Searching for space-time clusters: The CutL method compared to Kulldorff's scan statistic

Barbara Więckowska,[1] Ilona Górna,[2] Maciej Trojanowski,[3] Agata Pruciak,[1] Barbara Stawińska-Witoszyńska[4]

[1]Department of Computer Science and Statistics, Poznan University of Medical Sciences, Poznan; [2]Department of Bromatology, Poznan University of Medical Sciences, Poznan; [3]Greater Poland Cancer Centre, Greater Poland Cancer Registry, Poznan; [4]Department of Epidemiology and Hygiene, Poznan University of Medical Sciences, Poznan, Poland

Abstract

Both epidemiology and health care planning require analytical tools, especially for cluster detection in cases with unusually high rates of disease incidence. The aim of this work was to extend the application of the CutL method, which is used for detecting spatial clusters of any shape, to detecting space-time clusters, and to show how the method works compared to Kulldorff's scan statistic. In the CutL method, clusters with disease incidence rates higher than the one entered by the researcher are searched for. The way in which the space-time version of that method works is illustrated with the example of data simulating the distribution of people affected by health problems in Polish counties in the period 2013-2017. With respect to detection of irregularly shaped space-time clusters, the CutL method turned out to be more effective than Kulldorff's scan statistic; for cylinder-shaped space-time clusters, the two methods produced similar results. The CutL method has also the important advantage of being widely accessible through the PQScut and PQStat programmes (PQStat Software Company, Poznan, Poland).

Correspondence: Barbara Więckowska, Department of Computer Science and Statistics, Poznan University of Medical Sciences, Rokietnicka 7 St., Poznan 60-806, Poland.
Tel. +48.618452606.
E-mail: basia@ump.edu.pl

Key words: Space-time statistics; Cluster detection; Spatial epidemiology; Poland; Kulldorf; CutL.

Acknowledgements: the authors are grateful to Tomasz Więckowski (PQStat Software Company, Poznan, Poland) for his valuable assistance in building PQScut software.

Conflict of interest: the authors declare no potential conflict of interest.

Funding: Statutory research of Department of Computer Science and Statistics, PUMS, 2715.

Received for publication: 6 June 2019.
Revision received: 9 October 2019.
Accepted for publication: 10 October 2019

Geospatial Health 2019; 14:791
doi:10.4081/gh.2019.791

Introduction

In epidemiology, everything happens in a particular place at a particular time, which means that the collected data have express both a spatial and a temporal context. The search for environmental risk factors related to the occurrence of a disease can be more precise if the search area is narrowed down with respect to time and space. The tools for locating clusters with an increased disease incidence rate are an indispensable element of epidemiological analyses. Currently, a number of spatial cluster detection methods are used in epidemiology, to indicate geographic cluster boundaries. However, there are fewer useful methods for space-time cluster detection than methods indicating either the geographic or the temporal boundaries of clusters (Robertson *et al.*, 2010; An *et al.*, 2015).

Depending on the type of data, and on the aim of the analysis, specific approaches to cluster detection analyses are advised. In model-based approaches, Bayes's methodology, for example, one can easily take into account the influence of covariates such as age, sex or smoking with respect to the risk for a certain disease to develop. Bayes's models in spatial analyses, mainly developed by Lawson (2013), are used in many studies on disease mapping as mentioned by Robertson *et al.* (2010). Estimations based on those models, however, require that previous distributions be determined for every component of a given model and that samples of posterior distribution be collected with the use of Markov chain Monte Carlo methods (Lawson, 2013; Wakefield and Kim, 2013), which is time-consuming. For that reason, those methods are not popular for longitudinal supervision of disease incidence rates. Having said that, they represent a set of tools that are effective for the investigation of large sets of rare cases, where there are few other techniques that work (Khana *et al.*, 2018).

In approaches based on statistical tests, the ones that offer the widest spectrum of adaptability, are Kulldorff's scan statistic, first applied to spatial analyses (Kulldorff, 1997) and later expanded by the addition of time (Kulldorff *et al.*, 1998, 2005; Kulldorff, 2001). The scan approach proposes cluster analysis for distribu-

