pagepress

Geospatial Health 2017; volume 12:556

CutL: an alternative to Kulldorff's scan statistics for cluster detection with a specified cut-off level

Barbara Więckowska,[1] Justyna Marcinkowska[2]

[1]Department of Computer Science and Statistics, Karol Marcinkowski University of Medical Sciences, Poznan; [2]Department of Computer Science and Statistics, Karol Marcinkowski University of Medical Sciences, Poznan, Poland

Abstract

When searching for epidemiological clusters, an important tool can be to carry out one's own research with the incidence rate from the literature as the reference level. Values exceeding this level may indicate the presence of a cluster in that location. This paper presents a method of searching for clusters that have significantly higher incidence rates than those specified by the investigator. The proposed method uses the classic binomial exact test for one proportion and an algorithm that joins areas with potential clusters while reducing the number of multiple comparisons needed. The sensitivity and specificity are preserved by this new method, while avoiding the Monte Carlo approach and still delivering results comparable to the commonly used Kulldorff's scan statistics and other similar methods of localising clusters. A strong contributing factor afforded by the statistical software that makes this possible is that it allows analysis and presentation of the results cartographically.

Correspondence: Barbara Więckowska, Department of Computer Science and Statistics, Karol Marcinkowski University of Medical Sciences, 79 Dąbrowskiego St, 60-529 Poznan, Poland.
Tel: +48 61 854 68 08 - Fax: +48 61 854 69 43.
E-mail: basia@ump.edu.pl

Key words: Space clusters; Cut-off level; High incidence rate; Spatial pattern detection; Binomial test for one proportion.

Acknowledgments: the authors are grateful to Tomasz Wieckowski (PQStat Software Company, Poznan, Poland) for his valuable assistance in building PQScut software. Research projects supported with statutory funds, number: 502-01-022-32378-03519.

Received for publication: 2 February 2017.
Revision received: 18 May 2017.
Accepted for publication: 12 June 2017.

Geospatial Health 2017; 12:556
doi:10.4081/gh.2017.556

Introduction

The development of statistical methods in geographical analysis has accelerated rapidly with the development of technology in general. In the medical field, spatial cluster detection is an important tool in cancer surveillance, to identify areas of increased risk and to formulate hypotheses about cancer aetiology. A review of the literature draws attention to the particular focus of research in the epidemiology of leukaemia, which has a strong tendency to form clusters, a factor of increasing public attention (Hjalmars *et al*., 1996; Alexander *et al*., 1998; Michelozzi *et al*., 2002; Francis *et al*., 2012). As a result, cancer incidence rates are used in research and comparison of new methods for cluster detection (Turnbull *et al*., 1990; Wheeler, 2007; Huang *et al*., 2008; Lawson and Rotejanaprasert, 2014).

The application of statistical methods depends on initial assumptions about clusters. A method widely used for the detection of clusters employs the scanning window, introduced with Geographical Analysis Machine (GAM) by Openshaw *et al*. (1988), then further developed by Besag and Nowell (1991) and used for Spatial Scan Statistics (Kulldorff, 1997) and Flexible Scan Statistics (Tango and Takahashi, 2005, 2012). In spatial scan statistics, the assumed cluster is defined by showing a greater difference with regard to the observed and expected frequencies inside the window than outside. Over the past several years, development in spatial analysis has resulted in many data smoothing methods, including Bayesian methods (Besag *et al*., 1991; Kang *et al*., 2013). In particular, Bayesian partition model for cluster detection described by Wakefield and Kim (2013) has gained popularity. All these methods are still in progress but the essential aspect of their development and widespread use is accessible software that is developing in tandem with the advancement of these techniques, *e.g*., FleXScan (Takahashi *et al*., 2013), SaTScan (Kulldorff, 2015) and R Cran Packages such as SpatialEpi (Kim *et al*., 2014). The described methods allow the search for clusters without having to take into account the epidemiological expectations of the incidence rate, which the cluster should exceed. In order to search for clusters with an expected incidence rate, we have proposed the CutL method. The idea here is based on smoothing coefficients and searching for clusters with a higher incidence rate than the level of the-cut off level. We report here on simulation studies carried out to demonstrate the effect of this method where it was shown that the sensitivity and specificity of the CutL method are similar to Kulldorff's scan statistic and partly similar to flexible scan statistics and the Bayesian partitin model for cluster detection. We also present the application of this approach on a known dataset of leukaemia cases in New York, USA that were reported by Turnbull *et al*. (1990) and Waller and Gotway (2004).

Materials and Methods

Initial assumptions

The proposed method automatically searches for clusters based on the specified cut-off level value and the level of statistical significance. For example, one can looking for districts that have a incidence rate significantly higher than the specified incidence rate level (Figure 1). Incidence rate (r_i) within district (i) is then defined as the ratio of the number of patients (d_i) to the number of individuals in an exposed population (n_i) within this district:

$$r_i = \frac{d_i}{n_i}. \quad \text{Eq. 1}$$

It would be unrealistic to expect that the incidence rate will be exactly the same as the specified incidence rate level in each district. Particular districts will be characterised by various incidence rate, but the observed variability should be within a certain margin of error. Districts with incidence rates above this range may constitute the beginnings of potential clusters.

Here, a *cluster* is defined as a collection of districts (or a single district) with a significantly higher incidence rate than the specified incidence rate level. Any district that has the potential to build a cluster may independently form a cluster or, if it occurs in the vicinity of districts with high incidence rate, has the opportunity to join them and form a larger cluster. These designated potential clusters are tested for statistical significance in two separate steps: Step 1 - Locating the potential cluster(s) of greater incidence rate(s) than those of the reference; Step 2 - Testing the statistical significance of the potential cluster(s). The results of these two steps (*i.e.* localised clusters) are automatically displayed onto a map.

Step 1

In the proposed CutL method, potential clusters are located based on a specified incidence rate (cut-off level) above which the investigator expects the emergence of a cluster. This requires the investigator to define the neighbouring districts. In general, the Queen matrix is the standard used for defining contiguity of borders between districts (areas sharing any boundary point are taken as neighbors; Lloyd, 2010).

