Examining the impact of the number of regions used in cluster detection methods: An application to childhood asthma visits to a hospital in Manitoba, Canada

Mahmoud Torabi, Katie Galloway

Department of Community Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada

Abstract

The level of spatial aggregation is a major concern in cluster investigations. Combining regions to protect privacy may result in a loss of power and thus, can limit the information researchers can obtain. The impact of spatial aggregation on the ability to detect clusters is examined in this study, which shows the importance of choosing the correct level of spatial aggregation in cluster investigations. We applied the circular spatial scan statistic (CSS), flexible spatial scan statistic (FSS) and Bayesian disease mapping (BYM) approaches to a dataset containing childhood asthma visits to a hospital in Manitoba, Canada, using three different levels of spatial aggregation. Specifically, we used 56, 67 and 220 regions in the analysis, respectively. It is expected that the three scenarios will yield different results and will highlight the importance of using the right level of spatial aggregation. The three methods (CSS, FSS, BYM) examined in this study performed similarly when detecting potential clusters. However, for different levels of spatial aggregation, the potential clusters identified were different. As the number of regions used in the analysis increased, the total area identified in the cluster decreased. In general, potential clusters were identified in the central and northern parts of Manitoba. Overall, it is crucial to identify the appropriate number of regions to study spatial patterns of disease as it directly affects the results and consequently the conclusions. Additional investigation through future work is needed to determine which scenario of spatial aggregation is best.

Correspondence: Mahmoud Torabi, Department of Community Health Sciences, University of Manitoba, 750 Bannatyne Avenue, Winnipeg, Manitoba R3E 0W3, Canada.
Tel.: +1.204.272.3136 - Fax: +1.204.789.3905.
E-mail: Mahmoud.Torabi@umanitoba.ca

Acknowledgements: the authors would like to thank Monsur Chowdhury and Mili Roy for their help with the manuscript preparation.

Key words: Asthma; Bayesian computation; Geographic epidemiology; Spatial aggregation; Spatial cluster detection.

See online Appendix for additional materials.

Contributions: The analyses were performed by MT. MT and KG were involved in the study concept and design, acquisition of data, interpretation of data, and critical revision of the manuscript for important intellectual content.

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

Funding: Children's Hospital Research Institute of Manitoba (CHRIM), Natural Sciences and Engineering Research Council of Canada (NSERC).

Ethical statement: this study was approved by the Research Ethics Board of the University of Manitoba and the Health Information Privacy Committee. The Manitoba Health Information Privacy Committee approved this study for publication.

Received for publication: 10 April 2018.
Revision received: 21 August 2018.
Accepted for publication: 30 September 2018.

©Copyright M. Torabi and K. Galloway, 2018
Licensee PAGEPress, Italy
Geospatial Health 2018; 13:696
doi:10.4081/gh.2018.696

This article is distributed under the terms of the Creative Commons Attribution Noncommercial License (CC BY-NC 4.0) which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Non-commercial use only

Introduction

In the context of epidemiology, a spatial cluster is defined as a geographical region within the overall study area with unusual aggregation of disease (Lawson, 2006). The identification of clusters using cluster detection tools is helpful to initially identify regions with high rates of disease and are ultimately most at risk. Thus, the identification of these clusters plays an important part in epidemiological research and disease surveillance in public health. Cluster detection may uncover risk factors associated with disease and provide insight into disease aetiology (Jennings *et al.*, 2005). Moreover, public health agencies may use cluster detection to guide the implementation of disease interventions.

The majority of cluster detection procedures investigate whether the current rate of disease differs from the expected number of cases with respect to the population of the region of interest. One group of cluster detection methods are the general approaches, which include the Besag and Newell (Besag and Newell, 1991) and maximizing excess event test (Tango, 2000) approaches. The purpose of these procedures is to detect a potential cluster without specifying a specific area of interest. Another group of cluster detection approaches are the focused methods. These procedures aim to detect regions with high numbers of disease within a specified area and test the null hypothesis that no spatial clustering exists (Lawson *et al.*, 1999; Elliott *et al.*, 2001). This paper focuses on the circular spatial scan statistic (CSS) (Kulldorff, 1997),

flexible spatial scan statistic (FSS) (Tango and Takahashi, 2005) and Bayesian disease mapping (BYM) (Clayton and Kaldor, 1987; Besag *et al.*, 1991; Clayton and Bernardinelli, 1996) approaches, which are three of the most popular focused cluster detection methods.