tions based on various models, not only those developed by Bernoulli, Poisson and Gauss, but also permutation and multinomial models. The efficacy of these programmes results from the mechanism of its algorithm; space-time clusters are found with the use of circular or elliptic scanning windows of different diameters, which makes Kulldorff's scan statistic particularly suited to dealing with circular or ellipse-shaped clusters. The scanning windows are connected with a 3-D, cylindrical view whose height gives the period of time during which the cluster exists. The statistical significance of the proposed cluster is examined with the use of Monte Carlo simulations, which is a rather quick process when a circle-shaped window is used, but a much slower one when the window is elliptical. The main driving force behind the fast development and application of Kulldorff's scan statistic is the SaTScan programme developed by him {XE "software:SatScan"} (https://www.satscan.org) (Martin Kulldorff, Harvard Medical School, Boston, and Information Management Services Inc, Calverton, Maryland, USA), which makes it possible to make modifications and run a test, also devised by him, and a part of their modifications. However, the programme lacks the capability of searching for clusters of freely selected shapes and their intensity cannot be taken into account, as the tests fail to indicate which incidence rate of a given disease is typical and which should be viewed as too high. For that reason, sometimes clusters indicated with the use of that method are not characterised by a disease incidence rate which would qualify as unusually high in the eyes of an epidemiologist. Besides, when looking for clusters, the scan statistic compares the disease incidence rate within the scanning window with the rate outside of that window. Since there can also be clusters outside of the scanning window, it is difficult to locate potential clusters precisely and to indicate their statistical significance (Zhang *et al.*, 2010). That, in turn, obstructs the interpretation of the located clusters which do not, in such a case, constitute a sufficient narrowing down of the search for all environmental risk factors.

Another method that lends itself for cluster detection based on statistical testing is the CutL method (Więckowska and Marcinkowska, 2017) made available in the PQScut (http://pqscut.ump.edu.pl) and PQStat (http://www.pqstat.pl/en) programmes (Barbara Więckowska, Poznan University of Medical Sciences, Poznań, and PQStat Software Company, Poznań, Poland). It consists in searching for clusters with a disease incidence rate which is statistically more significant than the rate defined by the researcher. This approach provides the researcher with a unique level of control in defining the cut-off level in a given analysis. Furthermore, the possibility to define the cut-off level by the investigator allows for searches in areas where the frequency of an event is not alarmingly high but higher than the researcher expectation. That method allows searching for spatial clusters. In this work, an extension of the CutL method is presented which make it possible to look for space-time clusters. Also, the results of the use of the CutL method and of Kulldorff's scan statistic are compared, on the basis of simulation data.

Materials and Methods

The CutL method

The CutL method operates based on the population size n_i and data concerning the number of ill people d_i for particular O_i (i=1,…m) administrative areas. The method is used to search for clusters according to the cut-off level (X_{CutL}), which is set to the overall incidence rate by default if not given by the investigator. In case the investigator is interested in identifying clusters compared to the specified incidence rate, different areas or wider areas than those under study (*e.g.*, those reported in other countries), then the proposed cut-off level should be the incidence rate of the wider/different area. In the first step, anchoring points of the cluster structures can be found with the help of the cut-off level. They are those areas (countries) where the disease incidence rate is significantly higher than the cut-off level. In those calculations, the raw rate ratio $\frac{d_i}{n_i}$ is replaced with the smoothed coefficient based on local empirical Bayes smoothing. For the Bayes method, we assumed independent Poisson distribution for the observed count of events (conditional upon the risk parameter), and independent Gamma distribution for the prior of the risk parameter. In the Empirical approach, values for α and β of the prior Gamma distribution are estimated from the actual data. As a result, the rates for small counties (*i.e.* those with a small population at risk) tend to change/adjust considerably, whereas the rates for larger counties will barely change (Clayton and Kaldor, 1987; Anselin *et al.*, 2006). Next, clusters are constructed around the anchoring points. When overlapping, they later combine to form greater cluster agglomerates.

In order to extend the CutL method from spatial to space-time, particular steps of the spatial analyses have to change. The main change pertains to the definition of neighbourhood. In spatial analysis, object neighbourhood in space is indicated by a two-dimensional weight matrix, $W_{space}=[w_{ij}]$, where the value 1 means that objects are neighbours, and 0 the opposite. For the purpose of the space-time analysis, matrix W_{space} was extended to three dimensions – not only space, but also time $W_{space\text{-}time}=[w_{ijt}]$, where i=1,…m, j=1,…m, t=1,…,T with m represents the number of spatial objects and T the number of time layers. Neighbouring of objects in time is understood as direct neighbouring, that is if time is considered in years, then particular years (t_1, t_2, …, t_T) constitute time layers, and the objects that are neighbouring in time are the same objects, only located a year earlier or a year later on the time axis.

