A scoping review of spatial cluster analysis techniques for point-event data

Charles E. Fritz[1], Nadine Schuurman[1], Colin Robertson[2], Scott Lear[3]

[1]*Department of Geography, Faculty of Environment, Simon Fraser University, Burnaby, BC, Canada;*
[2]*Department of Geography and Environmental Studies, Wilfrid Laurier University, Waterloo, ON, Canada;*
[3]*Department of Biomedical Physiology and Kinesiology, Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada*

Abstract. Spatial cluster analysis is a uniquely interdisciplinary endeavour, and so it is important to communicate and disseminate ideas, innovations, best practices and challenges across practitioners, applied epidemiology researchers and spatial statisticians. In this research we conducted a scoping review to systematically search peer-reviewed journal databases for research that has employed spatial cluster analysis methods on individual-level, address location, or x and y coordinate derived data. To illustrate the thematic issues raised by our results, methods were tested using a dataset where known clusters existed. Point pattern methods, spatial clustering and cluster detection tests, and a locally weighted spatial regression model were most commonly used for individual-level, address location data (n = 29). The spatial scan statistic was the most popular method for address location data (n = 19). Six themes were identified relating to the application of spatial cluster analysis methods and subsequent analyses, which we recommend researchers to consider; exploratory analysis, visualization, spatial resolution, aetiology, scale and spatial weights. It is our intention that researchers seeking direction for using spatial cluster analysis methods, consider the caveats and strengths of each approach, but also explore the numerous other methods available for this type of analysis. Applied spatial epidemiology researchers and practitioners should give special consideration to applying multiple tests to a dataset. Future research should focus on developing frameworks for selecting appropriate methods and the corresponding spatial weighting schemes.

Keywords: spatial clustering, spatial epidemiology, cluster detection.

Introduction

Epidemiologists are keenly interested in understanding disease patterns in both space and time. John Snow's foundational investigation in 1849 in London, where cholera cases were visually clustered around a water pump suspected as the source of disease (at a time when many believed cholera transmission to be airborne), is now recognised as the beginning of spatial epidemiology (Johnson, 2006). Contemporary methods in spatial epidemiology are more advanced, and the field is growing increasingly multi-disciplinary. Public health, spatial statistics and geographical information systems (GIS) have contributed more recently to spatial epidemiology, creating an emphasis on inter-disciplinary collaboration and knowledge translation

(Moore and Carpenter, 1999; Elliott and Wartenberg, 2004; Wang et al., 2006; Beale et al., 2008). A commonly used methodology, and one that has been greatly enhanced by these linkages, is spatial cluster analysis (Openshaw et al., 1987; Besag and Newell, 1991). Defined by the Center for Disease Control and Prevention (CDC) as "an unusual aggregation, real or perceived, of health events that are grouped together in time and space", a cluster can occur in several health classifications and data types; population-based (e.g. disease rates) (Jacquez and Greiling, 2003), event-based (e.g. point locations) (Schuurman et al., 2009b), field-based (e.g. continuously distributed observations) (Rothman, 1990) or feature-based (e.g. points aggregated to boundaries) (Mostashari et al., 2003). For every data type numerous spatial cluster analysis methods exist and vary broadly with respect to assumptions and interpretation (Moore and Carpenter, 1999; Jacquez et al., 2005; Kulldorff, 2006). It is speculated that since the development of early algorithms, hundreds of new methods, and variants of existing ones have been introduced, providing researchers with more robust statistical and analytical capabilities (Kulldorff, 2006).

Corresponding author:
Charles E. Fritz
Department of Geography, Faculty of Environment
Simon Fraser University, 8888 University Drive
Burnaby, BC, V5A 1S6, Canada
Tel. +1 604 723 7942; Fax +1 778-782 5841
E-mail: charles.e.fritz@gmail.com

Spatial epidemiology is a large tent that encompasses many disciplines (e.g. disease surveillance, public health, veterinary epidemiology and disease mapping), each of which has been separately pursuing research in cluster analysis (Clark and Evans, 1954; Brown, 1982; Anselin, 1988; Gatrell et al., 1996; Getis, 2008). Building and transferring knowledge within spatial epidemiology and across other disciplines is imperative for applied researchers and practitioners to utilize the most recent developments in the field. The scoping review is one vehicle for such an information exchange.

Review papers (Moore and Carpenter, 1999; Chung et al., 2004; Elliott and Wartenberg, 2004; Paez and Scott, 2004) act as "state of the science" reports to highlight innovations and trends of the discipline. In a similar vein, method comparisons (Kulldorff et al., 2003; Ozonoff et al., 2005; Aamodt et al., 2006; Duczmal et al., 2011; Yao et al., 2011), and simulations highlight parameterization caveats (Sadahiro, 2003; Costa and Assunção, 2005), statistical power (Kulldorff et al., 2003), and practical issues and are beneficial for promoting knowledge transfer among users and developers. Though new methods are frequently tested, developed and released as packages or standalone applications, a remaining limitation is the lack of methods available through graphical user interface-based applications and accompanying documentation for them. Moreover, implementation outside of graphical user interface applications requires experience with advanced statistical programming tools. As most emerging methods are simulated using synthetic data (Wheeler, 2007; Meliker et al., 2009), determining the efficacy of methods when tested against real-data is less certain, and performance measures and implementation issues borne from the uncertainty and variation of real-data is not frequently readily assessed (Ozonoff et al., 2005; Meliker et al., 2009). Point-event data are also referred to as spatial point process data, for which many methods exist (Getis and Franklin, 2010); however few see use in epidemiological analysis since aggregated health data is more accessible to researchers (i.e. county level US states or Health Regions in Canada). Because there are relatively few examples of point-event methods used in applied spatial epidemiological studies when compared to studies using methods based on aggregate data, it is important for researchers to know how they work and understand issues that may arise, in order to effectively evaluate and optimize method selection, parameterization, and interpretation when such datasets are available (e.g., animal health surveillance data). This paper aims to review research that has used

spatial cluster analysis methods on individual-level, address location data, and highlight important issues for applied spatial epidemiology researchers and practitioners to consider when using this type of analysis.

Scoping reviews are a useful methodology for exploring a question or topic where little knowledge is currently established, highlighting research gaps and potential avenues for future studies (Arksey and O'Malley, 2005; Levac et al., 2010). As opposed to systematic reviews where quantitative analyses may be employed to glean trends in literature, scoping reviews assess the qualitative content of literature through concept and thematic mapping (Levac et al., 2010). Our objective was to review all published literature that utilized spatial clustering techniques for point-event data. An ancillary goal was to call attention to the basis for spatial cluster analysis method selection and the issues therein by proposing several key themes recurrent throughout the papers and also illustrate the application of a selection of methods using real data (e.g. non synthetic).