Cut-off level

Cut-off can be calculated automatically as the overall incidence rate or given by the investigator. The choice of cut-off level (X_{CutL}) has a significant impact on the interpretation of clusters. If the researcher is interested in showing the location of unusual clusters – only within the area under study – the proposed cut-off level will be calculated as a ratio of all patients to the entire population of that area. This is also known as the overall incidence rate and calculated as follows:

$$X_{CutL} = X_{\text{overall}} = \frac{\sum_{i=1}^{m} d_i}{\sum_{i=1}^{m} n_i} \quad \text{Eq. 2}$$

where m is the number of districts.

If the investigator is interested in identifying clusters compared to the specified incidence rate of wider areas or of different areas than those under study (*e.g.*, those referenced in other studies), then the proposed cut-off level should be the incidence rate of the wider/different area. Under specific situations, for example, when the whole area under study or a substantial portion makes up the cluster, a reference analysis using the externally specified incidence rate of the wider/different area is one way of locating such clusters.

Incidence rate smoothing

The method of data aggregation is most commonly associated with the administrative division of the investigated region, where subdivisions are the districts with varying populations, for example, cities (numerous), predominately urban (less numerous) and predominately rural (sparse). The stability of the incidence rate mainly depends on the number of people exposed. Districts with small populations are naturally characterised by high variability in the incidence rate, which has a tendency to include outliers. Because the task of the proposed method is to search for clusters with incidence rate values higher than the cut-off level (including outlier districts), incidence rate smoothing is used during the cluster detection stage. In the present study, incidence rate was smoothed using the Empirical Local Bayes Smoothing method (Waller and Gotway, 2004).

During smoothing, the incidence rate is determined based on the number of patients and individuals within the population of a given district and in its neighbouring districts (according to neighbourhood matrix). The potential value for a given district, *i.e.* the diagonal element of a neighbourhood matrix, is set as the sum of elements outside the diagonal of particular district. As a result, the given, smoothed district has the same influence on the result of smoothing as its adjacent districts altogether:

$$smooth(r_i) = \frac{smooth(d_i)}{smooth(n_i)} + C_i\left(\frac{d_i}{n_i} - \frac{smooth(d_i)}{smooth(n_i)}\right) \quad \text{Eq. 3}$$

where *smooth* (r_i) is the smoothed incidence rate within district (i); C_i the shrink factor:

$$smooth(d_i) = \frac{\sum_{i=1}^{m} w_{ij} d_i}{\sum_{i=1}^{m} w_{ij}} \text{ and } smooth(n_i) = \frac{\sum_{i=1}^{m} w_{ij} n_i}{\sum_{i=1}^{m} w_{ij}} \quad \text{Eq. 4}$$

where w_{ij} is the value of the neighbour matrix element of i and j and districts (that equals one when districts are neighbouring districts and zero when they are not); and $w_{ii} = \sum_{j=1, j\neq i}^{n} w_{ij}$.

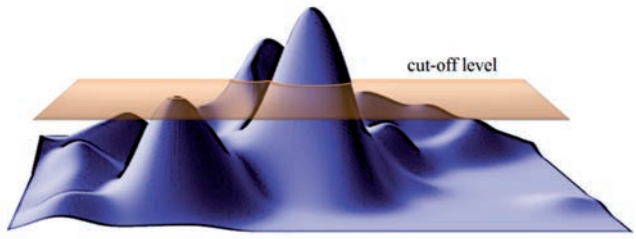

Figure 1. The idea of searching for clusters at a preset cut-off level.

Building of clusters

Localisation of anchoring points

The location of anchoring points and the determination of cluster size is automatic. Since the purpose of the analysis is localisation of cluster(s) with significantly higher incidence rate(s) than the cut-off level(s), anchoring points are districts that in themselves (alone) may constitute a cluster. The idea behind this procedure was taken from the method described by Choynowski (1959), which presents the probability on a map. For each district, a margin of error was built around the smoothed incidence rate. This margin was built based on the smoothed population size of a given district *smooth* (n_i) and the corresponding number of patients in the district calculated as *smooth* (n_i) • *smooth* (r_i). The location of *smooth* (r_i) along with the entire margin of error above the specified cut-off level X_{CutL} qualifies this district as an anchoring point in the building of a cluster. A 95% Clopper-Pearson (1934) confidence interval was chosen for the margin of error.

Cluster size determination

After identifying the anchoring points, further analysis moves from the smoothed incidence rate towards the actual incidence rate (unsmoothed) in order to determine the size of clusters. They can take place in one of three situations (Figure 2): i) the designated anchoring districts independently constitute a cluster – if there are no districts with high incidence rates in their surroundings; ii) the anchoring districts may be joined with neighbouring districts creating an aggregated cluster – if the neighbouring districts are characterised by a high incidence rate (a description of the algorithm is presented in the next section: (Increasing cluster); iii) the clusters built according to points i) and/or ii) above are combined into one aggregated area to form a larger cluster – if there are common districts.

Increasing cluster

As a result of finding anchoring points by use of the smoothed incidence rate, there may be neighbouring districts with high incidence rates that are not anchor points. This may occur especially when these districts are found on the border of potential clusters. In order to allow the joining of such districts, for each district (j) neighbouring with an anchoring district (i), the distance of the incidence rate from the cut-off level is calculated. The obtained difference is multiplied by the weight, which is the square root of the population size of the given district. In this way, the resulting difference is increased for districts with larger population sizes. This factor is given by:

$$c_j = \left(\frac{d_j}{n_j} - X_{CutL}\right)\sqrt{n_j} \qquad \text{Eq. 5}$$

Next, to the anchoring districts, successively, its neighbours starting with those neighbours whose value c_j is greatest, are joined. In aggregated areas, the coefficient:

$$c_i = \left(\frac{d_i + \sum d_j}{n_i + \sum n_j} - X_{CutL}\right)\sqrt{n_i + \sum n_j} \qquad \text{Eq. 6}$$

is re-set. The next neighbours are joined until their addition causes an increase in the coefficient c_i of a built cluster.