That matrix modification also demands that the smoothing method, that is the local empirical Bayes smoothing, be extended from two-dimensional to three-dimensional. In that way, the value of the smoothed incidence rate ratio of the event with respect to the object under study is made dependent on the value of the rates in the neighbouring objects and on the value of the rates in the phenomenon studied directly before and after the period for which the smoothing is performed (Eq. 1):

$$smooth(r_{it}) = \frac{smooth(d_{it})}{smooth(n_{it})} + C_{it}\left(\frac{d_{it}}{n_{it}} - \frac{smooth(d_{it})}{smooth(n_{it})}\right)$$

where $smooth(r_{it})$ – smoothed incidence rate within a county (i) in time (t),

$$smooth(d_{it}) = \frac{\sum_{j=1}^{m}\sum_{t=1}^{T} w_{ijt}d_j}{\sum_{j=1}^{m}\sum_{t=1}^{T} w_{ijt}},$$

Eq. 1

where $smooth(r_{it})$ – smoothed incidence rate within a county (i) in time (t),

$$smooth(d_{it}) = \frac{\sum_{j=1}^{m}\sum_{t=1}^{T} w_{ijt} d_j}{\sum_{j=1}^{m}\sum_{t=1}^{T} w_{ijt}},$$

$$smooth(n_{it}) = \frac{\sum_{j=1}^{m}\sum_{t=1}^{T} w_{ijt} n_j}{\sum_{j=1}^{m}\sum_{t=1}^{T} w_{ijt}},$$

C_{it} – shrink factor for county (i) in time (t).

The statistical significance of the clusters formed in that way is examined with the use of the binomial exact test for one proportion. That test compares the actual (unknown) incidence rate within a cluster ($R_{cluster}$), with the cut-off level (Eq. 2):

$$H_0 : (R_{cluster}) = X_{CutL} \quad \text{Eq. 2}$$

based on the known incidence rate within the cluster, $r_{cluster} = \frac{d_{cluster}}{n_{cluster}}$, where $d_{cluster}$ is the number of people affected by disease within the cluster, and $n_{cluster}$ the population size within the cluster. Because of the multiple testing, P-value of the detected clusters is corrected in accordance with the Benjamin-Hochberg correction (Benjamini and Hochberg, 1995).

A simulation for Polish counties of the number of people affected by disease

We use the population of Poland in 2013-2017 (N=192,278,050) as given by the Central Statistical Office of Poland (GUS, 2019) as the basis for the analysis. Poland has 380 counties. That division, in combination with a geographic information system, forms the basis of regional planning and health care. The administrative units differ greatly with respect to the number of inhabitants. The capital has the highest number of inhabitants: around 1,700,000 which only changes marginally. The least populated county has 20,000 inhabitants on average.

A simulation of the space-time distribution of the number of people who had very recently fallen ill was conducted. In order to ensure sufficiently great power to the analyses, the total number of recently affected people was set at the level d=19228. With that assumption, the overall incidence rate for Poland was $\frac{d}{N} = 0.0001$.

Three different space-time distributions were designed as follows:

1. No cluster, data distributed randomly in accordance with the multinomial distribution.

2a. Two separate clusters: a relatively circular one in central western Poland (2013-2015), which included 6.2% of the population in 2013 but gradually disappeared to include only 3.3% in 2015; and an oblong one, located in the South along the southern border of the country (2015-2017) that encompassed 1.5% of the population in 2015, while gradually growing to encompass 2.5% in 2017.

2b. A cluster following the course of the river which flows through the capital of Poland (2013-2017), which included 4.7% of the population but was not present the whole study period. It moved from the southern border in 2013 to the northern border in 2017.

The counties belonging to clusters 2a and 2b in the years indicated were accorded the status of true clusters. In order to obtain data illustrating those distributions, data concerning the number of affected people were generated on the basis of the multinomial distribution, according to the formula (Eq. 3):

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1T} \\ d_{21} & d_{22} & \cdots & d_{2T} \\ \cdots & \cdots & \cdots & \cdots \\ d_{m1} & d_{m2} & \cdots & d_{mT} \end{bmatrix} \sim multinomial \begin{bmatrix} RR_{11}\frac{d_{11}}{n_{11}} & RR_{12}\frac{d_{12}}{n_{12}} & \cdots & RR_{1T}\frac{d_{1T}}{n_{1T}} \\ RR_{21}\frac{d_{21}}{n_{21}} & RR_{22}\frac{d_{22}}{n_{22}} & \cdots & RR_{2T}\frac{d_{2T}}{n_{2T}} \\ \vdots & \vdots & & \vdots \\ RR_{m1}\frac{d_{m1}}{n_{m1}} & RR_{m2}\frac{d_{m2}}{n_{m2}} & \cdots & RR_{mT}\frac{d_{mT}}{n_{mT}} \end{bmatrix}$$

Eq. 3

where d_{it} is the number of affected people in the i^{th} county in year t; d the total number of affected people in counties m in years $T = \sum_{i=1}^{m}\sum_{t=1}^{T} d_{it}$ =192228, R_{it} the relative risk for county i in year t, depending on the indicated spatial distribution. Assuming that there were no clusters (1), that is, that the null hypothesis about constant relative risk is true, RR_{it}=1 for every county, each year. Assuming the presence of clusters (2a) and (2b), that is the truthfulness of the alternative hypothesis about differing risks, RR_{it}>1 was assumed for counties belonging to the clusters defined in a given year, that is for the clusters here defined as true.

