Loughborough University

This item was submitted to Loughborough's Research Repository by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Multiple-cluster detection test for purely temporal disease clustering: integration of scan statistics and generalized linear models

PLEASE CITE THE PUBLISHED VERSION

https://doi.org/10.1371/journal.pone.0207821

PUBLISHER

© The Authors. Published by Public Library of Science (PLoS)

VERSION

VoR (Version of Record)

PUBLISHER STATEMENT

This work is made available according to the conditions of the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. Full details of this licence are available at: http://creativecommons.org/licenses/ by/4.0/

LICENCE

CC BY 4.0

REPOSITORY RECORD

Takahashi, Kunihiko, and Hideyasu Shimadzu. 2019. “Multiple-cluster Detection Test for Purely Temporal Disease Clustering: Integration of Scan Statistics and Generalized Linear Models”. figshare. https://hdl.handle.net/2134/36197.

PLOS ONE

RESEARCH ARTICLE

Multiple-cluster detection test for purely temporal disease clustering: Integration of scan statistics and generalized linear models

Kunihiko Takahashi[1]*, **Hideyasu Shimadzu**[2]

1 Department of Biostatistics, Nagoya University Graduate School of Medicine, Nagoya, Japan,
2 Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire, United Kingdom

* kunihiko@med.nagoya-u.ac.jp

Check for updates

OPEN ACCESS

Citation: Takahashi K, Shimadzu H (2018) Multiple-cluster detection test for purely temporal disease clustering: Integration of scan statistics and generalized linear models. PLoS ONE 13(11): e0207821. https://doi.org/10.1371/journal.pone.0207821

Editor: Oliver Gruebner, University of Zurich, SWITZERLAND

Received: June 11, 2018

Accepted: November 5, 2018

Published: November 21, 2018

Copyright: © 2018 Takahashi, Shimadzu. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Daily ambulance records of out-of-hospital cardiac arrest (OHCA) cases were obtained from the All-Japan Utstein registry data of cardiopulmonary arrest patients provided by the Fire and Disaster Management Agency (FDMA). We obtained the data set from the Ministry of Internal Affairs and Communications, Japan. All the data set used in the present study is public data and available upon request to: fdma-goiken@ml.soumu.go.jp, referring to the All-Japan Utstein registry data 2005-2011.

Abstract

The spatial scan statistic is commonly used to detect spatial and/or temporal disease clusters in epidemiological studies. Although multiple clusters in the study space can be thus identified, current theoretical developments are mainly based on detecting a 'single' cluster. The standard scan statistic procedure enables the detection of multiple clusters, recursively identifying additional 'secondary' clusters. However, their *p*-values are calculated one at a time, as if each cluster is a primary one. Therefore, a new procedure that can accurately evaluate multiple clusters as a whole is needed. The present study focuses on purely temporal cases and then proposes a new test procedure that evaluates the *p*-value for multiple clusters, combining generalized linear models with an information criterion approach. This framework encompasses the conventional, currently widely used detection procedure as a special case. An application study adopting the new framework is presented, analysing the Japanese daily incidence of out-of-hospital cardiac arrest cases. The analysis reveals that the number of the incident increases around New Year's Day in Japan. Further, simulation studies undertaken confirm that the proposed method possesses a consistency property that tends to select the correct number of clusters when the truth is known.

Introduction

Whether the distribution of disease spreads randomly or clusters around particular epicenters, this has been a crucial concern in epidemiological studies. Any indication of disease clustering at an early phase of outbreak offers valuable insights into preventing us from a worse pandemic scenario or may provide us with clues to the etiology of the disease [1, 2]. A well-known monumental work would be one carried by John Snow, a British physician, in identifying the source of a cholera outbreak in Soho, London, during the middle of 19th century. He identified a water supply pump as the source of the epidemic by mapping the number of cholera victims which delineated a cluster around the pump [3].

Although the extent to which disease distributes is often discussed in a spatial context, like Snow's map, there is parallel to purely temporal events which also attract attention from a

Funding: KT acknowledges the support by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 17K00046. https://www.jsps.go.jp/english/e-grants/.

Competing interests: The authors have declared that no competing interests exist.

broad public today. For example, identifying the threat of emerging infections or the risk of bioterrorist attacks is a critical role of surveillance systems, the main focus of which is to monitor the incidence or prevalence of specific health problems over time within a well-defined population [4].

To identify meaningful clusters, in other words, to investigate a regional or temporal tendency in the presence of certain diseases, whether the disease risk is relatively high to other surrounding regions or subsequent time periods, a number of statistical tests have been proposed and are widely used [5]. These tests are classified based on their purpose. For example, focused tests have been developed to detect the existence of a local cluster around a predetermined point source, while general tests search for clusters without any preconceived assumptions about their location [6]. The general test framework further consists of two types of tests. Global clustering tests, such as those developed by Moran [7], Whitemore *et al* [8], Oden [9], Tango [10], Rogerson [11] and Bonetti and Pagano [12], detect the presence of clusters in a study area without determining the statistical significance of individual clusters. On the other hand, cluster detection tests (CDTs), such as those proposed by Besag and Newell [6], Turnbull *et al*. [13], Kulldorff and Nagarwalla [14], Kulldorff [15, 16], Tango [17, 18], Duczmal and Assunção [19], Tango and Takahashi [20, 21], Takahashi *et al*. [22], and Jung [23], possess the flexibility to accommodate spatial, temporal, or spatio-temporal data, and enables the determination of whether a disease pattern is completely random over the space of study without prior information while indicating regions or time periods with high disease prevalence.

The scan statistic is one of the most powerful elements of the CDT since it is based on a concrete statistical framework—i.e., the maximum likelihood ratio; examples include the circular scan statistic [15] along with the SaTScan software [24] and the flexibly shaped scan statistic [20] implemented in the FleXScan software [25]. These have widely been applied to important challenges in various fields such as the epidemiology of cancer and other diseases, infectious disease surveillance, parasitology, psychology, ambulance logistics, medical imaging, genome-wide association analyses, and drug and vaccine development [24, 26–28].

A shortcoming of these approaches is, however, the fact that most of them focus on 'single' cluster detection while investigating the extended study space or period within which more than one cluster is expected. To detect more than one cluster, the ordinary scan statistic procedure, including the circular and flexibly shaped ones, is iteratively applied after the identification of the first (primary) cluster; additional, mutually exclusive 'secondary' clusters are then sequentially detected by the likelihood ratio statistic—we hereafter refer to this conventional procedure as the secondary-cluster procedure. The procedure can only evaluate these clusters one by one, and each corresponding p-value is calculated as if the cluster were the primary one. This fact indicates that the current approach does not provide an accurate assessment of the selected multiple clusters.