Step 2

Clusters were identified in the first step and now in the second step the statistical significance of these clusters will be determined. This is done using the binomial exact test for one proportion. This test compares the real (unknown) incidence rate inside the cluster ($R_{cluster}$) to the cut-off level (X_{CutL}):

$$H_O: R_{cluster} = X_{CutL} \qquad \text{Eq. 7}$$

based on the known incidence rate in cluster:

$$r_{cluster} = \frac{d_{cluster}}{n_{cluster}} \qquad \text{Eq. 8}$$

where $d_{cluster}$ is the sum of patients within the cluster; and $n_{cluster}$ the population size within the cluster.

A one-tailed hypothesis test is then used due to the fact that a statistical verification of clusters with higher incidence rate than the given cut-off level is performed. The problem of multiple comparisons is solved using the criterion of false discovery rate (FDR) discussed by Benjamini and Hochberg (1995). The criterion is considered to be more effective in the detection of spatial clusters than the family-wise error rates (Caldas de Castro and Singer, 2006; Catelan and Biggeri, 2010). The Benjamin-Hochberg (1995) correction was used for a relatively small number of districts – only those indicated as potential clusters. In accordance with this method, clusters are sorted by descending P values. Next corrections are sequentially applied to decreasing number of the remaining hypotheses.

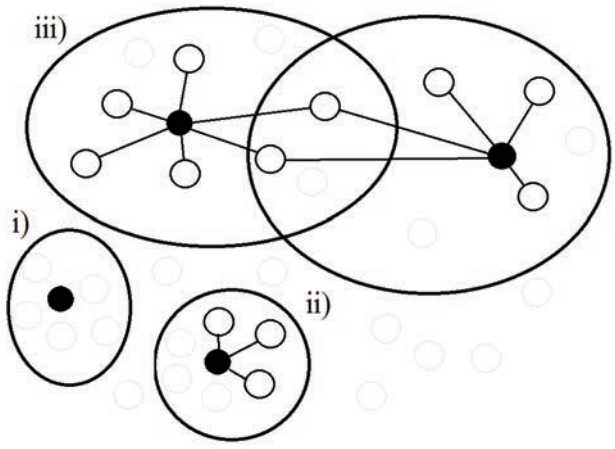

Figure 2. Possible situations encountered when determining the size of clusters around the anchoring district using the CutL method. i) The anchoring district independently constitutes a cluster; ii) the anchoring district joined with neighbouring districts creating an aggregated cluster; iii) the clusters combined into one aggregated area to form a larger cluster.

Pseudo-code

For readability purposes of the Methods section, we present the following procedure:

```
//Getting data
DataMap.pt = GetDataFromMap

//Creating neighbourhood matrices for Spatial Polygons (Queen matrix - calculated automatically, or matrix given by investigator)
DataMap.mx = GenMatrix(DataMap.pt)

//Cut-off level (automatically calculated from data or given by the investigator)
CutL = SumCases / SumPopulation

//Calculating smoothed incidence rates (Empirical Local Bayes Smoothing method)
DataSmooth = LocalEmpiricalBaysSmooth(DataSheet, DataMap.mx);

//Localisation of anchoring points (smoothed incidence rate of a district along with the entire margin of error (95%CI) above the specified cut-off level qualifies the district as an anchoring point)
for i=0 to Length(DataSmooth[0])-1
if (ClopperPearson_CI_Low(DataSmooth[i][1] / DataSmooth[i][0]) > CutL)
GenAnchoringPoints(i)

//Cluster size determination (based on ci coefficient)
for i=0 to Length(DataMap.mx)-1
for k=0 to Length(DataMap.mx[i])-1
MainMatrix[i][k]:=( (case / pop) - CutL) * Sqrt(pop)
 max := SearchMaxValueMainMatrix(MainMatrix)
 sumMatrix := SearchNeighboursAboveCutL(MainMatrix, CutL, max)
CLUST_ID := BuildSubClusters(sumMatrix)
CLUST_ID := SearchAndExtendConnectedSubClusters(CLUST_ID)

//Statistical significance analysis for clusters
test_Binom = ComputeBinomExTest(CLUST_ID)
P value_multi = ComputeMultiCompareTest(test_Binom,BenjaminHochberg)
```

Simulation studies

The project

As the basis for our simulations, the population of Wielkopolskie Province in 2013 (3,467,016) was used, which was provided by the Central Statistical Office of Poland, Local Data Bank). The province is divided into 315 municipalities, which constitute the smallest units of administrative division. This division in is the basis of the regional planning using Geographic Information System (GIS) and the perceived need for medical care. The municipalities vary greatly by the number of inhabitants. The largest municipality by population (provincial capital) had 548,028 residents in 2013, while the least numerous one had 1,454 residents at this time. The median and quartiles were respectively: 6,298 (4,462; 9,621) residents, the number of patients (d) was set at 3,467. The cut-off level beyond which we searched for clusters, was therefore the overall incidence rate: d/N=0.001.

To show accuracy of the presented methods, three different spatial distributions were tested: (1) the null hypothesis of no clustering, *i.e.* the data distributed randomly in accordance with multinomial distribution; (2a) two separate clusters, one round in the northern part of the province (5% of the population), a second elongated located along the western border of the province (1% of the population); and (2b) a cluster following the course of the river through the capital of the province (28% of the population). Districts that belong to the defined clusters (2a) and (2b) received the status of a *true cluster*.

We simulated regional count data sets ($d_1, d_2, \ldots, d_m$) based on multinomial distribution:

$$(d_1, d_2, \ldots, d_m) \sim multinomial\left(RR_1 \frac{d_1}{n_1}, RR_2 \frac{d_2}{n_2}, \ldots, RR_m \frac{d_m}{n_m}\right) \quad \text{Eq. 9}$$

where

$$d = \sum_{i=1}^{m} d_i = 3{,}467.$$

(1) In the absence of a cluster, under the null hypothesis of constant relative risk, RR_i is the relative risk at geographic unit i, that is set to 1.

(2) In the presence of a cluster, under the alternative hypothesis, RR_i is the relative risk at geographic unit i, which is higher for units that belong to defined clusters (set as *true clusters*).