In order to obtain distribution (1) and distributions (2a) and (2b), the data simulation procedure was repeated 500 times for each of the assumptions about the location and intensity of the clusters. For distributions representing clusters (2a) and (2b), the relative risk (RR) for the clusters was RR=1.5 and RR=2.5 and RR=4, respectively.

Cluster recognition

The planned space-time clusters in Poland in 2013-2017 were searched on the basis of a simulated number of affected people and of the population size in particular counties for each year. The tools for the search were the CutL method and Kulldorff's scan statistic based on Poisson's model, applied with the use of the PQStat programme and the SaTScan programme, respectively. Since a researcher usually does not know the size or shape of clusters being looked for, we used the default settings of Kulldorff's scan statistic. The only change was the maximal time of the existence of a cluster, which was increased from 50% to 60% in the case of the detection of the 2a clusters because one of the planned clusters existed for 3 years, which is 60% of the period of time analysed. In the CutL method, we have used a standard Queen adjacency matrix, and we have assumed the default settings with the cluster cut-off threshold selected on the basis of the overall incidence rate which was 0.0001 for the collected data. The counties with P-values smaller than the 0.05 significance level was classified as belonging to particular clusters with the use of the selected methods.

Summary of the cluster outcomes

The main goal of the analyses presented here was to determine the precision of the location and the size of the indicated clusters. To that end, we checked the degree to which the clusters located/found with the CutL method and Kulldorff's scan statistic overlap with the planned clusters (that is, the counties here called true clusters). Traditionally we mark true-positive values by TP, true-negative values by TN, false-positive values by FP and false-negative values by FN. The measures which described the accuracy of the detected clusters were: sensitivity = TP/(TP+FN), that is the ability of the method to correctly detect counties within the planned clusters; specificity = TN /(TN+FP), that is the ability of the method to correctly exclude the belonging of a given counties to a given cluster; positive predictive values (PPV),

PPV=TP/(TP+FP), that is proportions indicating which part of the counties assigned to the clusters by the given method really would have the status of a true cluster; negative predictive values (NPV), NPV=TN/(TN+FN), that is proportions indicating which part of the counties excluded by the given method from the clusters really would not have the status of a true cluster, and accuracy =(TP+TN)/(TP+TN+FP+FN), that is the general proportion of correctly classified counties. In the formulas making it possible to determine those measures, TP signified the number of counties indicated as belonging to the clusters and, at the same time, constituting true clusters, FP – the number of counties indicated as belonging to the clusters but not constituting true clusters, TN – the number of counties indicated as not belonging to the clusters and not constituting true clusters, and FN – the number of counties indicated as not belonging to the clusters but constituting true clusters.

The percentage of counties classified as clusters when the relative risk was 2.5 times higher within the planned true clusters (2a) and (2b) is presented on the maps with the use of the PQScut programme. Both simulated data and the files with maps and the obtained results can be downloaded from http://pqscut.ump.edu.pl.

Results

The simulated data obtained in this study allowed the creation of a situation in which the location of the space-time clusters was known, as well as the RR for people falling ill within the planned clusters. That made it possible to use both the CutL method and Kulldorff's scan statistic, and to determine their precision with the use of sensitivity, specificity, PPV, NPV, and accuracy (Table 1).

If a researcher wants to confirm that an area, in this case a county, belongs to a cluster (to detect true clusters but not to exclude that possibility of belonging), then the most important measures for evaluating the capability of the cluster detection methods are sensitivity (first) and PPVs (second).In the presented analyses of the simulated data, the values differed greatly for particular values of the RR, which made it possible to compare the methods used for cluster detection. In each case, the PPVs were higher for the CutL method, while the sensitivity of that method was lower than that of Kulldorff's scan statistic, but only for RR=1.5. The specificity and the NPVs responsible for the exclusion of a county from the clusters remained at a high level – exceeding 97% – for the whole time.

Both methods turned out to be highly effective for detecting two simultaneously existing clusters for RR≥2.5. The sensitivity and PPV for the CutL method exceeded 90%, but these scores were only a little lower for Kulldorff's scan statistic. The location of the two planned clusters in space and time as well as the results illustrating the precision of the indication of those clusters with the CutL method and Kulldorff's scan statistic for RR=2.5 are shown in Figure 1. However, the CutL method showed a clear advantage in the case of clusters located along the river where, for RR≥2.5, the obtained sensitivity and PPV at >90% were nearly twice as high as those obtained with Kulldorff's scan statistic. The CutL method correctly indicated the planned shape of the clusters, while Kulldorff's scan statistic mainly indicated cities within the area of the planned clusters, such as Cracow in 2013 and Warsaw, the capital city, in 2015 (Figure 2).