Materials and methods

Selection of search terms and papers

Studies that fulfilled the following criteria were included in the review: the use of at least one spatial cluster analysis technique that analysed individual-level, address location data, or data derived from real-world geographical coordinates. Operationalization of the term *spatial cluster analysis* is imbued with ambiguity and results would have been superfluous if specific definition of the term was not used in the selection criteria. By constraining our search to this term, we aimed to select all papers used in applied spatial epidemiology since 2000. We included all papers that employed the use of spatial statistical and geostatistical techniques. Local and global methods were included, as both classes of methods analyse the data at the individual level, and only differ in the scale at which they are evaluated. Methods were then further categorised based on the type of data: areal, point and line. All studies that used point features to represent individual level, address location data in their analysis were included. Fig. 1 illustrates the process of the scoping review.

Our review was constrained to papers that applied a spatial clustering method to real data as our focus was on highlighting practical issues and impediments. Methodology papers were used as a means to develop a background for comparison of each study and to generate a table illustrating common analysis themes.
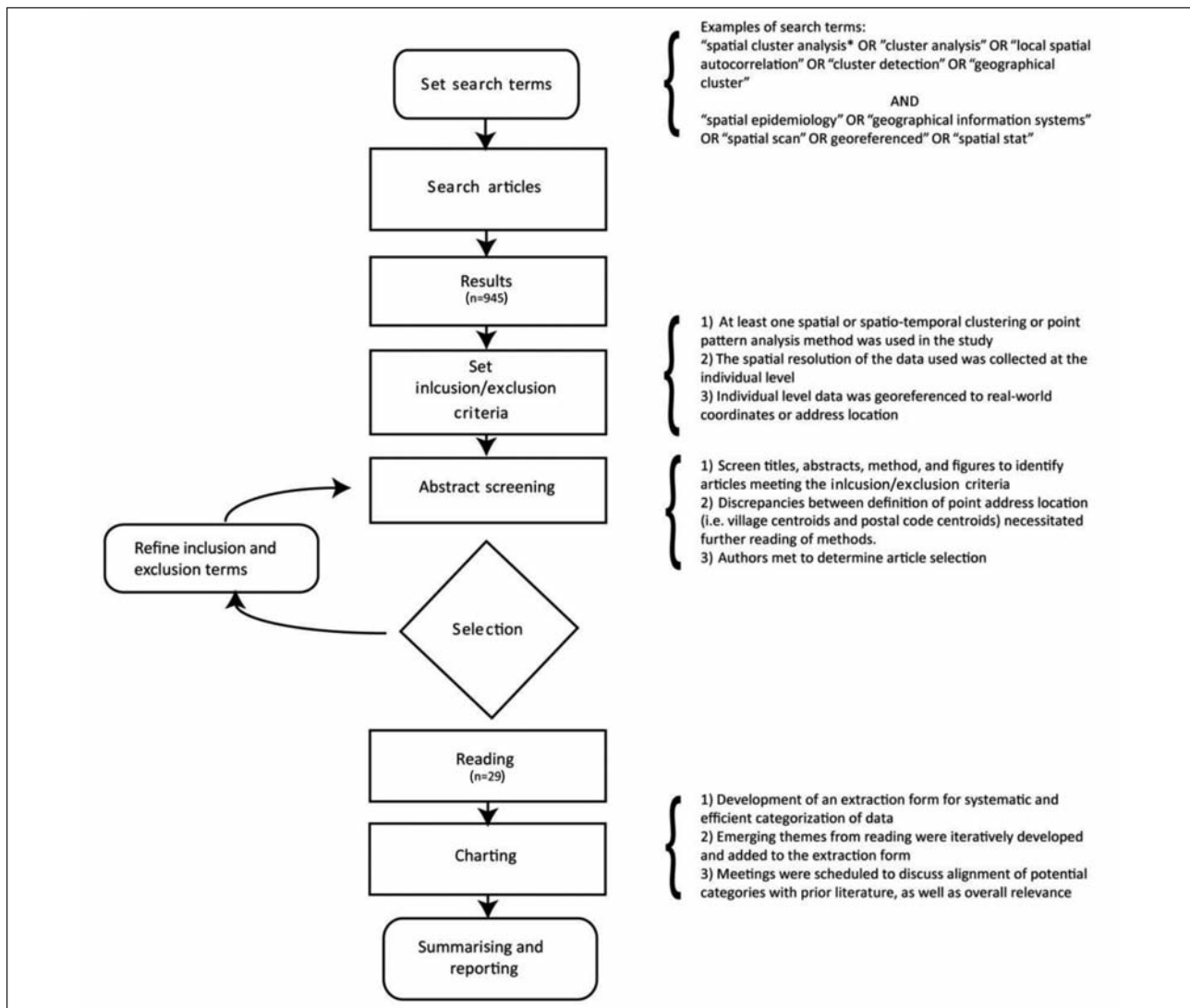
Fig. 1. Overview of scoping review. Flow chart illustrating scoping review process.

Search terms were extracted from a range of academic papers that employed clustering methods in order to ensure coverage of diverse disciplinary language regarding methodology. Our temporal window for acceptance was set from the year 2000 to present.

Once databases were selected (Medline, Web of Science, Science Direct, Academic Search Premier, Jstor, Criminal Justing Abstracts, Global Health, BIOMED Central, CINAHL, TOXNET and Environment Complete), we started our searching process beginning with the formulation of search terms. Final results were compiled and independently reviewed by reading abstracts and certain sections of the articles (i.e. methods, introduction, results and figures) that revealed relevant details regarding spatial data resolution. Articles were then selected based on a set of inclusion and exclusion criteria. Systematic screening of the content was performed; we determined which articles to include in

the charting and collating stage. Themes were iteratively identified and extracted based on author's knowledge and published go-to papers (i.e. review or "state of the science" papers). Consensus was reached for inclusion of themes over several meetings and subsequent content was collaboratively generated. Our goal in identifying themes was to summarise overall patterns of each paper's method implementation aspects. The identification of themes was also guided by the authors expert knowledge and key review papers. A detailed description of stages can be found in Fig. 1.