The procedure was repeated by drawing 200 times for the random data (1) and 200 times for each value of the relative risk with a spatial pattern forming clusters (2), and thus for RR=1.5, RR=2.5 and RR=4. These simulation data can be downloaded from the website http://pqscut.ump.edu.pl.

Since searching for clusters aims to identify both the location and size of clusters and investigate the statistical significance of selected locations, usually checking the quality of the analysis takes into account both of these aspects. Clusters presented in this study range from 6% of the population (197,712 individuals) - in the case of the first two clusters - and up to 28% (959,045 individuals) for clusters located along the river. The power of the CutL analysis and Kulldorff's scan statistic is over 99.9%, when RR>1.5 and type I error=0.05. Thus, both analyses are powerful enough to detect statistically significant clusters. The main focus of this study is the aspect of precision in detected location(s) and cluster size(s).

If we denote the number of municipalities that are truly clusters (*i.e.* correctly detected as clusters) as *TP*; the municipalities that are not clusters (*i.e.* incorrectly detected as clusters) as *FP*; the municipalities that are not clusters (and not identified as such) as *TN*; the municipalities that are truly clusters (and not identified as such) as *FN*; the compatibility of the location and size of clusters detected with actual clusters can be examined by using five measures: Sensitivity - the proportion of municipalities identified as clusters to those that are true clusters: $TP/(TP+FN)$; Specificity - the proportion of municipalities identified as non-clusters among those that are non-clusters: $TN/(TN+FP)$; Positive Predictive Value - the proportion of municipalities that are true clusters among all those identified as clusters: $TP/(TP+FP)$; Negative Predictive Value - the proportion of municipalities that are non-clusters among all those identified non-clusters: $TN/(TN+FN)$; and Accuracy - the proportion of correctly classified municipalities: $(TP+TN)/(TP+TN+FP+FN)$.

This study compared the results obtained with the CutL method to the most popular method, *i.e.* Kulldorff's scan statistic.

The flexible scan statistic and the Bayesian partition model for cluster detection are both more flexible than Kulldorff's scan statistic and allows for searches of any shape. However, this comparison was only done for RR=2.5 because the flexible scan statistic algorithm and the Bayesian partition model for cluster detection are slow and it is time-consuming to carry out repeated analysis for simulation data. During the analysis, the default settings of these methods were not changed.

For statistical analyses the significance level α=0.05 was assumed. The PQScut program was used for the CutL method and for plotting data on the map. Kulldorff's scan statistic was calculated in SatScan, the flexible scan statistic by the FlexScan program and the Bayesian partition model for cluster detection in R (SpatialEpi package).

For the simulation study, sensitivity, specificity, positive predictive value, negative predictive value and accuracy were all calculated to describe the precision of detected clusters 2a and 2b (compliance with the municipalities set as *true clusters*). Specificity was calculated for the null hypothesis (1) representing the absence of clustering.

Results

The results of the simulations for all tested levels of relative risk of random data (1), with two clusters (2a), and clusters located along the river (2b) are shown in Table 1, For RR=2.5, results are also illustrated on a map (Figures 3 and 4). For RR=1 (*i.e.* absence of clusters) both methods obtained very high results of designated measures where they reached 99%.

Both CutL and Kulldorff's scan statistic yielded satisfactory results in the analysis for simultaneously detecting two regions forming clusters (2a). All the designated measures remained at a high level from RR=2.5. However, compared to the Kulldorff's scan statistics, CutL characterised higher accuracy at each level of RR. The individual measures (sensitivity, specificity, positive predictive value, and negative predictive value) were also higher with the exception of RR=1.5, at which the Kulldorff's scan statistic yielded a higher sensitivity and a higher negative predictive value. Compared to the Bayesian partition model for cluster detection, CutL reached a higher value for all of the designated measurements at RR=2.5. In comparison to the flexible scan statistic, the