Discussion

Because of the great computational complexity of spatial and space-time analyses, they cannot be used on a large scale if not offered in computational tools. According to Robertson and Nelson (Robertson and Nelson, 2010), the SaTScan programme, which offers analyses of spatial cluster detection and space-time cluster detection, would be the best for an automated surveillance system. The field of spatial and space-time analyses is developing with increasing speed, which means that there is a need for new user-friendly methods. As has been noted in the summary of the article cycle on software for spatial analyses *Software for Spatial Statistics* (Pebesma *et al.*, 2015), the focus is free-license software, which indicates that that field is still in its early stages of development. In this work, we used the Polish version of the proprietary PQScut programme (also available in English) with the CutL method, which allows the detection of spatial and space-time clusters. The results obtained with both the CutL method and Kulldorff's scan statistic are of direct practical use when presented on a map.

The results of the comparison of the CutL method with Kulldorff's scan statistic for two regularly-shaped clusters were similar. For the use of Kulldorff's scan statistic, it is beneficial if clusters exist invariably in the same place. However, this method ignores the fact that the clusters grow or diminish over time (Figure 1), which is not a problem for the CutL method that manages to locate fragments of clusters regardless of population size and irregular cluster shape. Kulldorff's scan statistic had a big problem with locating the irregular shape of clusters, when they contained only counties with a small population. On the other

Table 1. Comparisons between the CutL method and Kulldorff's scan statistic for cluster detection based on 500 replications.

Cluster - RR		Sensitivity		Specificity		PPV		NPV		Accuracy	
		CutL	Scan*	CutL	Scan*	CutL	Scan*	CutL	Scan*	CutL	Scan*
(1)°	1.0	-	-	0.999	0.999	-	-	-	-	-	-
(2a)#	1.5	0.279	0.665	0.999	0.988	0.891	0.703	0.970	0.986	0.969	0.974
	2.5	0.903	0.859	0.998	0.984	0.951	0.701	0.996	0.994	0.994	0.979
	4.0	0.991	0.878	0.999	0.981	0.985	0.667	1.000	0.995	0.999	0.977
(2b)§	1.5	0.200	0.385	0.998	0.977	0.774	0.360	0.973	0.979	0.972	0.957
	2.5	0.909	0.572	0.997	0.972	0.923	0.416	0.997	0.985	0.994	0.959
	4.0	0.900	0.589	0.999	0.967	0.981	0.383	1.000	0.986	0.999	0.955

RR, relative risk; PPV, positive predictive values; NPV, negative predictive values; *Kulldorf's scan statistic; °null hypothesis with no clustering; #two clusters; §clusters located along the river.

hand, when the clusters included heavily populated counties (such as the country capital in 2016), this method built wrongly too large clusters – far exceeding the planned scope (Figure 2).

The lack of flexibility of Kulldorff's scan statistic has been noticed and discussed many times. Among the first analyses based on this statistic but with more flexible scanning windows have been proposed by Tango and Takahashi (2005), who developed a method together with the FleXScan software (Tango, 2008; Tango and Takahashi, 2012). Efforts have been made to extended this approach was for use with space-time clusters (Takahashi *et al.*, 2008), but the extension is not made available in the FleXScan software. As regards purely spatial data, flexible scan statistic is used for searching for relatively small clusters because when greater cluster sizes are desired, the time needed for running the test rapidly mounts with the size targeted.

Assunção *et al.* (2006) also tried to make Kulldorff's scan statistic more flexible by use of information about the neighbourhood structure, *e.g.*, adding subsequent neighbours through a common boundary. The method adds neighbours sequentially to the cluster whenever a neighbour maximizing the likelihood function is detected, and the process is repeated until the maximal cluster size is reached. This procedure usually creates large clusters which are close to the maximal cluster size and with very high probability values and thin geometric connections known as the octopus effect (Duczmal and Assunção, 2004).