While an extensive number of cluster detection methods exist, protecting the privacy and confidentiality of patients is a concern for epidemiological researchers and public health officials (AbdelMalik *et al.*, 2008), no matter what cluster detection approach is utilized. Numerous studies have addressed privacy and confidentiality concerns in population health and spatial analyses (Cox, 1996; Armstrong *et al.*, 1999; AbdelMalik *et al.*, 2008; O'Keefe and Rubin, 2015). Data modification, where counts are rounded up to get a sufficient number of cases and data aggregation, where numerous years of data are combined to get adequate counts are some techniques used to protect the privacy and confidentiality of disease cases (Cox, 1996). An alternative approach is spatial aggregation, which is the process of combining regions in order to get a suitable number of disease cases so that the cases are unidentifiable (Cox, 1996; Armstrong *et al.*, 1999). Conversely, by combining years of data or regions, public health officials or researchers are limited in the information they can obtain from spatial analyses. There is evidence that there is a loss of power when regions are aggregated. It has also been shown that there is an increase in the number of regions that are falsely detected as being part of a cluster when the regions are aggregated (Olson *et al.*, 2006; Ozonoff *et al.*, 2007).

Previously, three focused cluster detection methods (CSS, FSS and BYM) were applied to a dataset containing childhood asthma visits to a hospital in Manitoba, Canada in order to investigate potential clusters (Torabi, 2012). As an extension of that work, we studied the impact of spatial aggregation in this paper. In particular, we applied the previously mentioned approaches (CSS, FFS and BYM) to the same asthma dataset using three different levels of spatial aggregation. Specifically, we used three different numbers of regions (56, 67 and 220) in the analysis. The results were compared in terms of potential clusters for the three different numbers of regions.

Materials and Methods

This retrospective study was conducted using data provided by Manitoba Health through the Manitoba Centre for Health Policy. A yearly dataset of asthma visits to hospital by children (age ≤ 18) during the 2000-2010 fiscal years in the province of Manitoba, Canada was used in the analysis. During the study period, the delivery of health care services in the province was distributed across eleven Regional Health Authorities (RHAs). These regions were further divided into 56 smaller regions called Regional Health Authority Districts (RHADs). This is the geographical unit utilized in the analysis of 56 regions. In the analysis using 67 regions, the same RHADs are used, except the major urban centre in the province, the City of Winnipeg, is now divided into 12 sub-

Non-commercial use only

Table 1. The order of significant regions for the circular spatial scan statistic, flexible spatial scan statistic, and Bayesian disease mapping methods in the case of 56 and 67 regions.

Region	C_i	E_i	56 regions			67 regions		
			CSS	FSS	BYM	CSS	FSS	BYM
10	273	121	1	-	-	1	-	-
14	156	80	1	-	-	1	-	-
20	362	229	1	-	-	1	-	-
21	292	138	1	-	-	1	-	-
25	296	170	-	1	-	-	-	-
26	359	124	1	1	6	1	-	-
28	356	105	-	-	4	-	-	5
29	394	213	-	1	-	-	-	-
30	209	65	-	-	-	-	-	6
31	333	231	1	-	-	1	-	-
32	135	48	-	1	-	-	-	-
33	73	23	1	1	-	1	1	-
34	218	52	1	1	3	1	1	2
35	257	96	1	1	-	1	1	-
36	327	91	1	1	5	1	1	3
37	624	167	1	1	1	1	1	1
38	49	16	1	1	-	1	1	-
39	117	33	-	1	-	-	1	-
40	240	80	1	1	-	1	1	-
41	268	70	1	1	2	1	-	4

CSS: Circular spatial scan statistic, FSS: Flexible spatial scan statistic, BYM: Bayesian disease mapping; C_i: the observed number of cases in region i; E_i: the expected number of cases in region i.

regions. Manitoba municipalities, along with the 12 Winnipeg sub-regions comprise the units used in the analysis of 220 regions. The data used in this study was aggregated over the study period and linked to these non-overlapping geographical units, which are labeled as 1,2,…,56 (or 67, or 220) for simplicity. A population-based centroid was provided for each unit, although, these centroids are not necessarily geographical centres.

The observed as well as the expected number of asthma visits and the population size for each region are the necessary data components for each of the cluster detection methods used in this study. Here, the expected numbers of asthma visits to a hospital in each region were adjusted by year (1-10) and gender (male, female).