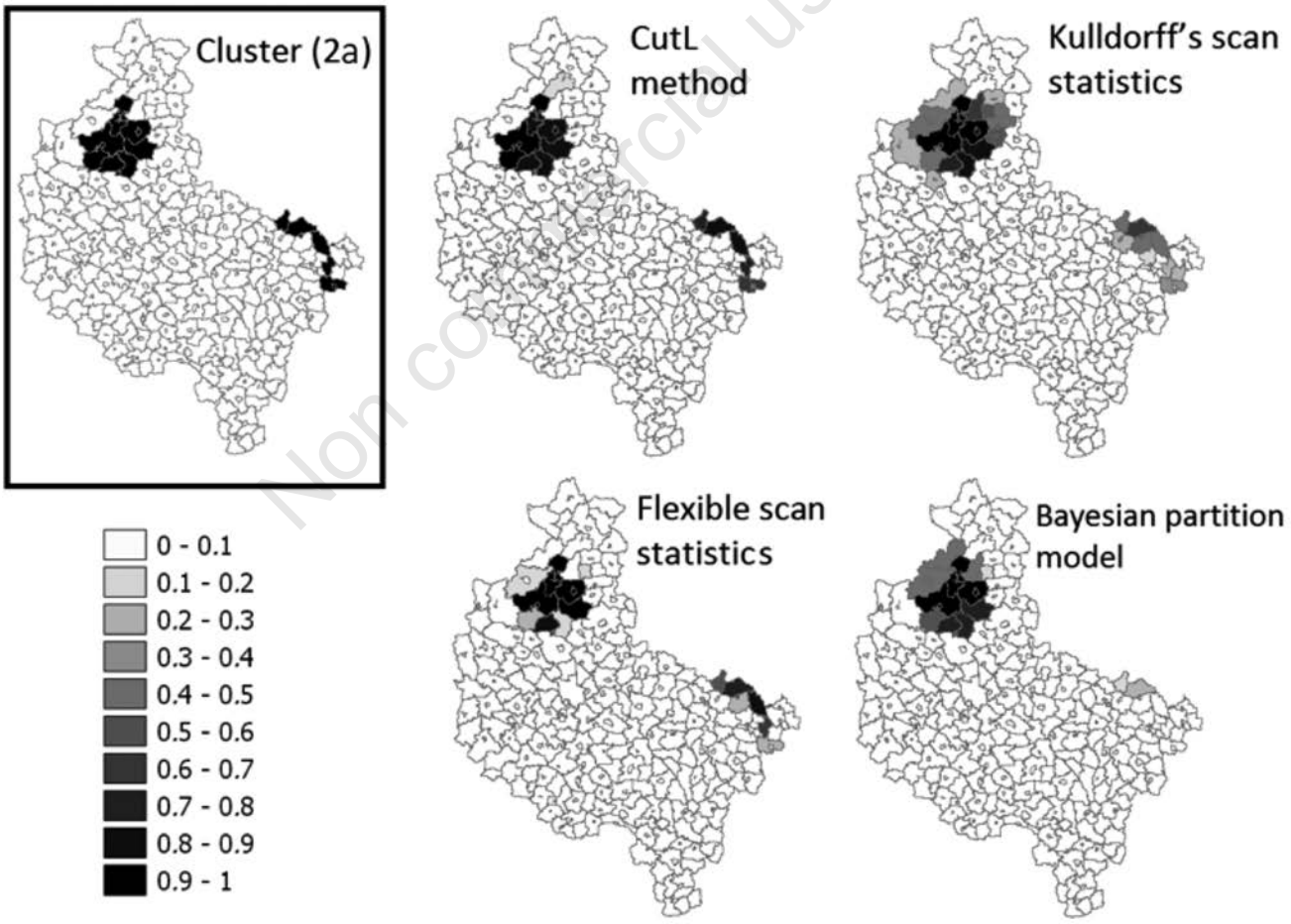

Figure 3. Simulation results comparing CutL, Kulldorff's scan statistics, flexible scan statistics and Bayesian partition model for cluster detection based on 200 replications when applied for two separate clusters. The percentage of municipalities classified as *true clusters* when the relative risk is 2.5 times higher with respect to clustering for the municipalities (2a); CutL cut-off level=overall incidence rate=0.001.

CutL method obtained higher values for sensitivity and negative predictive value.

Weaker results were obtained for clusters located along the river (2b). All the indicated measurements increased with increasing levels of RR. Compared to the Kulldorff's scan statistic, CutL revealed greater values for all the indicated measurements at each RR level. Measures designated for the flexible scan statistic concerned only those at RR=2.5 where they revealed slightly better results than the CutL method and much better results than both Kulldorff's scan statistic and the Bayesian partition model for cluster detection.

Cluster detection for standard data files (Turnbull *et al*., 1990) of leukaemia cases in New York from 1978-1982, was carried out at the overall incidence rate level. As described by Waller and Gotway (2004), these cases were georeferenced at the level of census block groups, but some of the cases could not be georeferenced at this resolution. These cases were originally allocated proportionally among the block groups, so that some of the resulting disease counts were not necessarily integers. That was a problem for scan statistics, therefore disease counts were rounded to the closest integer and all analysis performed on rounded data. The data included 574 leukaemia cases among 1,057,673 people at risk. The map of smoothed incidence rates and clusters location (detected by various methods) is presented in Figure 5.

Discussion

The presented CutL method serves to detect clusters with significantly higher incidence rates than the cut-off incidence rate specified by the investigator. This approach provides the researcher with a unique level of control in defining the cut-off level in a given analysis. For example, it is no longer necessary to compare the incidence rates within and outside clusters, but instead compare the incidence rate within a cluster at a specified incidence rate level. Comparisons of incidence rates within and outside clusters lead to difficulties in interpreting the results due to the lack of knowledge about the area with which the potential cluster is compared. Of concern is whether that area is free from any threat and therefore does not contain any cluster. A distinct advantage of the CutL method is that is offers the possibility to compare results obtained using the same cut-off level for different populations and geographical areas, which in turn facilitates comparison of results from various studies. Furthermore, the possibility to define the cut-off level by the investigator allows searches in areas where the frequency of an event is not alarmingly high, but higher than expected. For example, an area under study may be characterised by a high incidence rate of a certain illness, but does necessarily contain significant clusters at a specified level, then all you have to do is to decrease cut-off level. In this way, identification of areas in need of improved prophylactic measures to achieve the desired effect can be identified.

Another advantage of CutL is that this method permits the use of classic statistical methods. The binomial exact test for one proportion, used in this method, determines the p-value in an analytical way. In contrast to the Monte Carlo approach, the problems seen in multiple sampling techniques, disappears when using the CutL method. The selection of the type of generator, which is critical for each programme based on simulations (Gentle, 2003), represents one such problem. Another is the prolonged duration of calculations when many potential clusters are present. For example, one of our analyses using the Bayesian partition model for cluster detection took 66 minutes on a 2GHz Intel Core i7 processor with 8GB of 1333MHz DDR3 RAM. The problem of multiple comparisons encountered by all the cluster detection methods has been solved by use of one of the standard procedures: Benjamin-Hochberg (1995) correction. This correction is possible due to the detection of only a small number of potential clusters and the *p*-value is determined using a classical approach.

In contrast to other methods, such as those introduced by Turnbull *et al*. (1990), Besag and Newell (1991), Kulldorff (1997), Tango and Takahashi (2005), CutL does not require defining additional technical parameters, *e.g*., expected shape of a cluster or the maximum size of the scanning window. Appropriate selection of these parameters requires prior knowledge of the shape or size of clusters, which are typically not known. This problem does not appear with CutL, where the detected clusters can be of the any shape and size and therefore can be unknown before the analysis is performed. The most important parameter in the CutL methodology is the cut-off level, which can be defined by the investigator.

Applying CutL and Kulldorff's scan statistic on the same dataset provides a comparison of their levels of accuracy. In locations with round or oval clusters, as well as the unusual shape along the course of the river, CutL yielded the most accurate results (Table 1). This has to do with the fact that the results obtained from a known dataset by the application different methods cannot be exactly the same. In the current case, CutL and FlexScan both indicate the existence of a statistically significant cluster in the centre, but it covers 1 district with CutL *vis-à-vis* 6 districts with FlexScan. The Kulldorff scan statistic found the significant cluster in the South, but Bayesian partition model did not locate any statistically significant cluster. These results were reached by the default settings of each analysis, *i.e*. the change of output settings allows the detection of more clusters in the scan methods (Wakefield and Kim, 2013). In the CutL method, on the other hand, the location of a larger number of clusters can be achieved by lowering the cut-off level.

