14
12:	0	0	1	5	6
13:	0	0	0	12	12
14:	0	0	0	4	4
15:	0	0	0	5	5
Total	1	99	52	829	981

$P(p = 0.001) = 0.906,\ P(p \leqslant 0.01) = 0.957$

$P^{-}_{FP} = 0.754$

$l \backslash s$	0	1	2	3	Total
1:	0	271			271
2:	0	0	12		12
3:	0	0	0	*573	573
4:	0	0	0	88	88
5:	0	0	0	16	16
6:	0	0	0	5	5
7:	0	0	0	1	1
8:	0	0	0	1	1
9:	0	0	1	0	1
10:	0	0	1	1	2
11:	0	0	1	1	2
12:	0	0	0	0	0
13:	0	0	0	0	0
14:	0	0	0	1	1
15:	0	0	0	1	1
Total	0	271	15	688	974

$P(p = 0.001) = 0.853,\ P(p \leqslant 0.01) = 0.947$

$P^{-}_{FP} = 0.856$

More than 260

$\mu(A) = \{3.229, 5.347, 6.248\}$

likelihood ratio statistic λ

(iii) $\theta_{14} = 4.0, \theta_{15} = 3.0, \theta_{20} = 2.0$

$l\backslash s$	0	1	2	3	Total
1:	1	25			26
2:	1	0	325		326
3:	1	0	0	*369	370
4:	0	0	0	115	115
5:	1	0	0	29	30
6:	1	0	1	16	18
7:	0	1	0	6	7
8:	0	0	1	4	5
9:	0	0	2	3	5
10:	0	0	1	3	4
11:	0	0	2	2	4
12:	0	1	0	14	15
13:	0	0	0	8	8
14:	0	0	0	3	3
15:	0	0	0	8	8
Total	5	27	332	580	944

$P(p = 0.001) = 0.750,\ P(p \leq 0.01) = 0.842$

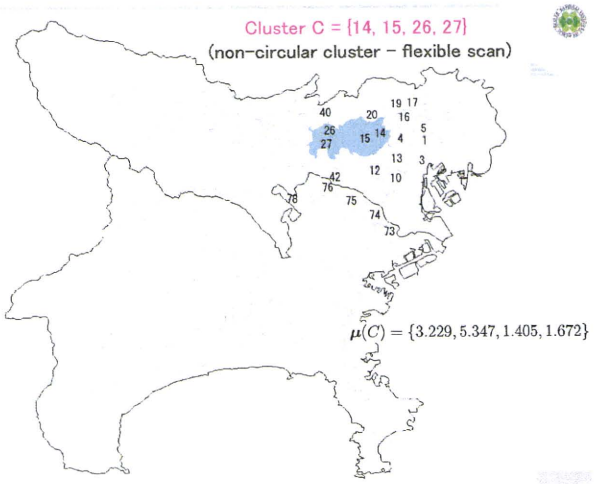
$P_{FP}^{-} = 0.719$

score-type statistic U

$l\backslash s$	0	1	2	3	Total
1:	1	26			27
2:	0	0	340		340
3:	1	1	0	*447	449
4:	0	0	0	98	98
5:	1	0	0	16	17
6:	0	0	0	3	3
7:	0	0	0	0	0
8:	0	0	2	3	5
9:	0	0	0	1	1
10:	0	0	0	2	2
11:	0	0	0	0	0
12:	0	0	0	2	2
13:	0	0	0	0	0
14:	0	0	0	1	1
Total	3	27	342	573	945

$P(p = 0.001) = 0.753,\ P(p \leq 0.01) = 0.882$

$P_{FP}^{-} = 0.813$



likelihood ratio statistic λ

(i) $\theta_{14} = \theta_{15} = \theta_{26} = \theta_{27} = 3.0$

$l\backslash s$	0	1	2	3	4	Total
1:	0	3				3
2:	0	0	11			11
3:	0	0	5	60		65
4:	0	0	8	47	*65	120
5:	0	1	6	49	77	133
6:	0	0	10	34	125	169
7:	0	0	14	30	105	149
8:	1	1	9	29	74	114
9:	0	1	7	18	46	72
10:	0	0	4	8	14	26
11:	0	0	2	3	4	9
12:	0	0	0	0	1	1
Total	1	6	76	278	511	872