The CSS, FSS and BYM methods were applied to a dataset containing childhood asthma visits to a hospital in Manitoba, Canada. The identification of clusters in this study is the first step in spatial modelling. A technical description of these approaches can be found in the Appendix. In the CSS and FSS approaches, the maximum number of regions included in a cluster is 15. However, this is not a requirement in the BYM procedure. In the BYM method, we use a gamma distribution as a prior for the inverse of variance component (to account for variation of spatial random effects), with shape and scale parameter 0.001. A Normal Distribution with mean 0 and variance 10^6 was used as a prior for the fixed effect. For the model-based cluster identification method (BYM), the deviance residual (McCullagh and Nelder, 1989) was checked to ensure the model fits the data well. If the deviance residual suggests that the model does not fit the data appropriately, the study results may indicate incorrect clusters.

Results

The comparison of the results for the CSS, FSS and BYM cluster detection methods is provided to identify the potential clusters of childhood asthma visits to hospital for the period of 10 years (2000-2010) in the province of Manitoba, Canada, for different levels of spatial aggregation (56, 67 and 220 regions). The deviance residual was checked for the BYM approach and no violation was observed in terms of the pattern of the residuals.

The population range of Manitoba was relatively steady during the study with 1.15 million people in 2000 and 1.20 million people in 2010. Over the study period, there was an average population of 336,000 children in Manitoba and there was a total of 14,691 childhood asthma visits to hospitals. In the case of 56 regions, the mean and median numbers of yearly cases per region were 26 and 17 (range 3 to 422), respectively. For 67 regions, the mean and median numbers of yearly cases per region were 21 and 19 (range 3 to 72), respectively. Finally, in the case of 220 regions, the mean and median numbers of yearly cases per region were 6 and 3 (ranging from 0 to 72), respectively.

In general, we found that the northern and central parts of the province of Manitoba contained possible clusters of childhood asthma visits to hospital. In Figures 1-3 the areas that were statistically significant (potential clusters) are shaded for each method (CSS, FSS, BYN) and number of regions (56, 67, and 220), respectively. Tables 1 and 2 present a summary of the results for each method (CSS, FSS, BYN) and number of regions (56, 67, and 220), respectively. The regions that were most likely to constitute a cluster are ordered and presented in these Tables as well.

Based on 56 regions, the CSS, FSS and BYM methods, respec-

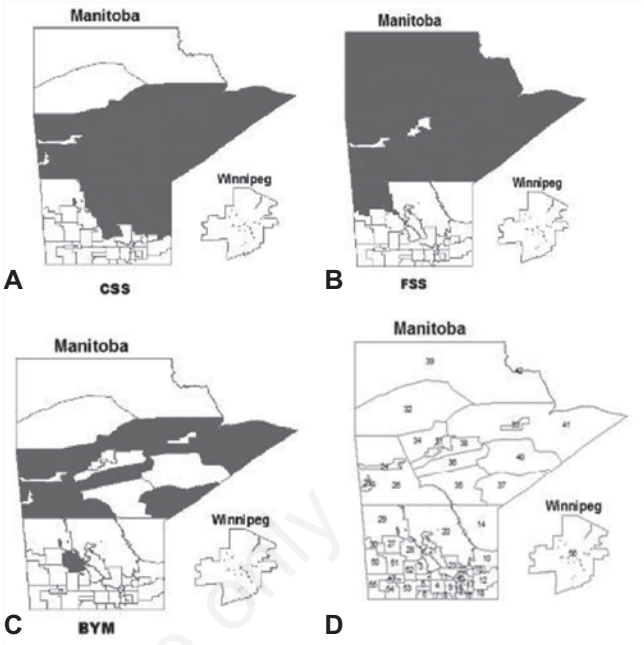

Figure 1. Potential clusters (shaded regions) identified by the three methods: (A) circular spatial scan statistic (CSS), (B) flexible spatial scan statistic (FSS), and (C) Bayesian disease mapping (BYM) in the case of 56 regions; (D) is the blank map with the region number.

Table 2. The order of significant regions for the circular spatial scan statistic, flexible spatial scan statistic, and Bayesian disease mapping methods in the case of 220 regions.