Method testing

Empirical spatial cluster analysis methods identified in the scoping review process were run on real data to supplement the thematic discussion produced by the scoping review and to provide a visual example of some

of their strengths and/or limitations. Methods were not tested for power to detect accurate clusters or for sensitivity. Our dataset was derived from a database of address level data of severe injuries for the province of British Columbia for the years 2001-2006 called the British Columbia Trauma Registry (BCTR). Our study area was limited to Metro Vancouver, British Columbia, Canada so as to restrict the included population to predominantly urban areas. All incidents in the study area were geocoded with 95% accuracy. The BCTR codes all severe injuries using the ICD-10 classification system, categorising injuries based on the nature of the injury (i.e. auto collision, pedestrian). Previous work by Schuurman et al. (2009a) detected significant clusters of severe pedestrian injury in Metropolitan Vancouver. Our exposure variable for this demonstration was also limited to severe pedestrian injury. For each method, analysis using pedestrian injury data was run where applicable, allowing comparison amongst multiple methods using the same dataset. All local and global methods were compared separately.

Where controls were required, they were estimated by randomly sampling all intersections and street segment midpoints. We do not recommend the use of this method for control sampling in pedestrian injury studies, and chose it based on convenience while considering the focus of this paper.

Results

In this scoping review our initial search returned 945 papers. After setting the inclusion and exclusion criteria and screening each abstract and method section of the papers for key words, we extracted 29 papers from our initial search. We found that point pattern methods, spatial clustering and cluster detection tests, and a locally weighted spatial regression model were most commonly used for individual level, address location data (n = 29). Table 1 provides a summary of each method outlining the application details associated with each method along with software, disciplines utilizing the method, and the relevant article citations from our search.

K-function

The K-function (Ripley, 1976) was the most used *global* clustering method (n = 9). The primary use of K-function analysis was exploring the presence and scale of spatial clustering of the selected exposure variables (Austin et al., 2005; Hillier et al., 2009; Day and Pearce, 2011). The K-function was also used to assess

the spatial structure of a distribution before conducting *local* analyses of spatial clustering (Han et al., 2004; Broman et al., 2006; Wheeler, 2007; Epp et al., 2010; Ngowi et al., 2010; Poljak et al., 2010). Knowing the scale and structure of the spatial dependency among data helps the user confirm whether local analyses are required as well as provide an approximation of spatial weight specifications. Among the reviewed papers two variations of the method were used; univariate K-function (Gatrell et al., 1996) and K-function difference (Cuthbert and Anderson, 2002). The univariate K-function method is best suited for case-event data, and the K-function difference, or bivariate K-function, is best suited for case-control data.

Our illustration of both K-function methods shows that both datasets are clustered. Fig. 2 illustrates the results of K-function methods applied to the BC injury data. Highlights of the outputs are the differences between results when comparing homogenous to inhomogeneous univariate k-function, and the graph as a visual utility to describe the spatial structure of the dataset. Analyses were done using the splancs (Rowlingson and Diggle, 1993) and spatstat (Baddeley and Turner, 2005) libraries of R statistical programming software(R Development Core Team, 2012).

Nearest neighbour statistics

Overall, nearest neighbour-based methods - nearest neighbour index (NNI) (Clark and Evans, 1954), nearest neighbour hierarchical (NnH) (Levine, 2006) and Cuzick Edwards test (Cuzick and Edwards, 1990) - were the second most common class of global methods used in the papers reviewed. Nearly all papers were published from the health-related research disciplines (Andrade et al., 2004; Wheeler, 2007; Pasma, 2008; Lai et al., 2009; Meliker et al., 2009; Epp et al., 2010). Papers that used case-event data utilized the NNI) and NnH methods. Papers that used case-control data utilized the Cuzick Edwards test (Wheeler, 2007; Pasma, 2008; Meliker et al., 2009; Epp et al., 2010). Since they are global methods, the tests do not identify locations of clustering, rather the potential scales at which the distribution may exhibit dependence or association. Overall, nearest neighbour-based methods provide a similar function to the K-function in the way of a global analysis, but differs based on the definition of spatial neighbours and scale, i.e. spatial weights.

Our test of the Cuzick Edwards method indicates significant global clustering for all levels of K nearest neighbours (k-NN) (Table 2). Simulations

Table 1. Reviewed methods. Table summarises the methods used by papers in the review. Other information on methods include data type and resolution required for analysis, software available for implementing the method and disciplines that utilized the method.

Method	Data type	Spatial resolution	Software	Discipline using method	References*
Global methods					
K-function[T]	Case event and case-control	Point	R[a], ArcGIS, MATLAB, ClusterSeer	Environmental public health; veterinary epidemiology; cancer epidemiology; disease surveillance	28, 2, 7, 12, 4, 8, 11, 16, 19
Cuzick-Edwards test[dT]	Case-control	Point or areal	Space Time Intelligence System (STIS), ClusterSeer, R[a]	Cancer epidemiology; veterinary epidemiology;	15, 28, 8, 17
Nearest neighbour index	Case event	Point	ClusterSeer, Crimestat[a]	Veterinary epidemiology; injury prevention; population health surveillance	15, 17
Nearest neighbour hierachical	Case event	Point	Crimestat	Injury prevention; population health surveillance	14, 1
Kernel density estimation[T]	Case event and case-control	Point	Crimestat[a], ArcGIS	Criminology; population health surveillance	1, 22
Local methods					
Kernel intensity function[dT]	Case event and case-control	Point	R[a]	Cancer epidemiology	28
Anselin's local Moran's *I*	Continuous	Point or areal	GeoDa[a], ArcGIS, SpaceStat, R[a]	Environmental science; transportation studies; epidemiology	10
Generalised additive model[T]	Continuous	Point	R[a], SPSS, S-PLUS	Veterinary epidemiology; environmental public health; cancer epidemiology	19, 21, 24, 25, 23
Spatial scan statistic[T]	Case-control, case-event and continuous	Point or areal	R[b], SatScan[a], ClusterSeer[b]	Veterinary epidemiology; surveillance[c]; injury prevention; epidemiology[c]	15, 28, 15, 11, 16, 19, 17, 1, 3, 5, 6, 9, 18, 20, 22, 26, 27, 29, 13

[a]Free software; [b]limited parameters and models; [c]refers to many types of discipline; [d]global and local; N = strengths and limitations are derived from the articles [T]Tested method.
*1) Andrade et al., 2004; 2) Austin et al., 2005; 3) Bautista et al., 2006; 4) Broman et al., 2006; 5) Brooker et al., 2004; 6) Chaix et al., 2006; 7) Day and Pearce, 2011; 8) Epp et al., 2010; 9) Ernst et al., 2006; 10) Gruebner et al., 2011; 11) Han et al., 2004; 12) Hillier et al., 2009; 13) Huang et al., 2009; 14) Lai et al., 2009; 15) Meliker et al., 2009; 16) Ngowi et al., 2010; 17) Pasma, 2008; 18) Polack et al., 2005; 19) Poljak et al., 2010; 20) Sarkar et al., 2007; 21) Siqueira-Junior JB et al., 2008; 22) Tanser et al., 2009; 23) Vieira et al., 2009; 24) Vieira et al., 2010; 25) Vieira et al., 2008; 26) Warden, 2008; 27) Westercamp et al., 2010; 28) Wheeler, 2007; 29) Winskill et al., 2011.