Kulldorff himself has been looking for a way to extend the scan statistic (Duczmal *et al.*, 2006), presenting an idea for changing Assunção's algorithm in such a way that clusters do not grow to their maximal sizes if the connections among them are weak, that is based on a possibly small number of connections with the constructed cluster (Costa *et al.*, 2012). That is the double-connected spatial scan statistic – which has a weaker punishment function allowing greater loosening up of clusters, and a maximum linkage (Mlink) to spatial scan statistic, which has a stronger punishment function and thus capable of detecting tighter clusters. Choosing the size of the punishment function is not easy as the user needs to know the tightness of the detected clusters. Like flexible scan statistic, the Mlink was later extended to include the space-time aspect, but this extension is only available for the permutation model (Costa and Kulldorff, 2014), which can be used for examining very short periods of time (usually <1 year) because it only takes into account the number of cases of the given illness, as it can be assumed that the population size does not change in such a short time. The methods of making scanning windows more flexible proposed by Kulldorff are decidedly slower than the methods using a round or elliptical window, although they are faster than the flex-

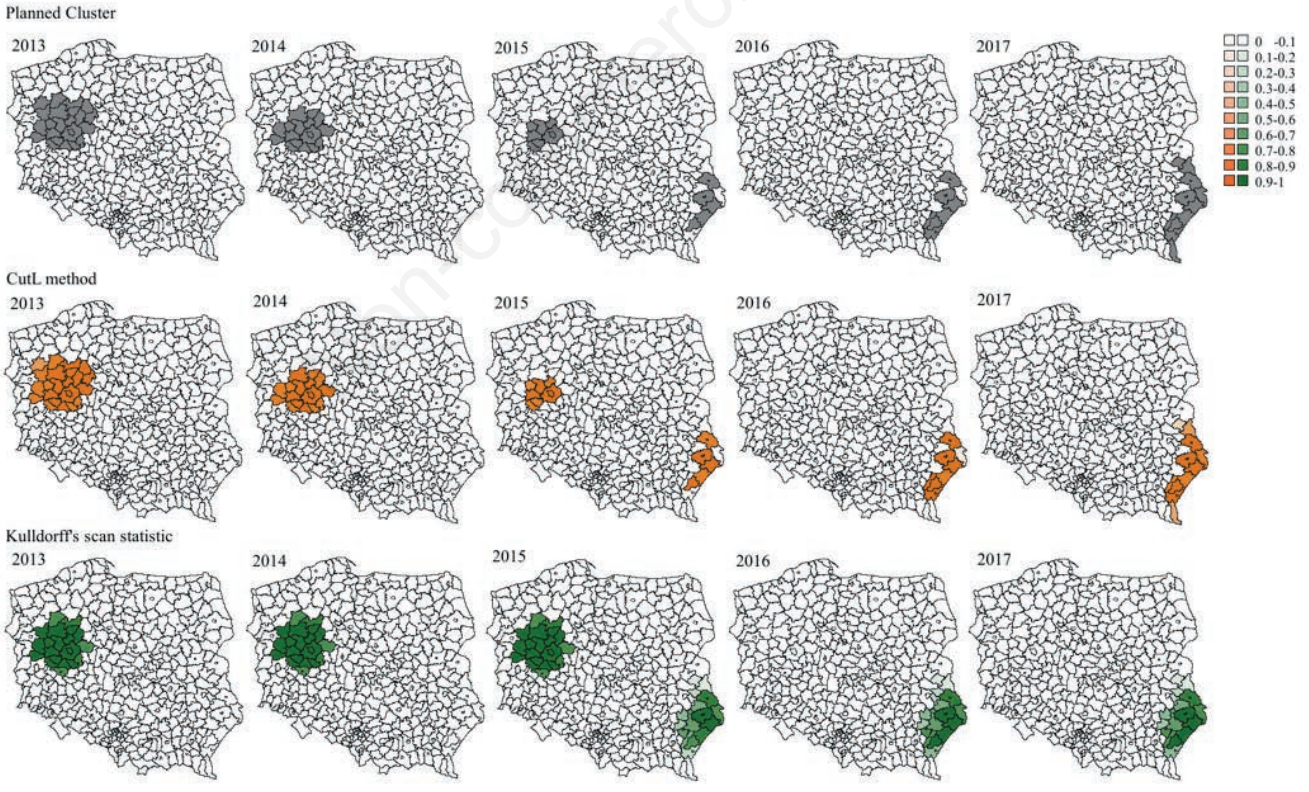

Figure 1. Simulation results comparing the CutL method and Kulldorff 's scan statistic based on 500 replications when applied to two separate clusters. The percentage of counties classified as clusters when the relative risk is 2.5 times higher within the planned true clusters (2a).

ible scan statistic (Costa *et al.*, 2012), which means that they are slower than the CutL method. Unlike the latter, neither of those two tests – although they had been tested by their author and designed in the C++ language – were not generally available in the programme, for example SaTScan, and have not come into general use. Iyengar (2005) has suggested, on the basis of his research on the influence of the use of flexible shapes on the located clusters of people suffering from brain cancer in New Mexico, that a more flexible shape of the window should result in greater insight into the clusters. On the other hand, irregular shapes are natural for random data distribution, so when clusters of any shape are looked for, clusters with test probabilities <0.05 can often be detected by accident. Our studies, however, did not justify those fears. In every possible case, PPVs and specificity, that is measures which take into account the number of falsely detected clusters (FP), were higher for the CutL method than for Kulldorff's scan statistic based on round windows. Therefore, the development of space-time cluster detection methods for detecting clusters of any shape appears to be a move in the right direction.