Region	C_i	E_i	CSS	FSS	BYM
199	460	163	1	1	1
200	176	70	1	1	-
201	18	2	1	1	-
202	244	98	1	1	-
203	51	14	1	1	-
204	78	19	1	1	-
205	229	109	1	1	-
206	213	63	-	1	2
207	104	43	1	1	-
208	9	3	1	1	-
209	44	9	1	1	-
210	103	32	1	1	-
211	54	29	1	1	-
212	24	13	-	-	-
213	45	23	-	-	-
214	39	13	-	-	-
215	38	12	-	-	-
216	18	4	-	-	-
217	104	43	-	-	-
218	15	13	-	-	-
219	55	13	-	-	-
220	98	23	1	-	-

CSS: Circular spatial scan statistic, FSS: Flexible spatial scan statistic, BYM: Bayesian disease mapping; C_i: the observed number of cases in region i; E_i: the expected number of cases in region i.

Non-commercial use only

tively, found 14, 13 and 6 regions as potential clusters. Regions 26, 34, 36, 37, and 41 were identified by all three methods as potential clusters. In the case of 67 regions, the CSS approach detected 14 regions, while the FSS method found 8 regions and the BYM technique identified 6 regions as potential clusters. Here we observed that most of the regions identified by the FSS method were also found in the CSS method and the majority of the regions detected by the BYM approach as potential clusters were also found by the CSS and FSS procedures. Based on 220 regions, 13 regions were identified as potential clusters by both the CSS and FSS methods, while the BYM method only identified 2 regions as potential clusters. It should be noted that region 199 was detected by all three methods, while region 206 was only identified as a possible cluster by the FSS and BYM methods.

The maps in Figures 1 and 2 display the results for the analyses with 56 and 67 regions, respectively. Here, the CSS method identified the same regions in central Manitoba in both analyses. Meanwhile, for the FSS and BYM approaches, there were some differences between the results with 56 and 67 regions. In Figure 1, the FSS method identified most of northern Manitoba as a potential cluster, whereas, in Figure 2, only some regions in northern Manitoba were found in the cluster. For the BYM procedure, in Figure 1, cluster regions in the north-central part of the province were found, with a small region in southern Manitoba identified as well. In Figure 2, most of the same regions were identified, with the exception of one region along the western Manitoba border. There was also an additional such region found in southern Manitoba. The maps in Figure 3 display the results for the analysis with 220 regions. Here, one large region in north-central Manitoba was found as a potential cluster with all three methods. The different approaches also identified various smaller regions contained within this one larger region. The results for 220 regions showed that a much smaller area was identified as a potential cluster when a lower level of spatial aggregation was used in the analysis.

Discussion

We employed three different focused cluster detection methods, namely the CSS, FSS and BYM approaches to detect potential clusters with high ratios of childhood asthma visits to hospital in the province of Manitoba, Canada. We also used three different numbers of regions (56, 67 and 220) to study their impacts in identifying potential clusters. We observed that using a different number of regions produced different potential clusters for each of the methods. In theory, if we would have different numbers of regions, we would also have different spatial information for the inference to detect potential clusters. In terms of the different approaches, the FSS method may have had a better ability to detect clusters in Manitoba due to the non-circular shape of the regions there. The BYM must also be deemed flexible, especially if the study interest goes beyond clustering to include spatial regression modelling.

All three approaches (CSS, FSS and BYM) identified the same three regions {34, 36, and 37} as potential clusters in the study when using both 56 and 67 regions in the analysis. Moreover, the regions {33, 34, 35, 36, 37, 38, and 40} were identified as potential clusters by the CSS and FSS procedures in the study when using both 56 and 67 regions in the analysis. The BYM procedure detected the same clusters {28, 34, 36, 37, 41} when using both 56 and 67 regions in the analysis, with the exception of regions 26 and 30. When utilizing 220 regions, northern Manitoba was shown to con-

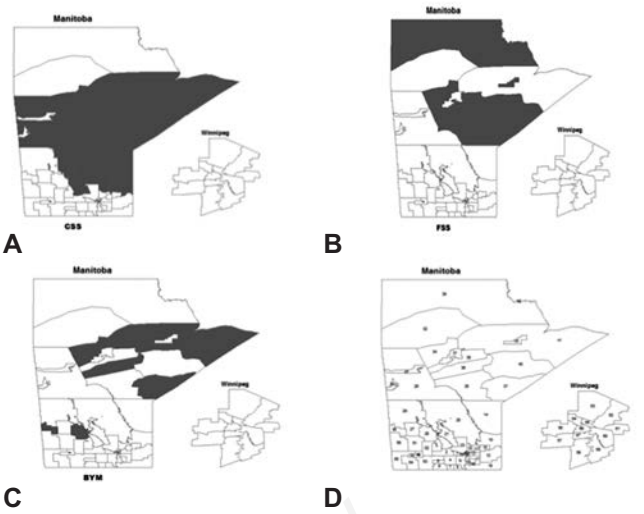

Figure 2. Potential clusters (shaded regions) identified by the three methods: (A) circular spatial scan statistic (CSS), (B) flexible spatial scan statistic (FSS), and (C) Bayesian disease mapping (BYM) in the case of 67 regions; (D) is the blank map with the region number.