(n = 999) were used for Monte Carlo significance testing. Compared to the K-function, Cuzick-Edwards defines spatial relationships in terms of nearest neighbours and not distance. This method offers an alternative exploratory approach for global clustering. Analysis was done in Clusterseer 2.3 (www.biomedware.com).

Local Moran's I

One paper utilized the local Moran's *I* (LMI) (Anselin, 1995) statistic (Gruebner et al., 2011).

Gruebner et al. (2011) applied a suite of Moran's *I* statistics (global univariate/bivariate Moran's *I* and local univariate/bivariate Moran's *I*) to explore self-rated WHO-5 mental health and health determining factor scores at the household level in Dhaka slums. Among the numerous variables tested, both global and local spatial autocorrelation was evident, with the intensity of the dependency decreasing as the amount of nearest neighbours increased. LMI is a powerful tool for detecting both spatial clusters and spatial clustering. Due to the type of data used for our method testing (case-control), we excluded LMI from that portion.
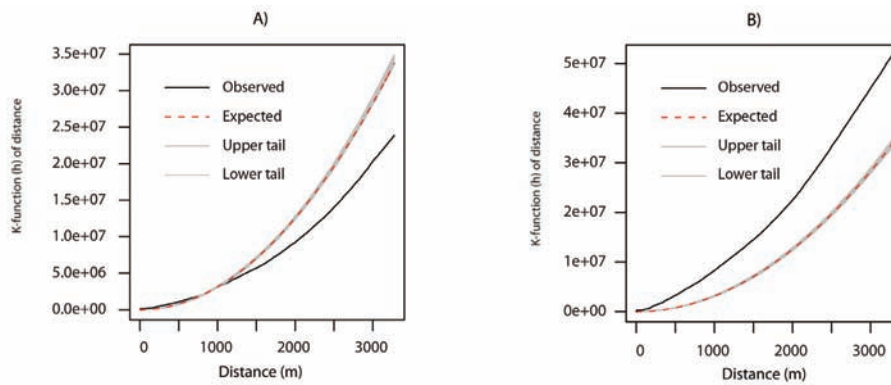
Fig. 2. K-function global clustering results-not that this has been changed. Image illustrates the use of the K-function as a multiscale global spatial clustering tool where the observed curve is above the theoretical, clustering is apparent at the corresponding distances. For significance testing the inhomogeneous k-function (A) assumes a non-stationary point process, whereas the ordinary k-function (B) assumes a stationary point process. For a more intuitive value for K, the transformation of the values to an L function is widely applied. Simulations (n=99) were used for significance testing.

Kernel estimation

A predominate use of kernel estimation approaches was for visual exploration of a dataset (Andrade et al., 2004; Wheeler, 2007). The visual analysis provided through the KDE (Silverman, 1986) method makes reference and communication of results intuitive (Andrade et al., 2004). Outputs from the Kernel density estimation (KDE) method also provide evidence of

visual "hotspots" to the researcher for subsequent analyses (Lu, 2006). Similar benefits can be realized from the Kernel intensity function (KIF) approach (Kelsall and Diggle, 1995). Visual outputs communicate log relative risk ratios of cases and control data (Wheeler, 2007).

Our tests using these two methods (Fig. 3) experimented with multiple bandwidth parameter settings; however for space reasons we chose to only include a
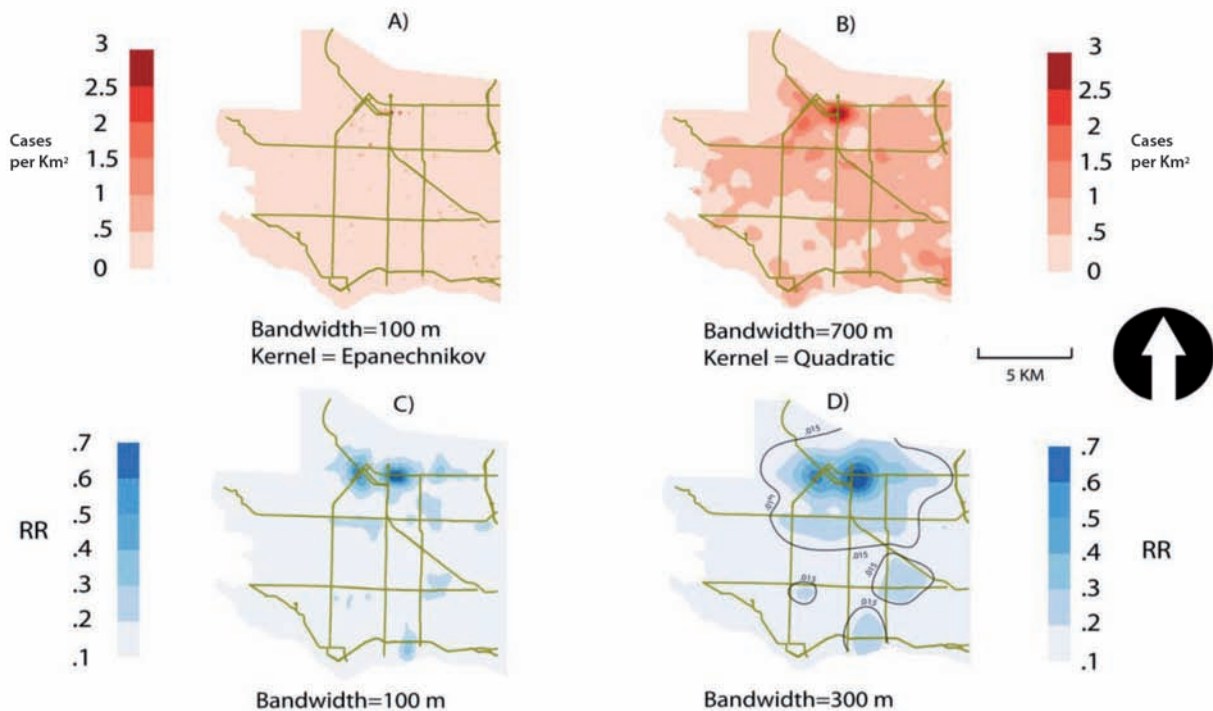

Fig. 3. Kernel methods results. Kernel density estimation (KDE) and kernel intensity function (KIF) surfaces generated using varying bandwidths (both). Representation illustrates the flexibility of using the KDE and KIF approaches serving as both visual and statistical tools. (A) KDE; (B) KDE; (C) KIF; (D) KIF.