In the present work, we construct a general test procedure that enables the simultaneous evaluation of multiple clusters, focusing on a purely temporal Poisson model. Combining generalized linear models (GLMs) and the ordinary scan statistic procedure, the new testing framework stands directly on the full-likelihood principle that can easily be amalgamated with an information criterion approach to select clusters via GLMs. This procedure becomes, as will be described in a later section, a natural extension of scan statistic—i.e., the conventional secondary-cluster procedure, and can accurately evaluate multiple clusters as a whole. An application study adopting the proposed procedure is then discussed in the context of a real-world example—i.e., temporal data on the daily incidence of out-of-hospital cardiac arrest cases in Japan [28]. The results are compared to those obtained by the secondary-cluster procedure. The consistency property of the proposed procedure, a desirable property, whether it tends to select the correct number of clusters is also examined via simulations studies.

Methods

In this section, we first present an overview of the ordinary scan statistic framework implemented in SaTScan and FleXScan for a single-cluster detection and then describe the conventional secondary-cluster procedure for the multiple-cluster detection proposed by Kulldorff [15] using a Poisson GLM as described in Jung [23]. We then propose a new test procedure that simultaneously evaluates multiple clusters using both GLMs and an information approach. We also show that the new procedure encompasses the secondary-cluster procedure as a special case.

Ordinary scan statistic for single-cluster detection and secondary-cluster procedure

A study space (area or time period) G consists of m segments, each of which corresponds to the smallest element of the space; this can be a county, month, or day depending on the type of data. The number of cases within segment i—i.e., Y_i—is presumed to follow an independent Poisson distribution, with an expected value μ_i—i.e., $Y_i|\mu_i \sim \text{Poisson}(\mu_i)$, which is henceforth denoted in lowercase as y_i, $i = 1, 2, \ldots, m$. Let $\mathcal{W}$ denote the set of all potential scanning zones (the sets of connected segments) of any size, construction of which relies on an employed scanning method. Assuming the window $w\ (\in \mathcal{W})$ as a hot-spot cluster in which the number of cases, y_i, is higher than in other parts of the study space, the expected number of cases can be modelled as

$$\log\ \mu_i = \log(\theta_i \mu_i^0) = \alpha + \beta z_i + \log\ \mu_i^0, \tag{1}$$

with the indicator variable $z_i = 1$ if $i \in w$ and $z_i = 0$ otherwise ($i \notin w$). For those segments that fall into a hot-spot cluster (w), a parameter of model (1) becomes $\theta_i = \theta_w = \exp(\alpha + \beta)$ for $\beta \geq 0$. On the other hand, for those that fall outside the cluster ($\bar{w}$), the parameter is $\theta_i = \theta_{\bar{w}} = \exp(\alpha)$. The constant term $\mu_i^0 \coloneqq \mu_i^0(\boldsymbol{x}_i)$ is often modelled as a function of other covariates $\boldsymbol{x}_i$, such as demographic or environmental factors; this yields the null model—i.e., the expected number of cases if there is no cluster in the study space such that $\beta = 0$. The null model is therefore $\log\ \mu_i = \alpha^0 + \log\ \mu_i^0$.

The likelihood function of model (1) is generated as follows: $f_i(y_i|\boldsymbol{z}, \boldsymbol{\psi}) = f(y_i|\mu_i^0, \boldsymbol{z}, \boldsymbol{\psi})$ is the probability function of $Y_i = y_i$ given the location of hot-spot window, $\boldsymbol{z} \coloneqq \boldsymbol{z}(w) = (z_1, z_2, \ldots, z_m)$, and the parameters $\boldsymbol{\psi} = (\alpha, \beta)$ and μ_i^0. The probability function $f(y_i|\mu_i^0, \boldsymbol{z}, \boldsymbol{\psi})$ can be expressed as either $f(y_i|\mu_i^0, \boldsymbol{z}, \theta_w)$ for $i \in w$ or $f(y_i|\mu_i^0, \boldsymbol{z}, \theta_{\bar{w}})$ for $i \in \bar{w}$, $w \cap \bar{w} = \emptyset$. Hence, the conditional log likelihood function is $l(\boldsymbol{\psi}|\boldsymbol{z}) \coloneqq \log\{\prod_{i=1}^m f_i(y_i|\boldsymbol{z}, \boldsymbol{\psi})\} = \log\left[\prod_{i=1}^m \{f(y_i|\mu_i^0, \boldsymbol{z}, \theta_w)\}^{z_i} \{f(y_i|\mu_i^0, \boldsymbol{z}, \theta_{\bar{w}})\}^{1-z_i}\right]$. This leads to the following statistical hypothesis test to detect a single cluster:

$$H_0 : \theta_w = \theta_{\bar{w}} \ \text{ for } \forall w \in \mathcal{W};\ \ H_1 : \theta_w > \theta_{\bar{w}} \ \text{ for } \exists w \in \mathcal{W}.$$

The null hypothesis H_0 states that there is no cluster—i.e., $\boldsymbol{z} = \boldsymbol{0}$, whereas the alternative hypothesis H_1 asserts that there is a hot-spot window w—namely, $\boldsymbol{z} = \boldsymbol{z}(w)$—in which the underlying disease risk is higher than in the other parts of the study space. It is important to note that the two hypotheses can also be expressed in equivalent forms with respect to β as H_0: $\beta = 0$ for $\forall w$ and H_1: $\beta > 0$ for $\exists w$.

The maximum log likelihood ratio (LLR) for a given w is

$$LLR(w) = l(\boldsymbol{\psi} = (\alpha, \beta)|\boldsymbol{z} = \boldsymbol{z}(w)) - l(\boldsymbol{\psi} = (\alpha^0)|\boldsymbol{z} = \boldsymbol{0}) \tag{2}$$

and the conventional scan statistic is defined as $\max_{w \in \mathcal{W}} LLR(w)$. A window w^* that attains

$\max_{w\in\mathcal{W}} LLR(w)$—i.e., $\max_{w\in\mathcal{W}}\ l(\boldsymbol{\psi}|\boldsymbol{z}=\boldsymbol{z}(w))$—, is identified as the most likely (primary) cluster. The Monte Carlo method is commonly used to evaluate statistical significance. A large number of random datasets are generated under the null hypothesis, and the test statistic is then calculated for each dataset; the upper or lower percentiles provide a Monte Carlo-based *p*-value.

The described procedure above was intended to identify only the primary cluster, $w_1^* = w^*$. For multiple cluster detection, Kulldorff [15] extended its use; the procedure was repeatedly used to identify other clusters, namely 'secondary' clusters, $w_2^*, w_3^*, \ldots$, among which there were no overlaps—i.e., $w_k^* \cap w_{k'}^* = \emptyset$ for $k \neq k'$ and consequently, their LLRs always followed a descending order, $LLR(w_1^*) > LLR(w_2^*) > \cdots > LLR(w_k^*) > \cdots$. The statistical significance of secondary clusters was evaluated in the same way as that of the most likely cluster; i.e., the log-likelihood ratio of each secondary cluster was compared to the maximum log-likelihood ratio calculated from randomly generated data sets. This procedure is simple and easy to use, although it is limited by the fact that it only evaluates clusters individually as if each one were the primary cluster, and does not assess multiple clusters as a whole.