Figure 3. Potential clusters (shaded regions) identified by the three methods: (A) circular spatial scan statistic (CSS), (B) flexible spatial scan statistic (FSS), and (C) Bayesian disease mapping (BYM) in the case of 220 regions; (D) is the blank map with the region number.

Non-commercial use only

tain many small regions contained within a larger region. Interestingly, the BYM method did not include the majority of these smaller regions in the cluster, contrary to the CSS and FSS approaches. There are advantages and disadvantages to the level of spatial aggregation used in the analysis. There is evidence that using a high level of spatial aggregation contributes to a loss of power. Subsequently, it has been suggested that this may lead to some regions being falsely identified as being part of a cluster (Olson *et al.*, 2006; Ozonoff *et al.*, 2007). When a high level of spatial aggregation is used in the analysis, the precision of the cluster is lost. If a large area is identified as a cluster in public health cluster investigations, it would be difficult for officials to narrow down the regions where further investigation is needed or where disease interventions are warranted. Alternately, when a large number of regions are used in the analysis there may be many regions with a small number of cases and privacy or confidentiality protection may become a concern for researchers or public health officials (Cox, 1996; Armstrong *et al*., 1999; AbdelMalik *et al*., 2008; O'Keefe and Rubin, 2015). One approach to solve this would be to use multiple years of data in the analysis. However, then the analysis would not be based on the most recent years of data available, which could impact the results. For these reasons, it is important to consider the trade-offs of the level of spatial aggregation used in the study prior to the start of the analysis.

Jeffery *et al*. (2009) studied the effect of the strength of the spatial disturbance in the disease cluster on level of spatial aggregation used in the study. They found that stronger spatial disturbances are better detected when a small level of spatial aggregation is used in the analysis. Conversely, when the spatial disturbance is weak, the cluster is better detected using higher levels of aggregation. However, when public health officials are conducting disease surveillance, the strength of the spatial disturbance or the strength of the disease cluster is unknown. Therefore, Jeffery *et al*. (2009) suggest looking at several levels of precision rather than choosing just one. It is of interest to consider an entirely different approach to protecting the privacy and confidentiality of disease cases while maintaining a low level of spatial aggregation. Geographic masking is a process, which alters the original location of the disease case by incorporating a degree of randomization into the geocoding process (Armstrong *et al.*, 1999). While the geographic masking approach avoids the need for spatial aggregation, it is still important to consider the masking technique and the level of displacement used to ensure the privacy and confidentiality of patients is maintained. Although the risk of identification is reduced with geographic masking, it is possible the masked location may be close to the original location. As well, the risk of identification increases if there are multiple versions of the masked dataset released (Zimmerman and Pavlik, 2008; Zandbergen, 2014).

Several studies have compared the performance of both focused and general cluster detection methods (Song and Kulldorff, 2003; Aamodt *et al.*, 2006; Waller *et al.*, 2006; Tango, 2008; Tango and Takahashi, 2012; Lemke *et al.*, 2013). It has been shown the performance of most tests depends on the shape of the cluster, size of the relative risk and location of the cluster. Aamodt *et al.* (2006) found the BYM method performs better when the relative risk is high, compared to the CSS approach, which performs better for low relative risks. Tango and Takahashi (2012) found the FSS method had better predictive ability than the CSS approach when the true clusters were not circular in shape. This may explain the difference in regions identified as potential clusters for the CSS and FSS methods in our study. The FSS method had the ability to identify the non-circular shaped regions in Manitoba as potential clusters, whereas the CSS method could not.