Table 2. Cuzick-Edwards global clustering method results. The first column (k) indicates the amount of spatial neighbours used. The test statistic is denoted by T(k), and describes the amount of neighbouring cases from k. The expected value of the test statistic T(k) is denoted as E(Tk). The variance of the test statistic around the mean value is denoted as Var(T). Bonferroni and Simes corrections are applied for multiple testing. The null hypothesis of complete spatial randomness was rejected at all levels of k-nn.

k-nn	T(k)	E(k)	Var(T)	z	Monte Carlo P-value
1	357	148.812	157.013	16.6145	0.001
2	654	297.625	326.757	19.7149	0.001
3	937	446.437	500.548	21.9266	0.001
4	1,200	595.25	677.412	23.2354	0.001
5	1,463	744.062	859.402	24.5241	0.001

Bonferroni P-value: 0.005
Simes P-value: 0.001

few different settings. In general, results for KDE and KIF approaches are similar; both readily identified where elevated intensities of pedestrian injury were. A minor difference in outputs is seen in the KIF method however, as the underlying population distribution is included (controls), depicting a slightly different risk surface than the KDE approach. Statistical analyses were carried out using the splancs (Rowlingson and Diggle, 1993) and spatialkernel (Zheng and Diggle, 2009) libraries of the R statistical programming software (R Development Core Team, 2012) All images were imported to ArcGIS 10 (ESRI, 2009) for representation and layer overlay operations. A slight variant of the KIF approach, the spatial relative risk function, is also accessible through the R package Sparr (Davies et al., 2011).

Generalized additive model

Generalized additive models (GAM) are a type of generalized linear model (GLM) (Hastie and Tibshirani, 1987; Kelsall and Diggle, 2002) that extend GLMs by adding a smoothing function to account for geographical space. GAMs have most recently been used in spatial epidemiology and disease risk mapping (Ozonoff et al., 2005). Of the 29 papers reviewed, five used GAMs. All five papers utilized the GAM approach for exploratory purposes (Siqueira-Junior et al., 2008; Vieira et al., 2008, 2009, 2010; Poljak et al., 2010). A primary goal in epidemiology is the explanation of processes generating spatial and temporal patterns of disease and disease risks. GAMs can be applied in a space only, spatial-temporal, or time only frameworks (Vieira et al., 2009; Poljak et al., 2010) and have been cited to be particularly useful for analysing longitudinal data that incorporate residential history patterns (Vieira et al., 2010). The spatial output of the variety of analyses are communicated through a smoothed risk surface, aiding visual recognition of patterns, a technique often used as an exploratory spatial data analysis (Poljak et al., 2010). The prime advantage of GAMs is the ability to control for the underlying population distribution from spatial control locations – similar to the KIF – as well as covariates.

Illustration of the GAM yielded a similar visual display as the KIF and KDE, mainly due to smoothing parameters used in the method (Fig. 4). The smoother chosen was the locally weighted scatter plot smoother (LOESS), and was applied to the x and y values of each case and control location. Akaike information
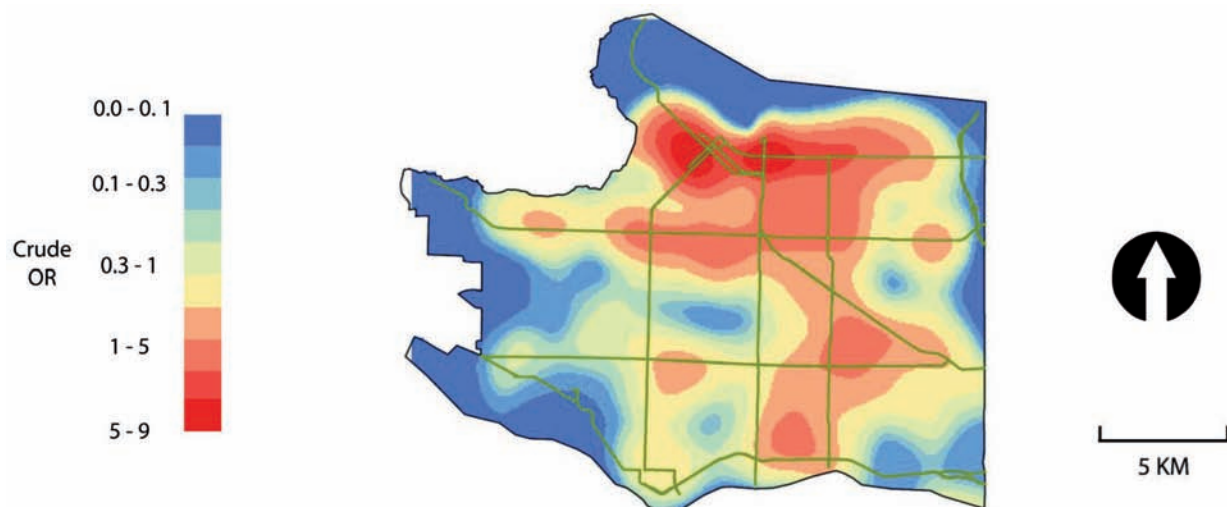

Fig. 4. Generalized additive model surface generated using optimal span size of .50. Statistically significant (P = 0.05) hotspots (high odds ratios) are coloured in red, and coldspots are in blue.

criterion (AIC) was used to determine the optimal span size of .50. Only crude odds ratios were calculated as covariates were not available. Statistical analysis was carried out using the gam (Hastie, 2011) library of R statistical programming software (R Development Core Team, 2012).

Spatial scan statistic

Over half of papers in this review applied the spatial scan statistic to examine the spatial patterns of address location data (Andrade et al., 2004; Brooker et al., 2004; Han et al., 2004; Polack et al., 2005; Bautista et al., 2006; Chaix et al., 2006; Ernst et al., 2006; Pollack et al., 2006; Sarkar et al., 2007; Wheeler, 2007; Pasma, 2008; Warden, 2008; Huang et al., 2009; Meliker et al., 2009; Tanser et al., 2009; Epp et al., 2010; Ngowi et al., 2010; Poljak et al., 2010; Westercamp et al., 2010; Winskill et al., 2011). Of the reviewed articles, 83% applied a Bernoulli model spatial scan statistic to case-control data; two of those articles also used other models in SatScan that can be applied to address location data, the discrete normal continuous model (Huang et al., 2009) and the discrete poisson continuous model (Ngowi et al., 2010), ordinal model (Westercamp et al., 2010) and the multinomial model (Westercamp et al., 2010). A clear

categorical distinction between applications was the scale of the study area in which the spatial scan statistic was used; large-scale (state, regional and metropolitan areas) versus small-scale (small areas, neighbourhoods and villages).