A new framework for multiple-cluster detection test procedure

To construct a new procedure, we first show here that the multi-cluster model can be formulated in the form of mixture Poisson GLMs and then derive the likelihood function. This new formulation recognizes the cluster selection problem as a model selection problem for which we can embrace the information criterion paradigm. To choose an appropriate number of clusters, we then propose a new criterion delivered from the likelihood function in the same manner as BIC. The computational aspect of the proposed methods is reasonably straightforward utilizing existing statistical software although, some technical details are explained at the end of this section.

Multiple cluster model and its likelihood. Assuming that there are K clusters [$\boldsymbol{w} = (w_1, w_2, \ldots, w_K)$] in a space G, each mutually exclusive window w_k contains a set of adjacent segments as a cluster—i.e., $w_k \cap w_{k'} = \emptyset$ for $k \neq k'$. Note that $K = 0$ and $K = 1$ indicate that there is no cluster and a single cluster in the study space, respectively. The expected number of cases μ_i of segment i can be modelled as an extension of (1) as follows:

$$\log \mu_i = \log(\theta_i \mu_i^0) = \alpha + \sum_{k=1}^{K} \beta_k z_{ki} + \log \mu_i^0 \tag{3}$$

for $K \geq 1$ and $\log \mu_i = \alpha^0 + \log \mu_i^0$ for $K = 0$. Here, the indicator variable $z_{ki} = 1$ if $i \in w_k$ and $z_{ki} = 0$ otherwise. Note that $\beta_k > 0$.

Recalling the notation in the previous section, $f(y_i|\mu_i^0, \boldsymbol{z}, \boldsymbol{\psi})$ is the probability function of $Y_i = y_i$ given $\boldsymbol{z} := \boldsymbol{z}(w) = (z_{ki})$—which is now a $K \times m$ matrix—and the parameters $\boldsymbol{\psi} = (\alpha, \beta_1, \beta_2, \ldots, \beta_K)$. The conditional log likelihood function can be written as $l(\boldsymbol{\psi}|\boldsymbol{z}) := \log\left[\prod_{i=1}^{m}\prod_{k=0}^{K}\{f(y_i|\mu_i^0, \boldsymbol{z}, \boldsymbol{\psi})\}^{z_{ki}}\right]$, where $z_{0i} = 1$ if $i \notin \cup_{k=1}^{K} w_k$, and $z_{0i} = 0$ otherwise. We assume $\boldsymbol{z}$ to be randomly selected in accordance with a probability function $h(\boldsymbol{z})$. The complete (full) log likelihood function of $\boldsymbol{\psi}$ is then expressed as:

$$l(\boldsymbol{\psi}) = \log L(\boldsymbol{\psi}) = \log\left[\prod_{i=1}^{m}\prod_{k=0}^{K}\{f(y_i, z_{ki}|\mu_i^0, \boldsymbol{\psi})\}^{z_{ki}}\right] = l(\boldsymbol{\psi}|z) + \log\{h(\boldsymbol{z})\}$$

where $L(\boldsymbol{\psi})$ is the likelihood function of $\boldsymbol{\psi}$. The maximum log likelihood ratio statistic $LLR(w)$ (2) is equivalent to the maximum of $l(\boldsymbol{\psi})$ for $K = 1$ with $h(\boldsymbol{z})$, which assumes a constant probability under the alternative hypothesis H_1.

Candidates of multiple clusters $\boldsymbol{w}$. The multiple-cluster model (3) depends on the choice of $\boldsymbol{z}$—i.e., $\boldsymbol{w}$ among a large number of combinations of sets in $\mathcal{W}$. Although the choice is, indeed, a search problem in a general context, we here focus on extending the secondary-cluster procedure due to its natural interpretation. We are interested in evaluating whether each of candidate clusters is able to reject the null hypothesis on its own strength, significance in other words, in terms of the likelihood statistic value. In some sense, it is like a regression analysis where each variable is entered in a separate regression model and then evaluated without adjusting for other variables [29]. Thus, we adopt the conventional scan statistic approach to select candidate clusters. First, the primary cluster w_1^* is selected as candidate w_1 in $\boldsymbol{w}$; then, with a pre-defined maximum number of candidates, among which there are no overlaps—i.e., K_{max} (≥ 1)—($K_{max} - 1$) secondary clusters, $w_2^*, w_3^*, \ldots, w_{K_{max}}^*$, are selected one by one, regardless of their significance. In practice, we predefine the maximum number of candidates (e.g., $K_{max} = 10, 20, \ldots$) or a P-value threshold, p_s, (e.g., $p_s < 0.5, 0.8, 1.0$). It should be noted that $K_{max} = 1$ corresponds to detection of a single cluster using the conventional scan statistic procedure. Candidate selection may differ depending on the scanning method that is adopted (e.g., circular, flexible, etc.).

Selecting an appropriate K and significance test. To select an appropriate number of clusters, $K(\leq K_{max})$, we propose a new information criterion approach that chooses K in favor of the largest marginal likelihood, $ML(\boldsymbol{y}, \boldsymbol{z}) = \int \exp\{\log L(\boldsymbol{\psi})\}\, g(\boldsymbol{\psi}) d\boldsymbol{\psi}$, where $g(\boldsymbol{\psi})$ is a prior probability function of parameter $\boldsymbol{\psi}$. This can be achieved by the same manner in deriving the Bayesian information criterion (BIC). Applying the Taylor expansion and Laplace approximations to the integral above [30], the log marginal likelihood is approximated as

$$\begin{aligned} -2\log ML(\boldsymbol{y},\boldsymbol{z}) \quad \approx \quad & -2\sum_{i=1}^{m}\sum_{k=0}^{K} z_{ki}\{\log f(y_i|\mu_i^0, \boldsymbol{z}, \hat{\boldsymbol{\psi}})\} - 2\log(h(\boldsymbol{z})) \\ & + q\log m + \log|J(\hat{\boldsymbol{\psi}})| - q\log(2\pi) - 2\log(g(\hat{\boldsymbol{\psi}})) \end{aligned}$$

where $\hat{\boldsymbol{\psi}}$ is the maximum likelihood estimator of $\boldsymbol{\psi}$,

$$J(\hat{\boldsymbol{\psi}}) = -\frac{1}{m}\frac{\partial^2 l(\boldsymbol{\psi}|z)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$$

and $q = K + 1$. The model evaluation criterion can then be obtained by omitting terms with an order less than $O(1)$ with respect to the large sample size m; that is,

$$C(K) = -2l(\hat{\boldsymbol{\psi}}|\boldsymbol{z}) - 2\log(h(\boldsymbol{z})) + (K+1)\log m, \quad (K \geq 1). \tag{4}$$

To select an appropriate number of clusters, K, we define a relative difference statistic based on the criterion $C(K)$ as

$$RDC(K) = (C_0 - C(K))/C_0,$$

where $C_0 = C(0)$, the criterion under the null model. Appropriate multiple clusters are selected from the set of candidates $\tilde{\boldsymbol{w}} = (w_1, w_2, \ldots, w_K)$ with respect to $\max_K RDC(K)$. It should be noted that there is a clear link to the conventional cluster detection approach: the candidate cluster $\tilde{\boldsymbol{w}}$ attains $\max_w LLR(w)$ if $K = 1$ under $K_{max} = 1$.