A limitation to this study is that each of the cluster detection methods used had user-chosen settings. In the CSS and FSS methods, the choice of the maximum number of regions to include in a cluster can affect the results. In the BYM approach, decision rules used to determine a cluster can vary, which would impact the results. Different decision rules may be defined where the estimated relative risk (in terms of the credible/prediction interval) would be larger or smaller than two (Richardson *et al.*, 2004). The exceedance probability could also be employed when establishing the significance of a cluster (Banerjee *et al.*, 2004). Finally, since the current study is an application of the methods to a real dataset of asthma visits to hospital, we do not know where the true clusters lie and thus, we cannot draw conclusions on which method and which scenario of spatial aggregation performs best in our study. Future work to resolve this issue would include a simulation study where the true cluster is known in order to make direct comparisons on the performance of each cluster detection method. As well, a future study would allow us to draw conclusions on which level of spatial aggregation performs best. Another future study would include using spatial regression modelling to identify possible risk factors to explain the asthma clustering.

One strength of our study is the fact that we compared multiple cluster detection methods. Additionally, we examined multiple scenarios of different spatial aggregation. We started with 56 regions in the analysis since these are sub-regions of the eleven RHAs, which were responsible for health care delivery in the province during the time of this study. We looked at two similar scenarios in the analysis with 56 and 67 regions. The only difference between these two situations was the major urban centre in Manitoba (Winnipeg) was divided into sub-regions in the analysis with 67 regions. Since the second largest city in Manitoba has a population of less than 50,000, Winnipeg was the only city divided into sub-regions. Interestingly, even with the minor difference in the number of regions in the analyses with 56 and 67 regions, we still noticed a difference in the cluster results between these two scenarios. We also looked at a completely different scenario with 220 regions, where there was little spatial aggregation. Finally, this study includes an application of the methods to a real dataset of childhood asthma visits to a hospital in the province of Manitoba, Canada. The analysis with 220 regions yielded the smallest cluster region, which, in terms of health care interventions, would be the most practical to deal with. However, further investigation is needed to determine which scenario of spatial aggregation is best.

Conclusions

Overall, it is crucial to identify the appropriate number of regions to study spatial patterns of disease as it directly affects the results and consequently the conclusions. For these reasons, it is important to consider the trade-offs between privacy and the precision of the results in the decision regarding the level of spatial aggregation used in the study prior to the start of the analysis. Further exploration through our future work is needed to determine which scenario of spatial aggregation is best.

Non-commercial use only

References

Aamodt G, Samuelsen SO, Skrondal A, 2006. A simulation study of three methods for detecting disease clusters. Int J Health Geogr 5:15.

AbdelMalik P, Kamel Boulos MN, Jones R, 2008. The perceived impact of location privacy: A web-based survey of public health perspectives and requirements in the UK and Canada. BMC Public Health 8:156.

Armstrong MP, Rushton G, Zimmerman DL, 1999. Geographically masking health data to preserve confidentiality. Stat Med 18:497-525.

Banerjee S, Gelfand AE, Carlin BP, 2004. Hierarchical modeling and analysis for spatial data. London, UK: Chapman and Hall.

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

Besag JE, York JC, Mollìe A, 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). Ann Inst Statist Math 43:1-59.

Clayton D, Bernardinelli L, 1996. Bayesian methods for mapping disease risk. In: Elliott P, Cuzick J, English D, Stern R, eds. Geographical and environmental epidemiology: methods for small-area studies. Oxford, UK: Oxford University Press.

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

Cox LH, 1996. Protecting confidentiality in small population health and environmental statistics. Stat Med 15:1895-905.

Elliott P, Briggs D, Morris S, de Hoogh C, Hurt C, Jensen TK, Maitland I, Richardson S, Wakefield J, Jarup L, 2001. Risk of adverse birth outcomes in populations living near landfill sites. Br Med J 323:363-8.

Jeffery C, Ozonoff A, White LF, Nuno M, Pagano M, 2009. Power to detect spatial disturbances under different levels of geographic aggregation. J Am Med Inform Assn 16:847-54.

Jennings JM, Curriero FC, Celentano D, Ellen JM, 2005. Geographic identification of high gonorrhea transmission areas in Baltimore, Maryland. Am J Epid 161:73-80.

Kulldorff M, 1997. A spatial scan statistic. Comm Statist: Theor Meth 26:1481-96.

Lawson AB, 2006. Statistical methods in spatial epidemiology. London, UK: John Wiley & Sons, Ltd.

Lawson AB, Biggeri A, Williams FLR, 1999. A review of modeling approaches in health risk assessment around putative sources. In: Lawson AB, Biggeri A, Böhning D, Lesaffre E, Viel J, Bertollini R, eds. Disease mapping and risk assessment for public health. New York, NY: Wiley. pp 231-245.

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 12:54.

McCullagh P, Nelder JA, 1989. Generalized Linear Models. London, UK: Chapman and Hall.