For different scales, the spatial scan statistic allows the user to adjust this setting based on the minimum or maximum percent of the population to include in relative risk ratio calculations; radius of the scanning window; and proximity from the center of the circle. Other papers have examined the difference between window settings (Chen et al., 2008; Jackson et al., 2009), but none have addressed the ability for the spatial scan statistic to detect clusters of the same phenomena at different scales.

Analysis of the injury data using the spatial scan statistic returned several significant clusters, with the most significant located in the same general region as previous tests of KDE, KIF and GAM (Figs. 5 and 6). For this test a similar approach to the KDE and KIF methods was used; exploring the results using varying definitions for maximum cluster size. We defined maximum cluster size in two ways; maximum percentage of population in scanning window and maximum scanning window radius. Cluster maps were imported to ArcGIS for overlay with streets and study area boundary.
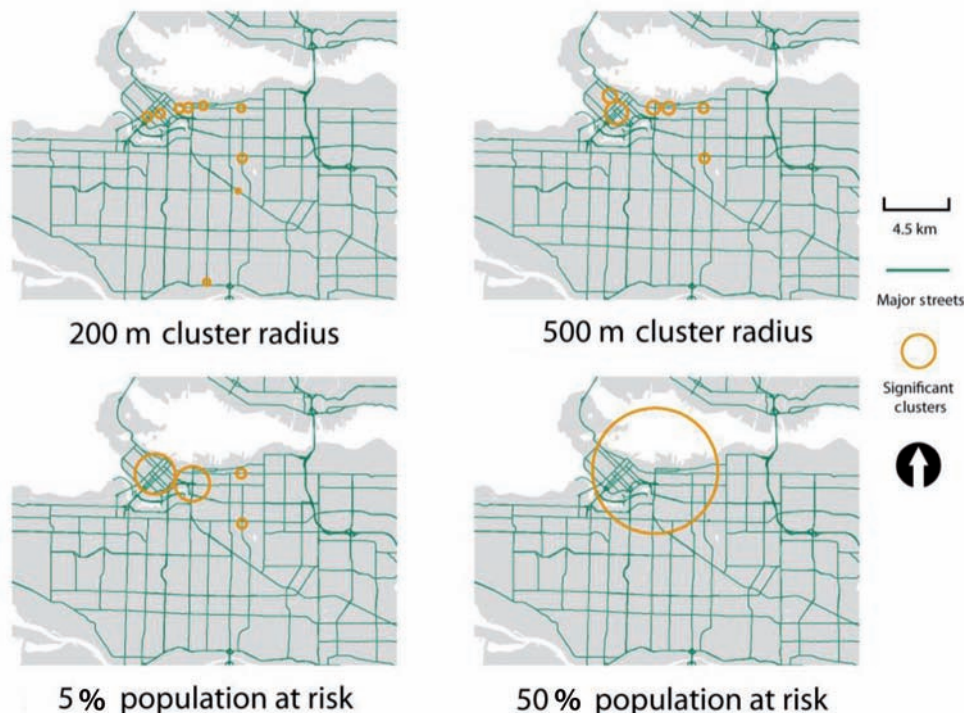

Fig. 5. Spatial scan statistic surface generated using different maximum cluster size definitions based on maximum percentage of population at risk (top two images) and maximum distance of scanning window radius (bottom two images) assuming a Bernoulli probability model. Significance of pedestrian injury cluster tested at P = 0.05 and indicated with orange circles.
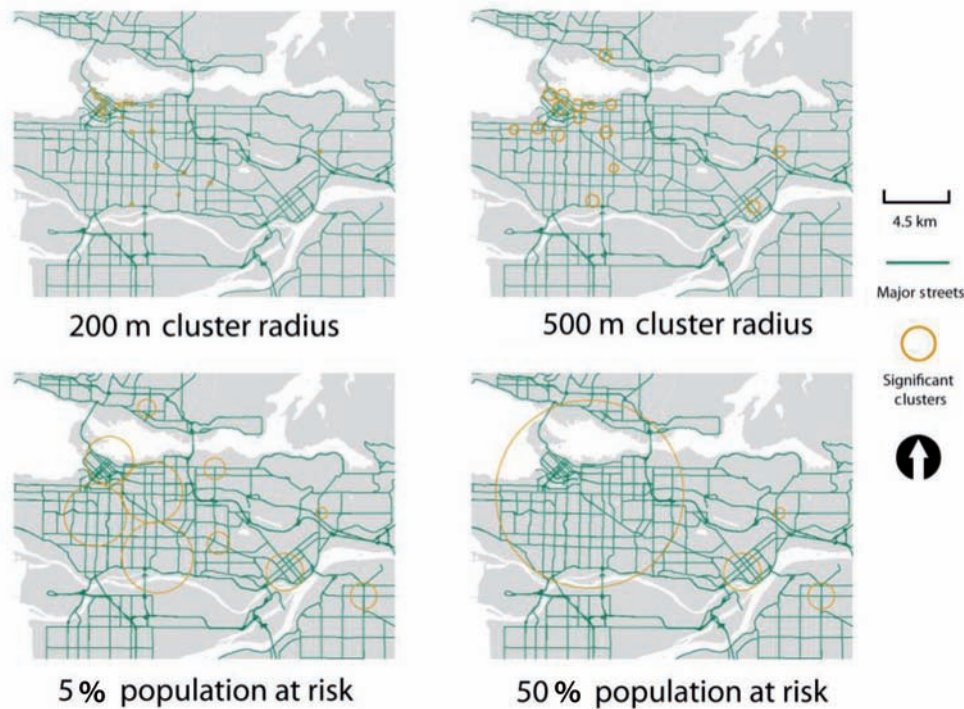
Fig. 6. Spatial scan statistic surface generated using different spatial weight conceptualizations based on maximum percentage of population at risk and maximum distance of radius assuming a Bernoulli probability model. Significance of pedestrian injury cluster is tested at P = 0.05 and indicated with orange circles. Representation illustrates how scale of study area can affect spatial cluster analysis results, suggesting that spatial window settings should be tailored to the unique study objective in order to retain the method's efficacy.

Themes

Our results summarise all peer-reviewed journal articles from the year 2000 to 2010 that met our inclusion criteria of using the term "spatial cluster analysis" to examine the spatial patterning of point-event data. Application of spatial cluster analysis to point-event data is mainly carried out through means of exploratory methods, emphasising the power of visualization. Authors seem to be aware of data resolution issues, but we maintain that consideration of other terms such as exploratory analysis, visualization, aetiology, scale, spatial weights and method selection should also be considered. These themes are not to be taken as strict guidelines for conducting spatial cluster analysis, but rather, they are recommendations from the authors' knowledge informed from key review papers and expert knowledge. Themes are summarised in Table 3.