To calculate the proposed criterion (4), the probability function $h(\boldsymbol{z})$ needs to be specified; for purely temporal cases, we recommend to use $h(\boldsymbol{z}) = (1/m)^K$ as an approximation of the probability of selecting locations $\boldsymbol{w}$ when the window size is relatively very small, $\#\{i|i \in \boldsymbol{w}\} \ll m$, with respect to the whole data size m. Hence, a temporal cluster selection

criterion is given as

$$C(K) = -2l(\hat{\boldsymbol{\psi}}\,|z) + (3K+1)\log m \qquad (K \geq 1).$$

The statistical significance of appropriate models is evaluated by the Monte Carlo hypothesis testing procedure in the same manner as with the standard scan statistic. Under the null hypothesis, a large number of random datasets are generated; however, for each of these, $\max_K RDC(K)$ is instead calculated as a test statistic.

Computation. All the numerical computation can easily be carried out by R [31] using the `glm` function that fits model (3) and then calculates the log-likelihood ratios. The Monte Carlo-based p-value [32] is computed in the following way. First, we calculate the observed test statistic based on the observed data set. Second, we compute the same statistic for a large number N_{rep} of data sets which are simulated independently under the null hypothesis of no clustering. The simulated p-value is obtained from the rank of the statistic among $N_{rep} + 1$ values of the statistic. If this rank is R, then p-value = $R/(N_{rep} + 1)$. The Monte Carlo simulated p-value is not necessary the same value all the time due to the randomly simulated data. In other words, another independent set of N_{rep} realisations will result in a slightly different p-value. However, a larger number of replications N_{rep} consequently provides a more stable p-value. The number of replication N_{rep} is usually set as 999 or 9,999.

Application and simulation data

To illustrate how the proposed procedure performs in detecting temporal multiple-clusters, we apply it to real-world data, the daily out-of-hospital cardiac arrest (OHCA) cases in Japan and compare the results with those obtained from the conventional secondary-cluster procedure. Further, the consistency of the proposed procedure is also investigated via a simulation study. The aspect of consistency, whether the proposed procedure tends to select the correct number of clusters when the truth is known, is an essential part in evaluating the performance of the proposed framework and the desired property as a reliable procedure.

An application: Detecting temporal clusters in daily incidence of out-of-hospital cardiac arrest cases

As an example, the Japanese OHCA data for the period of 2005–2011 (2,260 days) were studied and a total of 701,651 cases (MNC; male con-cardiac cases) were analyzed following a previous work [28]. The expected number of OHCA cases μ_i^0 on day i in model (3)—that is, the null expectation with no cluster—was estimated by a Poisson regression model for the observed OHCA cases y_i, while accommodating the following five factors into the model: year, month, day of the week, holidays, and temperature, in two-by-two stratification by sex (male/female) and the etiology of arrest (cardiac/non-cardiac). For more details see Takahashi and Shimadzu [28].

To determine whether OHCA incidence showed specific temporal clustering patterns, we calculated the scan statistic implemented with the restricted likelihood ratio [18, 21], which was an improved version of Kulldorff's standard scan statistic [15]. We set two default arguments of the program: the maximum temporal length of a cluster was 20 days and the pre-specified significance level for a restriction was $\alpha_1 = 0.2$. The significance level of the test was set as 0.05, and the p-value was calculated from 999 replications of the Monte Carlo hypothesis test.

Data and data accessibility. Daily ambulance records of OHCA cases were obtained from the All-Japan Utstein registry data of cardiopulmonary arrest patients provided by the Fire and Disaster Management Agency (FDMA). This was a nationwide and population-based registry system of OHCA cases available since 2005, in accordance with the Utstein guidelines.

Table 1. Assumed cluster periods with expected counts under the null model in simulation studies.

	period	# days	expected counts
A	2006/01/01–03	3	(115.10, 131.43, 122.34)
B	2005/01/01–03	3	(103.08, 108.53, 124.18)
C	2005/04/01–03	3	(76.44, 77.10, 84.35)
D	2005/02/01	1	(87.27)
E	2007/01/01–05	5	(127.77, 117.05, 114.54, 99.36, 100.34)
F	2007/04/01–05	5	(84.83, 81.85, 78.47, 79.52, 82.61)

https://doi.org/10.1371/journal.pone.0207821.t001

Since all the records were made anonymous by FDMA, according to the informed consent guidelines in Japan, we were exempt from obtaining informed consent from each patient to use this dataset. All the data set used in the present study is public data and available from the Ministry of Internal Affairs and Communications, Japan upon request at fdma-goiken@ml.soumu.go.jp, referring to the All-Japan Utstein registry data 2005–2011.

Simulation studies

We carried out simulation studies to assess the performance of the proposed framework in detecting multiple clusters from time-series data. The purpose of it is to check if the proposed framework possesses a consistent property that tends to select the correct number of clusters when the truth is known. Here, the expected count of daily OHCA MNC cases from the null hypothesis served as an example time series. We assumed six periods [A–F; Table 1], during which the incidence of OCHA cases—that is, the relative risk (RR) of OCHA—was set at a high value, and we introduced seven scenarios with different RR for each of the periods [Table 2]. For example, Scenario S2.1 had three cluster periods (A–C) for which the RR was set as 2.0; for the remainder of the time series, including periods D–F, the RR was 1.0, indicating that there was no cluster. We generated 1000 datasets for each scenario and compared the estimated power calculated from the two cluster detection tests with the significance level of 0.05.

Results

An application: Detecting temporal clusters in daily incidence of out-of-hospital cardiac arrest cases

Fig 1 shows the number of observed cases y_i and the expected counts μ_i^0 under the null hypothesis for male non-cardiac cases (MNC) with 185,819 cases.