O'Keefe CM, Rubin DB, 2015. Individual privacy versus public good: protecting confidentiality in health research. Stat Med 34:3081-103.

Olson K, Grannis SJ, Mandl KD, 2006. Privacy protection versus cluster detection in spatial epidemiology. Am J Public Health 96:2002-8.

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 6:52.

Richardson S, Thomson A, Best N, Elliott P, 2004. Interpreting posterior risk estimates in disease-mapping studies. Environ Health Persp 112:1016-25.

Song C, Kulldorff M, 2003. Power evaluation of disease clustering tests. Int J Health Geogr 2:9.

Tango T, 2000. A test for spatial disease clustering adjusted for multiple testing. Stat Med 19:191-204.

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 4:1-15.

Tango T, Takahashi K, 2012. A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. Stat Med 31:4207-18.

Torabi M, 2012. Spatial disease cluster detection: An application to childhood asthma in Manitoba, Canada. J Biometrics Biostat S7:010.

Waller LA, Hill EG, Rudd RA, 2006. The geography of power: Statistical performance of tests of clusters and clustering in heterogeneous populations. Stat Med 25:853-65.

Zandbergen PA, 2014. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. Adv Med 2014:567049.

Zimmerman DL, Pavlik C, 2008. Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. Geogr Anal 40:52-76.

Non-commercial use only

OPEN ACCESS

Geospatial Health
Health Applications in Geospatial Science

Università degli Studi di Napoli Federico II
Dipartimento Medicina Veterinaria Produzioni Animali
pagepress

Appendix

Geospatial Health #696

Examining the impact of the number of regions used in cluster detection methods: an application to childhood asthma visits to a hospital in Manitoba, Canada

The CSS, FSS, and BYM spatial focused cluster detection methods are outlined below.

Circular Spatial Scan Statistic (CSS)

Kulldorff proposed the spatial scan statistic (Kulldorff, 1997), which has been used in a variety of ways in the field of epidemiology. A circular window S is fixed on each region in the circular spatial scan statistic. The radius of the circle ranges from zero to a pre-specified maximum distance d or a pre-specified maximum number of regions J to be included in the cluster. The window composed of the $(j-1)$-*th* closest neighbours to region i is defined as $S_{i:j}(j = 1,2,\ldots,J)$. Then, $S_1 = \{S_{i:j}; i = 1,2,\ldots,m; j = 1,2,\ldots,J\}$ denotes the set of all windows the circular spatial scan statistic is to examine. Based on the number of observed and expected cases inside and outside the circle, a likelihood ratio statistic is calculated for each circle. Here, L_0 signifies the likelihood under the null hypothesis, where it is assumed no cluster exists in region i under the null hypothesis. The alternative hypothesis implies there is a cluster in region i based on the j-*th* nearest neighbours, where $L_i(i = 1,2,\ldots,m)$ represents the likelihood under the alternative hypothesis. Then the likelihood ratio statistic is given by

$$\max_i \frac{L_i}{L_0} = \left(\frac{C_i}{E_i}\right)^{C_i} \left(\frac{N - C_i}{N - E_i}\right)^{N - C_i} I(C_i > E_i), \qquad (1)$$

where the observed and expected number of cases within the circle are denoted by C_i and E_i, respectively. Similarly, the observed and expected number of cases outside of the circle are represented by $(N - C_i)$ and $(N - E_i)$, respectively. The indicator function $\mathrm{I}(\mathrm{C_i} > \mathrm{E_i})$ is equal to 1 when $C_i > E_i$ and 0 otherwise. Potential clusters are identified as circles with high likelihood ratios (Kullforff, 1997).

Non commercial use only

The SaTScan (Kulldorff *et al.*, 1998) or FleXScan (Takahashi *et al.*, 2006) software can be used to implement the CSS method. Usually, J is chosen to include 50% of the population at risk in a potential cluster. In this study, $J = 15$ was used as this is the FleXScan default. Our study used aggregate data and therefore, the radius of the circle had to include the region centroid for that region to be considered part of the cluster.