We categorised the themes into two divisions that may affect the way methods are employed by the researcher, including (i) methodological focus and (ii) data. Methodological focus refers to the reasons for undertaking research using a particular method. Data

refers to how the actual dataset may dictate the way a method is selected, or the results of subsequent analyses. Several papers utilized multiple methods to investigate spatial phenomena with a closer lens (n = 24), what we are terming as an *exploratory analysis*. Application of a variety of methods is not too dissimilar from an exploratory spatial data analysis approach (ESDA), wherein the researcher applies multiple spatial analysis techniques to glean spatial patterns from the dataset, and identify associations to be later incorporated into a model, or confirmatory analysis (Haining et al., 1998). A caution to this approach is the idea that "data snooping" or "data dredging" of the dataset via use of multiple methods or testing may lead to spurious post-hoc conclusions about underlying processes (Selvin and Stuart, 1966; Sullivan et al., 1999). There is a tradeoff between exploratory and confirmatory approaches that can be mediated by the intended objective of the analysis. Where the objective is to delineate spatial clusters to guide further study or generate hypotheses of risk factors, multiple methods may reinforce findings and provide confidence that the located areas are in fact "unusual". However, if the object of analysis includes

Table 3. Identified themes. Table provides an overview of themes identified from each paper. Each theme can also be interpreted as issues to consider before choosing a spatial clustering test, as well the subsequent synthesis of results.

Methodological focus	
Exploratory analysis	Several papers utilized multiple methods to investigate the spatial phenomena with a closer lens. Adopting an ESDA provides more in-depth analysis because multiple spatial cluster analysis methods are generally adopted. Being able to compare the dataset among various spatial methods enhances the researchers understanding of the data, better informing their inference of patterns that may arise in the dataset.
Visualization	Spatial cluster analysis methods that incorporate visualization in their outputs are advantageous in research and practice settings. Kernel density estimation, kernel intensity function and generalized additive models were adept for achieving this objective. Visualization is also an important step in an ESDA process.
Data	
Spatial resolution	Spatial resolution is an important component in the process of selecting methods for analysis and has positive and negative implications for subsequent analysis. Acquisition of individual-level, point-event data is absolutely necessary to prevent obfuscation of local spatial clustering. MAUP and EF are two common spatial data issues that may arise due to coarseness in the dataset.
Aetiology	Various data types and spectrums of spatial resolution can reflect the known aetiology of the studied phenomena. In some situations a single address level measurement (e.g. home address) may not be an accurate surrogate for exposure. Residential histories that include data on amount of times moved, and years spent at specific locations have proved an adequate surrogate for latency in disease.
Scale	An overarching goal of spatial cluster analysis is to understand the spatial structure, or scale of the studied processes. Equally important is determining optimal scale to analyse the data, or study area boundary extent, as it has been suggested in the literature that variable units of space can yield different results.
Spatial weights	Varying conceptualizations of space will yield different results for spatial clustering and cluster detection. Likewise, each spatial clustering method will define space with different parameters. For most methods, selecting an appropriate spatial weight conceptualization remains a decision based on researcher discretion and should be heavily considered when synthesising results.

relationships among variables, for example in the context of fitting a spatial point-process model, a more restrained approach is recommended. The tendency of papers to use multiple methods suggests that the methodological focus of research using individual level, point-event data and moreover spatial cluster analysis, is closely aligned with ESDA. Fig. 7 illustrates the use of multiple methods on the same dataset with an emphasis on exploring the dataset. Moreover, *visualization*, an ancillary utility of ESDA, was also a commonly used analytical support tool or supplement to research (n = 29). Among the methods recorded in the review, KDE, KIF and GAM approaches produced flexible and intuitive visual outputs, advantageous for knowledge transfer in research and practice settings. Fig. 7 provides an example of the tested methods visualization outputs.

Spatial data resolution, *aetiology*, *scale*, *spatial weights* and *method selection* are themes that fall in the data category. *Spatial resolution* of data largely governs which methods can be used for analysis. Because spatial health data is normally aggregated to

protect individual confidentiality, it restricts the selection of methods to those that handle data of coarser spatial resolution. It has been shown elsewhere that aggregated data can obfuscate local heterogeneity (Ozonoff et al., 2007; Meliker et al., 2009; Meliker and Sloan, 2011) caused by known spatial data issues, such as ecological fallacy (EF) (Openshaw, 1984) and modifiable areal unit problem (MAUP) (Fotheringham and Wong, 1991). *Aetiology* of a disease process or health event is related to the spatial data resolution as well, but barely acknowledged in the papers reviewed. Like spatial data resolution, aetiology can affect which methods are chosen for subsequent analysis. It was suggested among some of the reviewed papers that an individuals' home address may not be an accurate surrogate measure for some exposures (Huang et al., 2009; Vieira et al., 2009), and that census tracts or dissemination areas may reflect the level of exposure more accurately. Incorporating a sense of an individual's activity space (Orellana and Wachowicz, 2011; Zenk et al., 2011), mobility (Signorino et al., 2011) or latency period (Vieira et al., 2010) to an exposure may
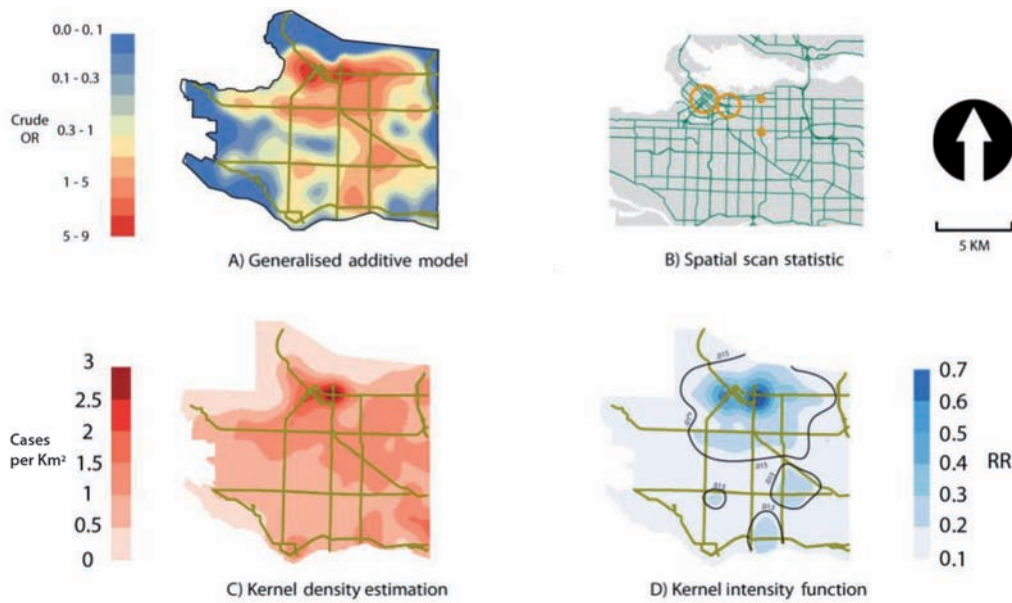
Fig. 7. Exploratory analysis example. Representation compares all visualization-capable methods applied to the same dataset. A) Statistically significant (P = 0.05) hotspots (high odds ratios) are coloured in red, whereas coldspots are in blue. B) Test assumes a Bernoulli probability model with a scanning window of 5% of the population at risk. Significant clusters (P = 0.05) highlighted with orange circles. C) Gaussian kernel with a bandwidth of 300, applied to cases only. D) Bandwidth of 500 applied to cases and controls. Statistically significant regions (P = 0.015) outlined by black contour line.