Due to the nature of cluster detection when based on data smoothing and combining neighbouring clusters, information from the neighbourhood matrix is particularly important. This matrix may be based on contiguity of borders such as the Queen matrix. It is also possible to use a binary matrix based on distances, for example Euclidean distance, where neighbours are objects within a predetermined radius. The default matrix in CutL analysis is the Queen matrix, but this poses a difficulty because it may take a relatively long time to build this matrix if unusual great detail can be required to describe borders of neighbouring districts. In this case, we recommend building a matrix prior to analysis and chose that matrix during the CutL analysis. Furthermore, the way in which the neighbourhood is defined influences the results of each analysis, including CutL, and Kulldorff's scan statistics, which warrants performing future studies on the influence of the matrix type on the accuracy and power of CutL analysis in locating clusters of any shape.

Besides a number of advantages offered by CutL, there is a clear disadvantage in what regards the sensitivity for the level of data aggregation. This is a familiar problem that affects many methods as is evident from its discussion by many authors (Fotheringham and Wong, 1991; Amrhein, 1995; Openshaw and Alvanides, 2005; Ozonoff *et al*., 2007; Lemke *et al*., 2013; Luo, 2013; Jeffery *et al*., 2014). The level of data aggregation is especially important when comparing studies from different areas.

Table 1. Comparison among CutL Kulldorff's scan statistics, flexible scan statistics and Bayesian partition model for cluster detection based on 200 replications.

Spatial distributions		Methods	Sensitivity	Specificity	PPV	NPV	Accuracy
(1) The null hypothesis of no clustering	RR°=1	CutL method	-	0.990	-	-	-
		Kulldorff's scan statistics	-	0.997	-	-	-
(2a) Two clusters	RR°=1.5	CutL method	0.234	0.999	0.631	0.955	0.956
		Kulldorff's scan statistics	0.473	0.970	0.491	0.968	0.942
	RR°=2.5	CutL	0.838	0.995	0.890	0.990	0.985
		Kulldorff's scan statistics	0.714	0.968	0.579	0.982	0.954
		Flexible scan statistics	0.710	0.995	0.892	0.983	0.979
		Bayesian partition model	0.679	0.994	0.854	0.985	0.981
	RR°=4	CutL method	0.986	0.996	0.939	0.999	0.996
		Kulldorff's scan statistics	0.698	0.965	0.545	0.981	0.949
(2b) Clusters located along the river	RR°=1.5	CutL method	0.195	0.991	0.764	0.894	0.890
		Kulldorff's scan statistics	0.185	0.972	0.492	0.891	0.872
	RR°=2.5	CutL method	0.522	0.995	0.944	0.935	0.935
		Kulldorff's scan statistics	0.335	0.973	0.647	0.910	0.892
		Flexible Scan Statistics	0.587	0.997	0.961	0.943	0.945
		Bayesian partition model	0.116	0.998	0.886	0.886	0.886
	RR°=4	CutL method	0.744	0.999	0.986	0.964	0.966
		Kulldorff's scan statistics	0.431	0.969	0.666	0.921	0.900

RR, relative risk; PPV, positive predictive value; NPV, negative predictive value. °The degree of RR for the presence of *true cluster(s)* set at 1 or 1.5, 2.5, 4 times higher for all municipalities, respectively. The cut-off level=overall incidence rate=0.001.