In addition to the methods of scanning space when searching for space-time clusters, other lattice-based local cluster methods have been extended. Examples include Local Indicators of Spatial Association (LISA) advanced by Anselin (1995), statistics including local Moran's statistic - extended from univariate to bivariate analysis of which space-time association is a special case (Anselin *et al.*, 2002) and the Geary's C method (Anselin, 2019) – extended to multi-dimensions analysis. In contrast to the Kulldorff's scan statistic, these methods do not assume a specific shape of future clusters, thus they can detect time-spatial clusters with greater accuracy. Therefore, further research is needed to compare methods CutL also with those techniques.

The CutL method, although it is much more effective than Kulldorff's scan statistic when irregularly-shaped clusters are searched for, and although the two methods yield comparable results in the case of tight clusters, is clearly weaker when it comes to detecting clusters of small intensity, that is for RR=1.5. Kulldorff's scan statistic is more than twice as sensitive as the CutL method. That is because, with the CutL method, the cluster construction begins from a county with an appropriately incidence rate ratio, together with its confidence interval. In the case of RR as low as 1.5, the CutL method may not be strong enough to begin the needed construction. Improvement of the efficacy of the CutL method will be attempted in further research.

Conclusions

Identification of geographical spatial clusters characterised by an excessive incidence of disease allows for narrowing the field of searching for environmental factors that can cause disease in these