$P(p = 0.001) = 0.497,\ P(p \leq 0.01) = 0.703$

$P_{FP}^{-} = 0.139$

score-type statistic U

$l\backslash s$	0	1	2	3	4	Total
1:	0	8				8
2:	0	0	79			79
3:	0	0	9	121		130
4:	0	0	10	64	*103	177
5:	0	0	6	29	79	114
6:	0	0	7	33	85	125
7:	0	0	4	26	61	91
8:	2	2	4	16	38	62
9:	0	1	6	12	27	46
10:	0	1	0	3	10	14
11:	0	0	1	1	1	3
Total	2	12	126	305	404	849

$P(p = 0.001) = 0.399,\ P(p \leq 0.01) = 0.647$

$P_{FP}^{-} = 0.311$

likelihood ratio statistic λ

(ii) $\theta_{14} = 3.0, \theta_{15} = 2.0, \theta_{26} = \theta_{27} = 4.0$

$l\backslash s$	0	1	2	3	4	Total
1:	0	13				13
2:	0	1	30			31
3:	0	0	6	19		25
4:	0	1	3	13	*55	72
5:	0	0	7	25	98	130
6:	0	0	13	31	123	167
7:	0	0	9	14	129	152
8:	1	0	6	22	88	117
9:	0	2	2	13	44	61
10:	0	0	0	3	34	37
11:	0	0	2	1	5	8
12:	0	0	0	0	3	3
Total	1	17	78	141	579	816

$P(p = 0.001) = 0.388,\ P(p \leq 0.01) = 0.606$

$P_{FP}^{-} = 0.117$

score-type statistic U

$l\backslash s$	0	1	2	3	4	Total
1:	0	11				11
2:	0	0	15			15
3:	0	0	4	72		76
4:	0	0	2	22	*130	154
5:	0	0	4	33	116	153
6:	0	1	9	39	119	168
7:	0	1	4	13	91	109
8:	1	0	2	14	65	82
9:	0	1	1	13	22	37
10:	0	0	1	4	14	19
11:	0	0	1	1	3	5
Total	1	14	43	211	560	829

$P(p = 0.001) = 0.368,\ P(p \leq 0.01) = 0.628$

$P_{FP}^{-} = 0.228$

likelihood ratio statistic λ

(iii) $\theta_{14} = 2.0, \theta_{15} = 3.0, \theta_{26} = 4.0, \theta_{27} = 2.0$

$l\backslash s$	0	1	2	3	4	Total
1:	0	8				8
2:	0	1	36			37
3:	1	3	17	53		74
4:	0	0	25	46	*33	104
5:	0	2	17	65	35	119
6:	2	1	15	72	58	148
7:	1	1	12	46	58	118
8:	0	1	7	38	50	96
9:	1	2	5	25	26	59
10:	0	0	3	12	13	28
11:	0	0	2	2	3	7
12:	0	0	0	0	1	1
Total	5	19	139	359	277	799

$P(p = 0.001) = 0.339,\ P(p \leq 0.01) = 0.565$

$P_{FP}^{-} = 0.130$

score-type statistic U

$l\backslash s$	0	1	2	3	4	Total
1:	0	24				24
2:	0	0	93			93
3:	1	2	26	106		135
4:	1	0	19	60	*39	119
5:	0	0	12	57	29	98
6:	0	1	5	34	47	87
7:	0	0	5	22	43	70
8:	0	0	7	24	37	68
9:	1	1	6	14	16	38
10:	0	0	2	7	6	15
11:	0	0	0	1	3	4
Total	3	28	175	325	220	751

$P(p = 0.001) = 0.260,\ P(p \leq 0.01) = 0.491$

$P_{FP}^{-} = 0.262$

Conclusions

- We proposed a maximum score-type scan statistic without assuming the hot-spot model using a new definition of the "most significant cluster".
- The statistic is derived from applying the Glaz and Zhang's procedure for i.i.d. variables.
- Several scenarios of simulation showed that the proposed statistic had shorter tails, and detects core areas of the assumed cluster without FPs.

Conclusions

- To satisfy the i.i.d. condition, the Anscombe's variance stabilization transformation is applied.
- It has been well known that the transformation is better for the Poisson model.
- The maximum score-type statistic in the approach is derived as the same statistic for combining p-values using inverse normal method.
- Although we did not assume "common p-value" within the window, we can examine to obtain U_i from $\Phi^{-1}(q_i)$ using *mid-p* values of $n_i \sim \text{Poisson}(\mu_i)$ without transformation. (Their simulation result were similar, but further research is needed)

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

研究成果の刊行物・別刷

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

© 2010, The International Biometric Society

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

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

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.