Table 2. Assumed relative risks in simulation studies.

scenario	# clusters	# days	Relative Risk (RR)*					
			A	B	C	D	E	F
S0	0	0						
S1	1	3	1.5					
S2.1	3	9	1.2	1.2	1.2			
S2.2	3	9	1.3	1.3	1.3			
S2.3	3	9	1.5	1.5	1.5			
S2.4	3	9	2.0	2.0	2.0			
S3.1	6	20	2.0	2.0	2.0	2.0	2.0	2.0
S3.2	6	20	1.3	1.5	2.0	2.0	1.3	2.0

* All the blanks should read as Relative Risk (RR) is 1.

https://doi.org/10.1371/journal.pone.0207821.t002

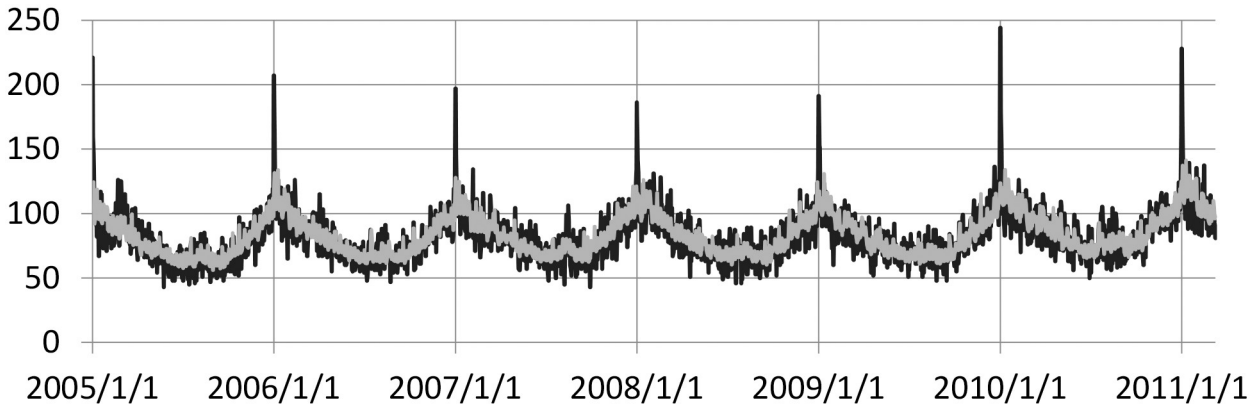


Fig 1. The black solid line represents the daily out-of-hospital cardiac arrest (OHCA), male non-cardiac cases (MNC), in Japan from 1 January 2005 to 10 March 2011 (m = 2, 260 days); total 185,819 cases, and maximum and minimum of daily counts were 244 and 43, respectively. Grey line overlays the null expected counts (Takahashi and Shimadzu [28]).

https://doi.org/10.1371/journal.pone.0207821.g001

The conventional secondary-cluster procedure first detected the most likely cluster of 2-day length and then seven other secondary clusters (Table 3). It should be noted that since random numbers were used in the Monte Carlo procedure, our p-values slightly differed from those previously reported [28].

We then applied our proposed framework to the same data. Candidate clusters were selected with a threshold p-value $p_s < 1.0$, which was equivalent to setting $K_{max} = 25$ for the observed data. The suggested multiple cluster model had eight clusters, $K = 8$ [$C_0 = 17433.89$, $C = C(8) = 17041.46$, $RDC(8) = 0.02251$].

Fig 2 shows the null distribution of $RDC(K)$ generated by the Monte Carlo-simulated data with 999 replications; this was similar to the results obtained by the secondary-cluster procedure. However, the proposed procedure calculated a p-value of 0.001 for the multiple-cluster model from the upper fifth percentile of the statistic $RDC_{5\%} = -0.00025$. The selected multiple-cluster model by the proposed procedure is shown in Table 4.

Fig 3 shows the value of the proposed criterion C for each K calculated from the observed data along with other criteria such as −2log likelihood, Akaike information criterion (AIC), and Bayesian information criterion (BIC). Our criterion C reached the minimum value—i.e., $\max_K RDC(K)$—at $K = 8$, unlike the other criteria that monotonically decreased with an inflection point at around $K = 8$.

Table 3. Detected significant temporal clusters in daily incidence of out-of-hospital cardiac arrest (OHCA), male non-cardiac cases (MNC), by the secondary-cluster procedure.

rank	cluster	clustered period	cases	expects	RR	p-value p_s
1	w_1	2010/01/01—2010/01/02	423	232.43	1.82	0.001
2	w_2	2005/01/01—2005/01/02	381	211.61	1.80	0.001
3	w_3	2011/01/01	228	111.40	2.05	0.001
4	w_4	2008/12/31—2009/01/03	630	444.54	1.42	0.001
5	w_5	2006/01/01	207	115.10	1.80	0.001
6	w_6	2008/01/01—2008/01/04	614	446.72	1.37	0.001
7	w_7	2007/01/01—2007/01/02	344	244.83	1.41	0.001
8	w_8	2011/01/02—2011/01/06	711	589.98	1.21	0.011

RR: relative risk

https://doi.org/10.1371/journal.pone.0207821.t003

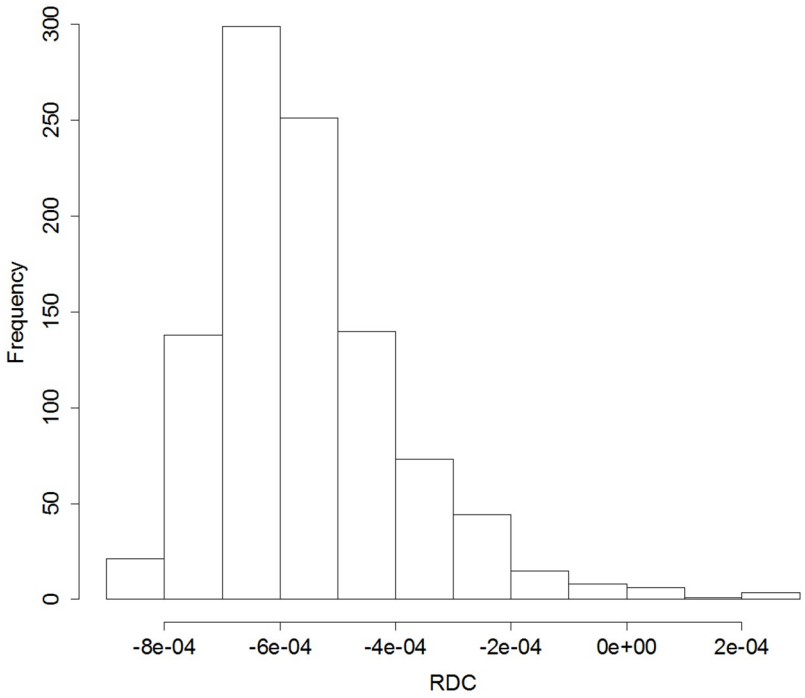


Fig 2. Histogram of the null distribution of *RDC*(*K*).

https://doi.org/10.1371/journal.pone.0207821.g002

Takahashi and Shimadzu [28] also investigated temporal clusters for other three groups such as male cardiac (MC) with 224,661 cases, female non-cardiac (FNC) with 128,491 cases, and female cardiac (FC) with 162,680 cases. Using the conventional secondary-cluster procedure, nine clusters (MC and FNC) and seven clusters (FC) were identified at the significance level $p_s < 0.05$ based on max$LLR(w)$. On the other hand, our proposed procedure yielded six clusters (MC and FNC) and five clusters (FC), both with p-value of 0.001. Most of the periods excluded by the proposed procedure among the secondary clusters of the conventional procedure were not around New Year's Day with $p_s > 0.01$.