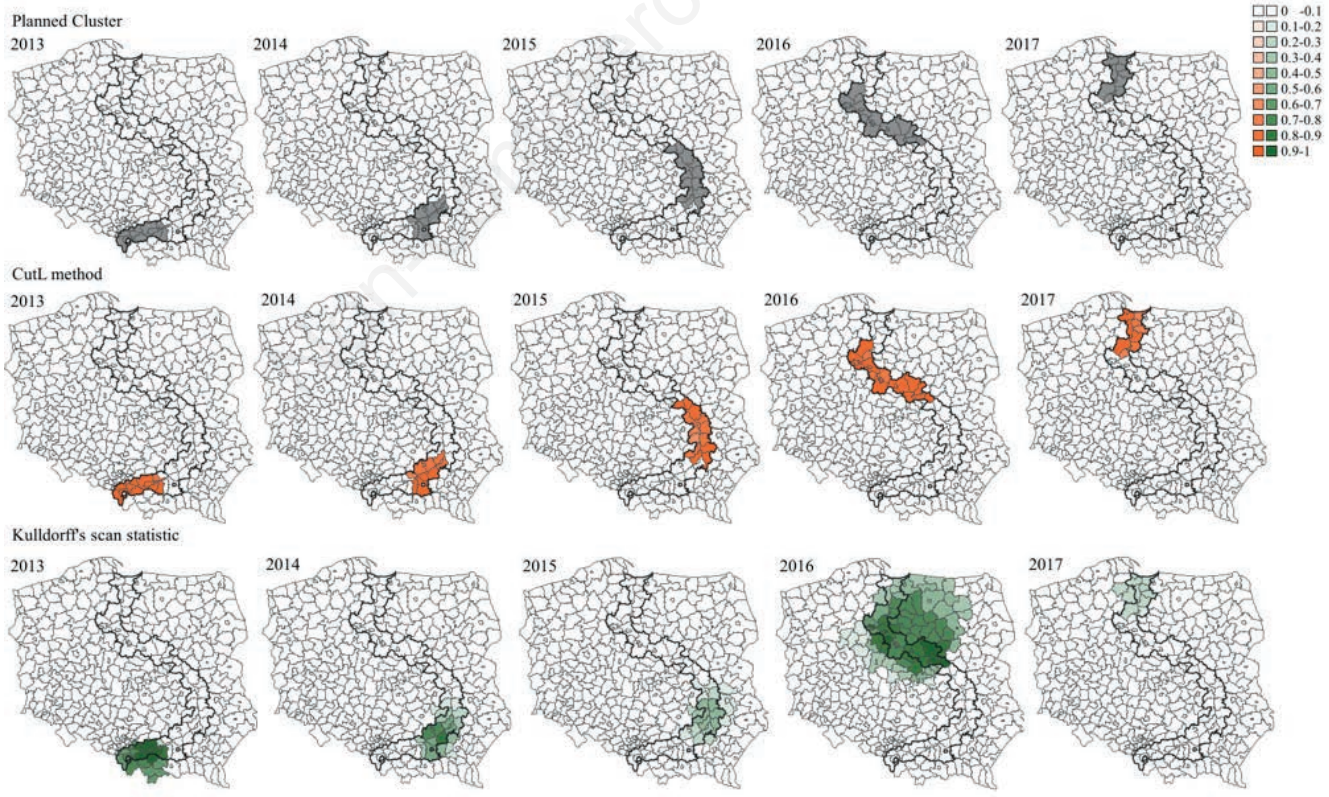

Figure 2. Simulation results comparing the CutL method and Kulldorff's scan statistic based on 500 replications for a cluster following the course of the river through the country capital. The percentage of counties classified as clusters when the relative risk is 2.5 times higher within the planned true clusters (2b).

clusters. The results of this study show the high efficacy of the new CutL method in the detection of time-space clusters of any shape.

References

An L, Tsou M-H, Crook SES, Chun Y, Spitzberg B, Gawron JM, Gupta DK, 2015. Space-time analysis: concepts, quantitative methods, and future directions. Ann Assoc Am Geogr 105:891-914.

Anselin L, 1995. Local indicators of spatial association – LISA. Geogr Anal 27:93-115.

Anselin L, 2019. A Local indicator of multivariate spatial association: extending Geary's c. Geogr Anal 51:133-50.

Anselin L, Lozano-Gracia N, Koschinky J, 2006. Rate transformations and smoothing. Technical report. Spatial Analysis Laboratory, Department of Geography, University of Illinois, Urbana, IL, USA.

Anselin L, Syabri I, Smirnov O, 2002. Visualizing multivariate spatial correlation with dynamically linked windows. In: Anselin L, Rey S, eds. New tools for spatial data analysis: proceedings of the specialist meeting. Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara, USA.

Assunção R, Costa M, Tavares A, Ferreira S, 2006. Fast detection of arbitrarily shaped disease clusters. Stat Med 25:723-42.

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

Clayton D, Kaldor J, 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43:671-81.

Costa MA, Assunção RM, Kulldorff M, 2012. Constrained spanning tree algorithms for irregularly-shaped spatial clustering. Comput Stat Data Anal 56:1771-83.

Costa MA, Kulldorff M, 2014. Maximum linkage space-time permutation scan statistics for disease outbreak detection. Int J Health Geogr 13:20.

Duczmal L, Assunção R, 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Comput Stat Data Anal 45:269-86.

Duczmal L, Kulldorff M, Huang L, 2006. Evaluation of spatial scan statistics for irregularly shaped clusters. J Computat Graph Stat 15:428-42.

GUS, 2019. Główny Urząd Statystyczny. Central Statistical Office of Poland, Poland. Available from: http://stat.gov.pl/ Accessed: February 6, 2019.

Iyengar VS, 2005. Space-time clusters with flexible shapes. MMWR-Morbid Mortal W 54:71-6.

Khana D, Rossen LM, Hedegaard H, Warner M, 2018. A Bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with R-INLA. J Data Sci 16:147-82.

Kulldorff M, 1997. A spatial scan statistic. Commun Stat Theory Method 26:1481-96.

Kulldorff M, 2001. Prospective time periodic geographical disease surveillance using a scan statistic. J R Stat Soc Ser A Stat Soc 164:61-72.

Kulldorff M, Athas WF, Feurer EJ, Miller BA, Key CR, 1998. Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health 88:1377-80.

Kulldorff M, Heffernan R, Hartman J, Assunçao R, Mostashari F, 2005. A space-time permutation scan statistic for disease outbreak detection. PLoS Med 2:e59.

Lawson AB, 2013. Bayesian disease mapping: hierarchical modeling in spatial epidemiology. Chapman and Hall/CRC Press, Boca Raton, FL, USA.

Pebesma E, Bivand R, Ribeiro PJ, 2015. Software for spatial statistics. J Stat Softw 63:1-8.

Robertson C, Nelson TA, 2010. Review of software for space-time disease surveillance. Int J Health Geogr 9:16.

Takahashi K, Kulldorff M, Tango T, Yih K, 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. Int J Health Geogr 7:14.

Tango T, 2008. A spatial scan statistic with a restricted likelihood ratio. Japan J Biometr 29:75-95.

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.

Wakefield J, Kim A, 2013. A Bayesian model for cluster detection. Biostatistics 14:752-65.

Więckowska B, Marcinkowska J, 2017. CutL: an alternative to Kulldorff's scan statistics for cluster detection with a specified cut-off level. Geospatial Health 12:556.

Zhang Z, Assunção R, Kulldorff M, 2010. Spatial scan statistics adjusted for multiple clusters. J Probab Stat 2010:642379.