Flexible Spatial Scan Statistic (FSS)

The flexible spatial scan statistic (Tango and Takahashi, 2005) performs in a similar manner to the circular scan statistic, although, the shape of the potential cluster in the FSS method is flexible, while still limited to a small neighbourhood of each region. With the flexible spatial scan statistic, an irregularly shaped window S is placed on each region by connecting the neighbouring regions. For any region i, the set of irregularly shaped windows of length j, containing j attached regions including region i, can vary from 1 to the previously set maximum length of the cluster J. Furthermore, to prevent improbable clusters the joined regions are restricted to the subsets of the regions i and $(J-1)$-*th* closest neighbours of region i. Then, $S_2 = \{S_{i:j(k)}; i = 1,2,\ldots,m; j = 1,2,\ldots,k_{ij}\}$ denotes the set of all windows the flexible spatial scan statistic is to examine. There are J circles examined by the circular spatial scan statistic, whereas the flexible spatial scan statistic observes J circles in addition to all the sets of connected regions whose centroids are found within the J-*th* largest concentric circle. Hence, the size of S_2 is larger than S_1, which is at most mJ. The likelihood ratio test for the FSS test statistic, under the Poisson assumption can be obtained by the equation given in (1), where the circle defined in (1) refers to S_2 instead of S_1. FleXScan software (Takahashi *et al.*, 2006) is used to perform the FSS method, under the default setting $J = 15$. The circles with the highest likelihood ratio values are identified as potential clusters, which is analogous to the circular spatial scan statistic.

Bayesian Disease Mapping (BYM)

First used by Besag *et al.* (1991), the Bayesian approach offers a flexible and robust method for spatial analysis and disease mapping. This method is done via a Bayesian framework using Markov chain Monte Carlo (MCMC). The cases are assumed to follow a Poisson distribution with an area specific parameter $\theta_i E_i$ in the first part of the BYM method

$$C_i \sim Poisson(\theta_i E_i),$$

Non commercial use only

where the observed number of cases in region i is denoted by C_i and expected number of cases in region i is given by E_i. The second part of the model is then expressed as

$$log(\theta_i) = \mu + \eta_i,$$

where the disease ratio (relative risk) in region i is given by θ_i, μ is the overall log of the mean ratio over the region and η_i signifies spatially correlated random effects. The spatial random effects η_i are modeled using a proper conditionally autoregressive (CAR) model. A variety of CAR models may also be used by taking a collection of mutually compatible conditional distributions $p(\eta_i|\eta_{-i}), i = 1,2, \dots, m,$ where $\eta_{-i}: j \neq i, j \in \delta_i$ and the set of neighbours for the i-th region is given by δ_i. The following general model for the spatial random effects η_i is given

$$\eta = (\eta_1, \dots, \eta_m)' \sim N(0, \Sigma_\eta),$$
$$\Sigma_\eta = \sigma_\eta^2 (I - \lambda_n D)^{-1} P^{-1},$$

where P is a $m \times m$ diagonal matrix with elements $P_{ii} = e_i$ with e_i as the number of regions adjacent to region i; D is a $m \times m$ matrix with elements $D_{ij} = 1/e_i$ if region i and j are neighbouring and $D_{ij} = 0$ otherwise; the spatial dispersion parameter is given by σ_η^2; the spatial association is measured by λ_η. Using non-informative priors, the parameters can be predicted in the Bayesian framework (MCMC). The posterior distributions for the parameters in the model are then produced. It is noted that the gamma distribution was used for the inverse of the variance components σ_η^2 with shape and scale parameters 0.001 and a Normal distribution was used with mean 0 and variance 10^6 for the fixed effects. Also, the uniform distribution was used as a prior for λ_η. For the BYM method, a cluster is defined as a region where the estimated disease ratio is significantly larger than two (in terms of their credibility sets). WinBUGS software (Spiegelhalter *et al.*, 2004) was used to apply this method in order to compute the estimated disease ratio values.

Non-commercial use only

References

Besag JE, York JC, Mollìe A, 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). Ann Inst Statist Math 43:1-59.

Kulldorff M, 1997. A spatial scan statistic. Comm Statist: Theor Meth 26:1481-96.

Kulldorff M, Rand K, Gherman G, Williams G, DeFrancesco D, 1998. SaTScan V2.1: Software for the Spatial and Space-time Scan Statistics. Bethesda, MD: National Centre Institute.

Spiegelhalter D, Thomas A, Best N, Lunn D, 2004. WinBUGS version 1.4 User Manual. London, UK: MRC Biostatistics unit, Institute of Public Health.

Takahashi K, Yokoyama T, Tango T, 2006. FleXScan: Software for the Flexible Scan Statistic. Japan: National Institute of Public Health.

Tango T, Takahashi K, 2005. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 4:1-15.

Non commercial use only