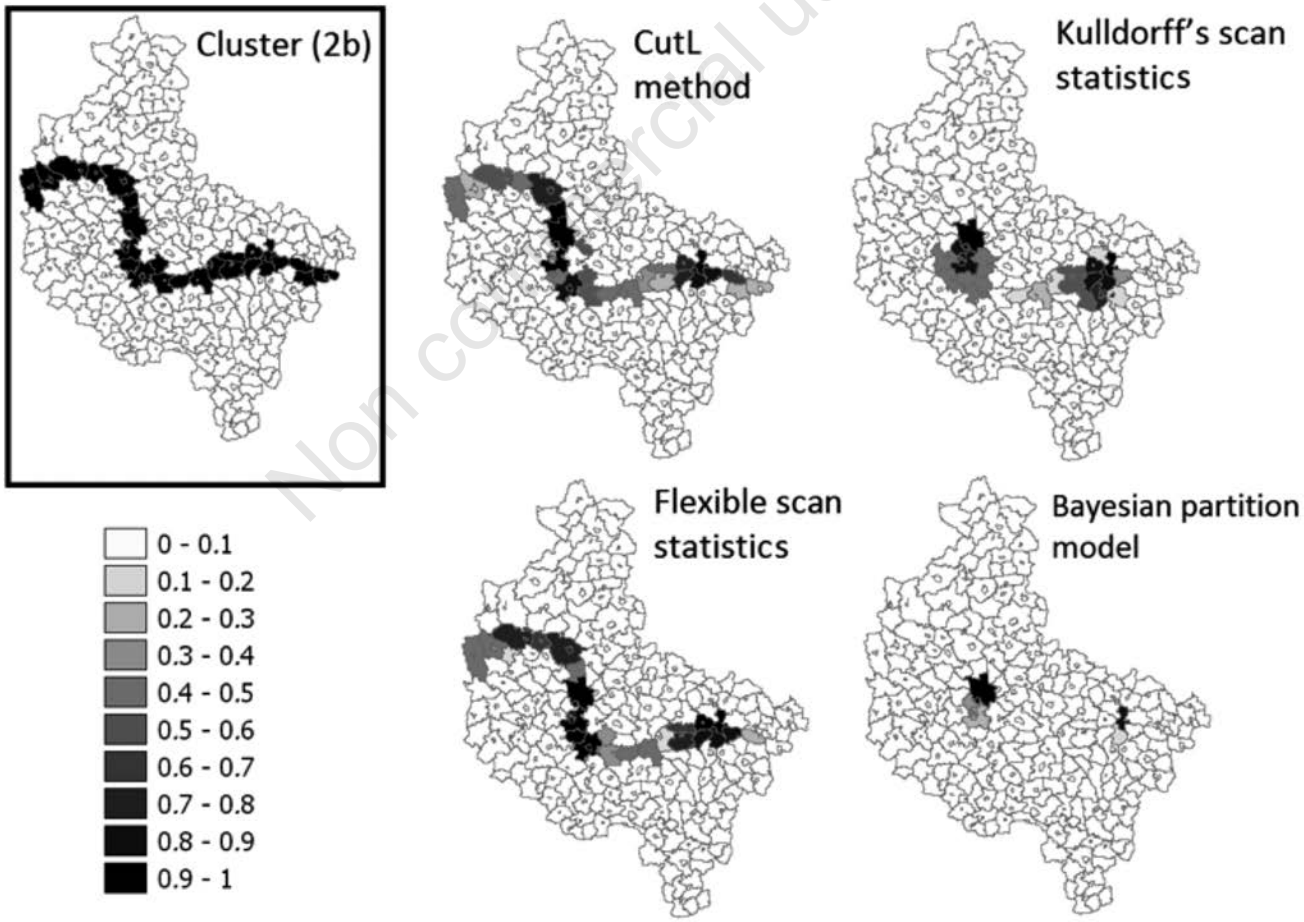

Figure 4. Simulation results comparing CutL, Kulldorff's scan statistics, flexible scan statistics and Bayesian partition model for cluster detection based on 200 replications when applied for a cluster following the course of the river through the provincial capital. The percentage of municipalities classified as *true clusters* when the relative risk is 2.5 times higher with respect to clustering for the municipalities (2b); CutL cut-off level=overall incidence rate=0.001.

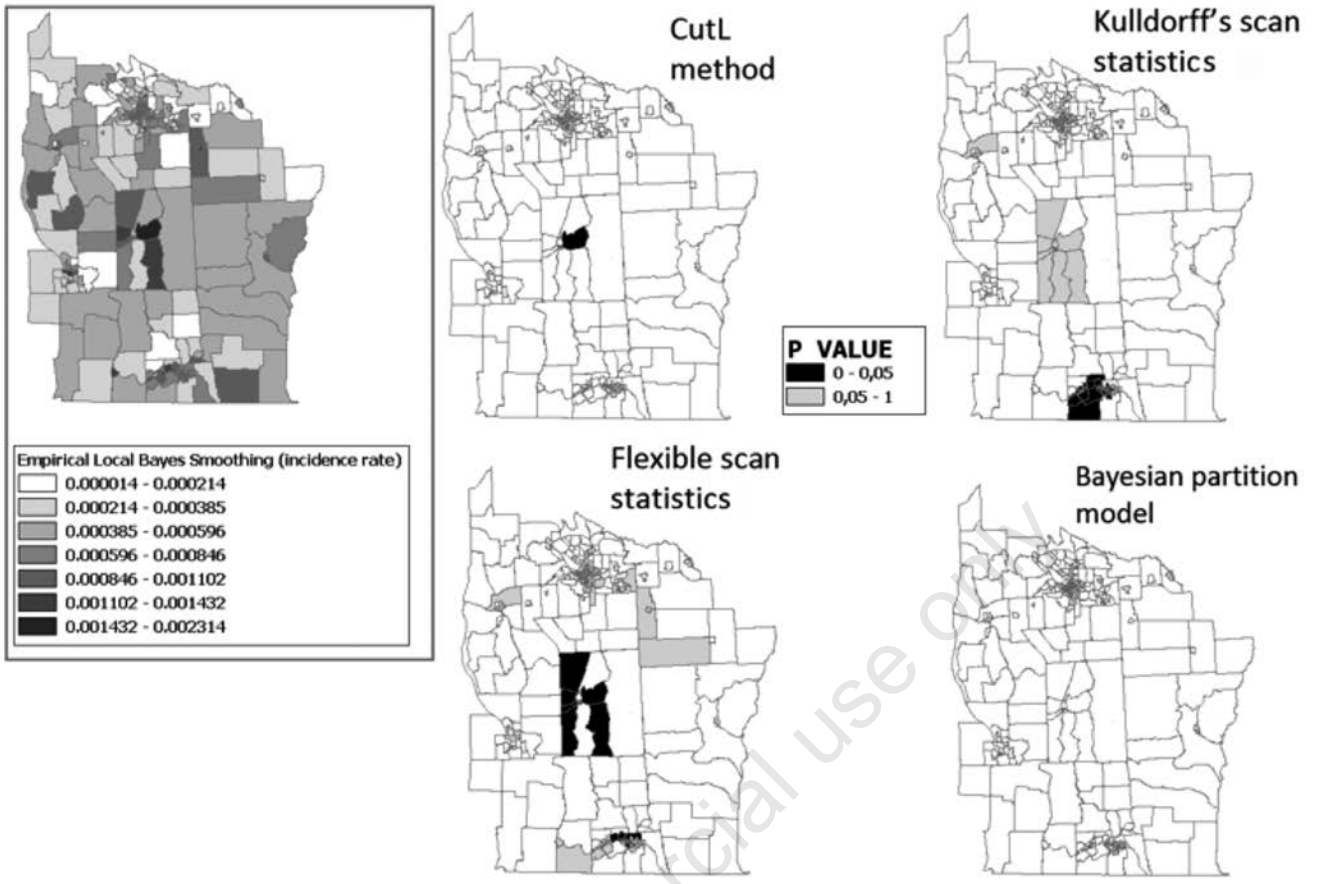

Figure 5. Comparison of CutL, Kulldorff's scan statistics, flexible scan statistics and Bayesian partition model for cluster detection when applied for leukaemia incidence based on a well-known dataset from New York, NY, USA. Leukaemia dataset from Turnbull *et al.* (1990) described in Waller and Gotway (2004); CutL cut-off level=overall incidence rate=0.0004.