Simulation studies

We generated 1,000 datasets for each scenario and compared the estimated power calculated from the two cluster detection tests with the significance level 0.05. Table 5 shows the total

Table 4. Detected multiple-cluster model by the proposed procedure.

	clustered period	coef.	OR	95%CI	p-value
intercept		−0.006			0.0077
w_1	2010/01/01—2010/01/02	0.605	1.831	(1.662, 2.012)	< 0.0001
w_2	2005/01/01—2005/01/02	0.594	1.812	(1.636, 2.000)	< 0.0001
w_3	2011/01/01	0.722	2.059	(1.803, 2.339)	< 0.0001
w_4	2008/12/31—2009/01/03	0.355	1.426	(1.317, 1.541)	< 0.0001
w_5	2006/01/01	0.593	1.810	(1.574, 2.068)	< 0.0001
w_6	2008/01/01—2008/01/04	0.324	1.383	(1.276, 1.500)	< 0.0001
w_7	2007/01/01—2007/01/02	0.346	1.414	(1.270, 1.569)	< 0.0001
w_8	2011/01/02—2011/01/06	0.193	1.213	(1.126, 1.304)	< 0.0001

coef.: estimated coefficients; OR: odds ratio; 95%CI: its 95% confidence interval; and p−value: the p-value of the estimated coefficient.

https://doi.org/10.1371/journal.pone.0207821.t004

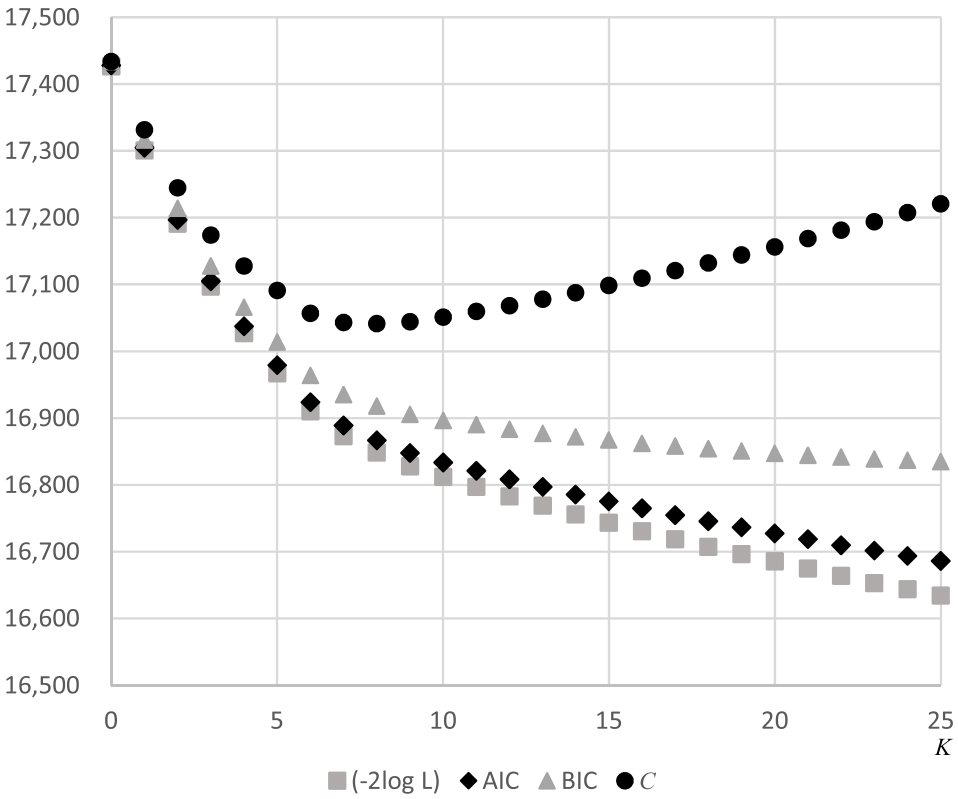


Fig 3. Values of the proposed criterion *C* for each *K*, with other criteria, −2log likelihood (−2log L), AIC and BIC.

https://doi.org/10.1371/journal.pone.0207821.g003

power, which was calculated as the sum of rejection rates over the number of clusters with $K > 0$. We calculated the sensitivity (Sen) and positive predictive value (PPV) of days detected as significant, as well as their averages and the number of detection with Sen = 1 and PPV = 1 among 1, 000 sets.

For the null scenario S0, the total power—i.e. probability of type I error—was 0.049 for both procedures and was very close to the significance level 0.05. In the single-cluster scenario [S1 (RR = 1.5)], both procedures had a total power equal to 1.0 but ours performed slightly better in the detection of a single cluster ($K = 1$). The total power for the three cluster scenarios [S2 (i)–(iv)] varied depending on RR. The power was lower—almost 0.57—for RR = 1.2, and increased to 0.991 for RR = 1.3. The total power for detection of $K = 3$ was also low for RR = 1.2 and 1.3; notably, they were lower for the proposed procedure as compared to the secondary-cluster procedure. On the contrary, the proposed procedure performed better in detecting $K = 3$ for RR = 1.5 and 2.0. Lastly, for the scenario with six clusters (S3), the total power was 1.0 for both procedures and all scenarios. The detection power of $K = 6$ clusters by the proposed procedure was slightly higher for scenario (i) with RRs of 2.0, while the secondary-cluster procedure showed higher power for scenario (ii) with different RRs. As such, the total power was almost the same for both procedures, and the detection power in terms of selecting the correct number of multiple clusters varied according to the scenario. The sensitivity of the secondary-cluster procedure was slightly higher than that of the proposed procedure, although the latter performed better in terms of PPV and the number detected with PPV = 1, suggesting that the secondary-cluster procedure tends to detect additional clusters.

Table 5. The power of the secondary-cluster and the proposed procedures in the simulation study, along with the sensitivity (Sen) and the positive predictive value (PPV) of days detected significantly.

	N.S. ($K = 0$)	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	total power	Sen (avg*)	Sen = 1	PPV (avg*)	PPV = 1
S0: null RR = 1.0													
s-c proc.	0.951	0.049							0.049	—	—	—	—
proposed proc.	0.951	0.049							0.049	—	—	—	—
S1: one cluster (three days) RR = 1.5													
s-c proc.		0.959	0.040	0.001					1.000	0.996	0.988	0.971	0.919
proposed proc.		0.994	0.006						1.000	0.996	0.988	0.986	0.954
S2.1: three clusters (nine days) RR = 1.2													
s-c proc.	0.430	0.436	0.123	0.011					0.570	0.361	0.003	0.870	0.376
proposed proc.	0.426	0.531	0.043						0.574	0.313	0.000	0.878	0.405
S2.2: three clusters (nine days) RR = 1.3													
s-c proc.	0.009	0.118	0.455	0.406	0.012				0.991	0.728	0.266	0.935	0.658
proposed proc.	0.009	0.315	0.469	0.206	0.001				0.991	0.605	0.140	0.946	0.734
S2.3: three clusters (nine days) RR = 1.5													
s-c proc.			0.004	0.962	0.034				1.000	0.992	0.940	0.978	0.847
proposed proc.			0.011	0.984	0.005				1.000	0.990	0.934	0.984	0.872
S2.4: three clusters (nine days) RR = 2.0													
s-c proc.				0.960	0.039	0.001			1.000	1.000	1.000	0.991	0.957
proposed proc.				0.990	0.010				1.000	1.000	1.000	0.997	0.987
S3.1: six clusters (20 days) RR = 2.0													
s-c proc.							0.976	0.024	1.000	0.999	0.999	0.997	0.972
proposed proc.						0.002	0.989	0.009	1.000	0.999	0.997	0.998	0.987
S3.2: six clusters (20 days) RR = 1.3, 1.5, 2.0													
s-c proc.					0.003	0.170	0.795	0.032	1.000	0.952	0.597	0.976	0.669
proposed proc.					0.012	0.276	0.704	0.008	1.000	0.934	0.526	0.979	0.699

* averages among the custers detected as $K > 0$; Sen = 1: #{$Sen = 1$}/1000; PPV = 1: #{$PPV = 1$}/1000; s-c proc.: secondary-cluster procedure.

https://doi.org/10.1371/journal.pone.0207821.t005

Discussion

We have proposed a general test procedure that enables the simultaneous evaluation of multiple clusters as an extension of the conventional secondary-cluster procedure, focusing on a purely temporal Poisson model. The Japanese OHCA data analysis has highlighted the most advantageous aspect of the new procedure, that is, it can evaluate the p-value for the whole multiple clusters, as opposed to the secondary-cluster procedure which cannot. This aspect becomes a key to the detection of meaningful multiple clusters. Our procedure has detected the equal or less number of clustered periods in the OHCA data, compared to ones of the secondary-cluster procedure [28]. This may seem to be a subtle difference although, it requires more careful interpretation regarding the p-value reported by Monte Carlo simulated data. Note that the individual p-values, p_s, of the clustered periods commonly identified by both of the procedures state high significance with $p_s = 0.001$. On the other hand, for those clusters being excluded by the proposed procedure indicate relatively large p-values. It is not difficult to imagine that such p-values may vary due to Monte Carlo simulated data with 999 replications. This fact clearly suggests that our proposed procedure seems to be more robust and is capable of detecting the genuinely significant clusters, which is not affected by the simulated random data in this application.

Several studies have reported the detection of multiple clusters using scan statistics. In one study proposing an adjusted p-value, a sequential approach is used [29]. Although it performs with higher power than the conventional approach, the relative sizes of adjusted p-value for secondary clusters are irrelevant to the order in which the clusters are sequentially detected; that is, the k-th cluster may have a smaller p-value than the previously detected $(k-1)$-th cluster. Moreover, this procedure can only evaluate the significance of individual clusters but not of multiple clusters as a whole.

In the spatial context, a multiple-cluster detection procedure using spatial scan statistics has also been proposed [33, 34]. However, these methods are also unable to assess the significance of multiple clusters as a whole. As a similar approach to us in terms of adopting a model selection procedure, Zhang et al. [35] have developed a generalized linear mixed model with Moran's I statistic and utilised a stepwise procedure that allows multiple-cluster evaluation, accounting for random spatial effects. They have, however, concluded that the power is lower than that of the standard scan statistic [35]. A recent study [36] proposed a multiple-cluster detection procedure in the spatial context, adopting the quasi-likelihood approach that deals with spatial correlation. One drawback is that quasi-likelihood is not able to produce the likelihood or information criterion. This fact indicates that the quasi-likelihood approach still suffers from the multiple testing problem, whereas our approach avoids the issue by utilising the model selection framework with the proposed information criterion that stands on the full-likelihood principle. We leave it for future work to undertake detail comparisons of these quasi- and full-likelihood approaches in the multiple-cluster detection context.

We have proposed an information criterion procedure to select an appropriate number of clusters for detection. This approach has been used for model selection in more general statistical modelling contexts—for instance, to estimate the number of multiple clusters [37, 38] and finite mixture [39] models, for which traditional AIC and BIC are used. However, in specific situations with large datasets, conventional information criteria including −2log likelihood, AIC, and BIC were unable to accurately select an appropriate number of clusters and performed poorly (Fig 3). The proposed criterion was derived from the marginal likelihood of the multiple-cluster model (3) and accounted for the probability distribution of selected candidate clusters. Our examples and simulations clearly demonstrate that the proposed criteria performed well in terms of identifying "true" multiple-clusters assumed in simulation studies as an appropriate multiple-cluster model.

A more conservative p-value is calculated by the secondary-cluster procedure as compared to the primary cluster procedure [29, 40]. Consequently, the former identifies fewer significant secondary clusters relative to true clusters. However, this was not the case in our simulation study, which showed that the secondary-cluster procedure detected extra clusters. Our example and simulation study demonstrated that the proposed framework performs well, although there are outstanding challenges to be addressed by future work. Firstly, multiple-cluster detection depends on the scanning method that is initially used; we adopted the conventional secondary-cluster procedure to pre-select candidates for a GLM. Therefore, choosing the optimal scan statistic with high detection accuracy is essential. Secondly, our illustration and simulation were carried out only for purely temporal clustering, although the proposed procedure can be applied to purely spatial and spatio-temporal cases with minor modifications. When applied to such cases, choosing the correct probability function, $h(\boldsymbol{z})$, is critical, and a detailed evaluation along with various simulations is required. Thirdly, our proposed framework is based on the likelihood ratio test statistic, but this is not the only option. For example, the statistical nature of Wald-based and other types of scan statistics have also been studied by other researchers [41, 42]. Some comparison study would be useful for understanding the deviation among those different scan statistics.

Here, it is also worth noting about the account for dependence structures in time series data. Although our likelihood approach assumes a (conditional) independent structure among observations, the approach also shares the notion of partial likelihood [43] that can cope with a dependence structure. It is known that under some reasonable conditions the partial likelihood is able to resemble the amount of information the full joint likelihood possesses. See Kedem and Fokianos [44] for the detail explanation and its theoretical elaboration in time series analysis adopting the GLM framework.

In the spatial context particular, it has been pointed out that the original scan statistic tends to produce more false-positives, when the data contain underlying overdispersion and/or spatial correlation, for which some modification has been proposed in the single-cluster detection test framework [45]. Further investigation on the utility of various multiple-cluster detection test statistics accounting for dependence structureswill bring new insights into the future research.