However, for methods that aim to reference results from studies of different areas this is especially important. So, if one would like to compare results of cluster detection, it is important that the degree of aggregation in the compared areas be similar. Another problem of CutL is that there is currently no possibility to add additional dimensions or confounding variables (*e.g.*, gender or age). However, because of the nature of the proposed method, it should be possible to develop this method and implement standardisation in future studies.

Conclusions

A new method, CutL, for analysing clusters characterised by significantly higher incidence rates than those specified by the investigator is presented. Without resorting to the Monte Carlo approach, sensitivity and specificity are preserved. A strong contributing factor afforded by the statistical software that allows analysis and presentation of the results cartographically. CutL has been implemented in PQScut free statistical software that can be downloaded from the http://pqscut.ump.edu.pl website, and PQStat software that is available at www.pqstat.com.

References

Alexander FE, Boyle P, Carli PM, Coebergh JW, Draper GJ, Ekbom A, Levi F, Mckinney PA, Mcwhirter W, Michaelis J, Peris-Bonet R, Petridou E, Pompe-Kirn V, Plisko I, Pukkala E, Rahu M, Storm H, Terracini B, Vatten L, Wray N, 1998. Spatial clustering of childhood leukaemia: summary results from the EUROCLUS project. Brit J Cancer 77:818-24.

Amrhein CG, 1995. Searching for the elusive aggregation effect. Evidence from statistical simulations. Environ Plann A 27:105-19.

Benjamini Y, Hochberg Y, 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B-Method 57:289-300.

Besag J, Newell J, 1991. The detection of clusters in rare diseases. J Roy Stat Soc A-Stat Soc 154:143-55.

Besag J, York J, Mollié A, 1991. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43:1-20.

Caldas De Castro M, Singer BH, 2006. Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. Geogr Anal 38:180-208.

Catelan D, Biggeri A, 2010. Multiple testing in disease mapping

and descriptive epidemiology. Geospat Health 4:219-29.
Choynowski M, 1959. Maps based on probabilities. J Am Stat Assoc 54:385-8.
Clopper CJ, Pearson ES, 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26:404-13.
Fotheringham AS, Wong DWS, 1991. The modifiable areal unit problem in multivariate statistical analysis. Environ Plann A 23:1025-44.
Francis SS, Selvin S, Yang W, Buffler PA, Wiemels JL, 2012. Unusual space-time patterning of the Fallon, Nevada leukemia cluster: Evidence of an infectious etiology. Chem Biol Interact 196:102-9.
Gentle JE, 2003. Random number generation and Monte Carlo methods, statistics and computing. Springer, Amsterdam, The Netherlands.
Hjalmars U, Kulldorff M, Gustafsson G, Nagarwalla N, 1996. Childhood leukaemia in Sweden: using GIS and a spatial scan statistic for cluster detection. Stat Med 15:707-15.
Huang L, Pickle LW, Das B, 2008. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. Stat Med 27:5111-42.
Jeffery C, Ozonoff A, Pagano M, 2014. The effect of spatial aggregation on performance when mapping a risk of disease. Int J Health Geogr 13:9.
Kang SY, Mcgree J, Mengersen K, 2013. The impact of spatial scales and spatial smoothing on the outcome of bayesian spatial model. PLoS One 8:e75957.
Kim AY, Chen C, Ross M, Wakefield J, 2014. Methods and data for spatial epidemiology. Available from: http://CRAN.R-project.org/package=SpatialEpi
Kulldorff M, 1997. A spatial scan statistic. Commun Stat Theory Method 26:1481-96.
Kulldorff M, 2015. Information Management Services Inc. SaTScan v9.4.1: Software for the spatial and space-time scan statistics. Available from: http://www.satscan.org/
Lawson AB, Rotejanaprasert C, 2014. Childhood brain cancer in Florida: A Bayesian clustering approach. Stat Publ Pol 1:99-107.
Lemke D, Mattauch V, Heidinger O, Pebesma E, Hense HW, 2013. Detecting cancer clusters in a regional population with local cluster tests and Bayesian smoothing methods: a simulation study. Int J Health Geogr 12:18.
Lloyd, CD, 2010. Chapter 3 in 'Local Models for Spatial Analysis', 2nd Edition. CRC Press, Boca Raton, FL, USA.
Luo L, 2013. Impact of spatial aggregation error on the spatial scan analysis: a case study of colorectal cancer. Geospat Health 8:22-35.
Michelozzi P, Capon A, Kirchmayer U, Forastiere F, Biggeri A, Barca A, Perucci CA, 2002. Adult and childhood leukemia near a high-power radio station in Rome, Italy. Am J Epidemiol 155:1096-103.
Openshaw S, Alvanides S, 2005. Applying geocomputation to the analysis of spatial distributions. In: Longley PA, Goodchild MF, Maguire DJ, Rhind DW (eds.) Geographic information systems: principles and technical issues. Abridged ed. New York: John Wiley and Son, 267-282 pp.
Openshaw S, Craft AW, Charlton M, Birch JM, 1988. Investigation of leukaemia clusters by use of a geographical analysis machine. Lancet 1:272-3.
Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M, 2007. Effect of spatial resolution on cluster detection: a simulation study. Int J Health Geogr 6:7.
Takahashi K, Yokoyama T, Tango T, 2013. FleXScan v312: Software for the flexible scan statistic. National Institute of Public Health. Available from: http://www.niph.go.jp/ soshiki/gijutsu/index_e.html
Tango T, Takahashi K, 2005. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 4:11.
Tango T, Takahashi K, 2012. A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. Stat Med 31:4207-18.
Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC, 1990. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. Am J Epidemiol 132:S136-43.
Wakefield J, Kim A, 2013. A Bayesian model for cluster detection. Biostatistics 14:752-65.
Waller LA, Gotway CA, 2004. Linking spatial exposure data to health events. In: Waller LA, Gotway CA (eds.) Applied spatial statistics for public health data. John Wiley and Sons, New York, NY, USA, pp. 325-443.
Wheeler DC, 2007. A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003. Int J Health Geogr 6:13.