Conclusion

We have proposed a new testing framework for the simultaneous evaluation of purely temporal multiple-clusters, combining GLM and information criterion approaches that directly stand on the likelihood principle. The framework can, thus, treat the cluster selection problem as the model selection problem and provide a single p-value to evaluate the selected multiple clusters as a whole. This feature is unique and beneficial compared to the conventional secondary-cluster procedure that recursively evaluates clusters one by one. Our simulation studies have demonstrated that the new procedure also enables the estimation and evaluation of multiple clusters with high detection power, possessing the consistency property that tends to select the correct number of clusters when the truth is known.

Acknowledgments

The authors are grateful to the editor and the anonymous reviewers for their helpful and constructive comments that have improved the manuscript greatly.

Author Contributions

Conceptualization: Kunihiko Takahashi.

Formal analysis: Kunihiko Takahashi.

Methodology: Kunihiko Takahashi, Hideyasu Shimadzu.

Writing – original draft: Kunihiko Takahashi.

Writing – review & editing: Hideyasu Shimadzu.

References

1. Lawson AB. Statistical methods in spatial epidemiology. 2nd ed. John Wiley & Sons Ltd; 2006.
2. Tango T. Statistical methods for disease clustering. Springer; 2010.
3. Show J. On the mode of communication of cholera. 2nd ed. London: John Churchill; 1855.
4. Lawson AB, Kleinman K.(eds.) Spatial and syndromic surveillance for public health. John Wiley & Sons Ltd; 2005.
5. Waller LA. Discussion: statistical cluster detection, epidemiologic interpretation, and public health policy. Statistics and Public Policy. 2015; 2: 1–8.
6. Besag J, Newell J. The detection of clusters in rare diseases. Journal of the Royal Statistical Society, Series A. 1991; 154: 143–155.

7. Moran PAP. Notes on continuous stochastic phenomena. Biometrika. 1950; 37: 17–23.
8. Whitemore AS, Friend N, Brown BW, Holly EA. A test to detect clusters of disease. Biometrika. 1987; 74: 631–635.
9. Oden N. Adjusting Moran's I for population density. Statistics in Medicine. 1995; 14: 17–26.
10. Tango T. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Statistics in Medicine. 1995; 14: 2323–2334.
11. Rogerson PA. The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic. Geographical Analysis, 1999; 31: 130–147.
12. Bonetti M, Pagano M. The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. Statistics in Medicine. 2005; 24: 753–773.
13. Turnbull B, Iwano E, Burnett W, Howe H, Clark L. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. American Journal of Epidemiology. 1990; 132: suppl. 136–143.
14. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Statistics in Medicine. 1995; 14: 799–810.
15. Kulldorff M. A spatial scan statistic. Communications in Statistic—Theory and Methods. 1997; 26: 1481–1496.
16. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society, Series A. 2001; 164: 61–72.
17. Tango T. A test for spatial disease clustering adjusted for multiple testing. Statistics in Medicine. 2000; 19: 191–204.
18. Tango T. A spatial scan statistic with a restricted likelihood ratio. Japanese Journal of Biometrics. 2008; 29: 75–95.
19. Duczmal L, Assunç~ao R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistics & Data Analysis. 2004; 45: 269–286.
20. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. International Journal of Health Geographics. 2005; 4: 11.
21. Tango T, Takahashi K. A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. Statistics in Medicine. 2011; 31: 4207–4218.
22. 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. 2008; 7: 14.
23. Jung I. A generalized linear models approach to spatial scan statistic for covariate adjustment. Statistics in Medicine. 2009; 28: 1131–1143.
24. Kulldorff M, Information Management Services, Inc. SaTScan v8.0: Software for the spatial and space-time scan statistics. 2009 [cited 10 June 2018]. Available from: http://www.satscan.org/.
25. Takahashi K, Yokoyama T, Tango T. FleXScan v3.1: Software for the Flexible Scan Statistic. 2010 [cited 10 June 2018]. Available from: https://sites.google.com/site/flexscansoftware/.
26. Glaz J, Pozdnyakov V, Wallenstein S. Scan statistics: Methods and applications. Birkhäuser; 2009.
27. Nishiyama T, Takahashi K, Tango T, Pinto D, Scherer SW, Takami S, et al. A scan statistic to extract causal gene clusters from case-control genome-wide rare CNV data. BMC Bioinformatics; 2011. 12: 205.
28. Takahashi K, Shimadzu H. The daily incidence of out-of-hospital cardiac arrest unexpectedly increases around New Year's Day in Japan. Resuscitation. 2015; 96: 156–162.
29. Zhang Z, Assunç~ao R, Kulldorff M. Spatial scan statistics adjusted for multiple clusters., Journal of Probability and Statistics. 2010; Article ID 642379, 11 pages.
30. Konishi S, Kigagawa G. *Information criteria and statistical modeling*. Springer; 2008.
31. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2017 [cited 10 June 2018]. Available from: https://www.R-project.org/.
32. Dwass M. Modified randomization tests for nonparametric hypotheses. The Annals of Mathematical Statistics. 1957; 28: 181–187.
33. Li XZ, Wang JF, Yang WZ, Li ZJ, Lai SJ. A spatial scan statistic for multiple clusters. Mathematical Biosciences. 2011; 233: 135–142.
34. Wan Y, Pei T, Zhou C, Jiang Y, Qu C, Qiao Y. ACOMCD: A multiple cluster detection algorithm based on the spatial scan statistic and ant colony optimization. Computational Statistics and Data Analysis. 2012; 56: 283–296.
35. Zhang T, Lin G. Cluster detection based on spatial associations and iterated residuals in generalized linear mixed models. Biometrics. 2009; 65: 353–360.

36. Lin PS, Kung YH, Clayton M. Spatial Scan Statistics for Detection of Multiple Clusters with Arbitrary Shapes. Biometrics. 2016; 72: 1226–1234.
37. Molinari N, Bonaldi C, Daures JP. Multiple temporal cluster detection. Biometrics. 2001; 57: 277–583.
38. Xie M, Sun Q, Naus J. A latent model to detect multiple clusters of varying sizes. Biometrics. 2009; 65: 1011–1020.
39. McLachlan G, Peel D. Finite mixture models. John Wiley & Sons; 2000.
40. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer clusters in the Northeast United States: a geographic analysis. American Journal of Epidemiology. 1997; 146: 161–170.
41. Zhang T, Lin G. Spatial scan statistics in loglinear models. Computational Statistics and Data Analysis. 2009; 53: 2851–2858.
42. Zhang T, Lin G. Family of power divergence spatial scan statistics. Computational Statistics and Data Analysis. 2014; 75: 162–178.
43. Cox DR. Partial likelihood. Biometrika. 1975; 62: 269–276.
44. Kedem B, Fokianos K. Regression models for time series analysis. Wiley; 2002.
45. Loh JM, Zhu Z. Accounting for spatial correlation in the scan statistic. The Annals of Applied Statistics. 2007; 1: 560–584.