paint a clearer picture of exposure as well. Similarly influential to spatial cluster analysis is the theme *scale*. Papers in this review were split in terms of the scale of study area for analysis; large scale (counties, metropolitan areas and provinces) and small scale (local neighbourhood areas). It has been suggested that a study area's boundary extent can affect the likelihood for detecting a true cluster (Jacquez and Greiling, 2003; Wheeler, 2007). Papers in our review illustrate this effect, which detected clusters that span several km in diameter throughout a rural area (Ngowi et al., 2010), or multiple contiguous neighbourhoods throughout a metropolitan area (Chaix et al., 2006). Our method test with pedestrian injury illustrated this particular effect when applying the spatial scan statistic to the same dataset only changed by an increased sample size and study area size (Fig. 6). *Spatial weights* refer to the measure of spatial neighbourhood relationships applied to spatial cluster analysis methods. In our review the most common issue with regard to spatial weights concerned the spatial scan statistic and its inability to detect irregularly shaped clusters (Brooker et al., 2004; Bautista et al., 2006; Chaix et al., 2006; Wheeler, 2007; Huang et al., 2009; Tanser et al., 2009). Since early developments of the spatial scan statistic options have been added to the software that deal with this issue (Kulldorff et al., 2006) and other scanning window methods have been developed

(Tango and Takahashi, 2005; Duczmal et al., 2011), however no reviewed papers applied them. Other methods such as KIF and GAM allow for flexible spatial conceptualizations of neighbourhood relationships through the application of smoothers (e.g. bivariate LOESS and spatial adaptive filtering), and bandwidth optimizations based on kernel functions. With regard to the spatial scan statistic, a fixed or variable scanning window setting is set *a priori* and its effect on cluster detection outcomes can be seen in our tests (Fig. 5).

Discussion

Our scoping review on spatial cluster analysis methods for individual level, address location data revealed that there has been an increased use of these methods among a range of research disciplines in the last decade. Based on our initial search for academic papers that fit our broad search criteria, a return of 945 papers used methods related to spatial cluster analysis. Without analysing the disciplines of our initial results, we may be able broadly assume that the final selection of 29 papers is somewhat reflective of the distribution, differing only on spatial data resolution (i.e. boundary *versus* address location). In a recent review of spatial analysis methods in spatial epidemiology (Auchincloss, 2012), the authors returned

an initial total – after applying broad search terms – of 5,641 papers, and eventually accepted 206. Their analysis drew from a variety of health disciplines and broadly surveyed all spatial analysis methods that were applied to problems in the respected area. While neither this paper nor the aforementioned provide a clear explanation for the poor ratio between papers searched and papers accepted, specific to our research objectives, we speculate this arises from data availability or inadequate methods to analyse individual level point-event data. Nonetheless the adoption of spatial cluster analysis methods within, and beyond health-related disciplines, continues to be a burgeoning trend (Costa and Kulldorff, 2009).

Since a 1999 review of spatial analysis methods in spatial epidemiology (Moore and Carpenter, 1999), prevailing methods NNI, Cuzick-Edwards, K-function and spatial scan statistic remain among the most prominent methods for conducting spatial cluster analysis of point-event data. In particular, our review found that the spatial scan statistic was utilized for spatial cluster detection in 19 of the 29 reviewed papers. Much has been discussed about the flexibility of the spatial scan statistic (Kulldorff, 1999; Kulldorff et al., 2006, 2009; Costa and Kulldorff, 2009), however, there has been equal amount of research highlighting some weaknesses of the method (Tango and Takahashi, 2005; Neill, 2009; Cançado et al., 2010). Recent uses in spatial clustering methods, GAMs (Vieira et al., 2009) and KIFs (Wheeler, 2007), illustrate the advantages of generating a smooth risk surface coupled with an intuitive visual output. While there has been an increase in the type of methods being adopted for analysis at the address level, there remain several unused methods for reasons unbeknownst to the authors (Kulldorff, 2006) (e.g. Turnbull's cluster evaluation permutation procedure, Besag and Newell's R, Tango's S flexibly shaped spatial scan statistic, Duczmal's simulated annealing method and Bayesian local likelihood method). Furthermore, as our review clearly illustrates, a number of themes germane to spatial data analysis are seldom considered, or at least mentioned by the authors in this review, even as review papers and comparison studies routinely identify and highlight issues to consider (Moore and Carpenter, 1999; Kulldorff et al., 2003; Ozonoff et al., 2005, 2007; Beale et al., 2008; Jackson et al., 2009; Meliker and Sloan, 2011). With consideration to these issues we recommend a brief series of research areas for those even tangentially involved in the discipline to ponder.

Future research recommendations

Based on the themes generated from our review, we make a few recommendations for future research in spatial cluster analysis, and spatial epidemiology more broadly. Firstly, with regard to *exploratory analysis,* we recommend that researchers utilizing spatial cluster analysis as an exploratory tool consider using multiple tests to gain a greater understanding of the dataset. Recent approaches proposed in spatial epidemiology (Berke, 2005; Jacquez, 2009) focus on using multiple methods as a means to explore every possible avenue of the data to rule out false positives or spurious clusters. Second, researchers should not discount the utility of visualization as a supplement to analysis and enhancement to communication and dissemination of results. Most methods reviewed in this paper produce visual outputs of results, yet using visualization as an analysis tool is an alternative that has yet to influence academic research (Chen et al., 2008; Grubesic, 2010).

Third, spatial data resolution and process aetiology are two inherently related themes that uniquely impact the design and results of research. It has been suggested in papers from this review that some levels of data resolution are not representative of the aetiology of some processes (Huang et al., 2009; Vieira et al., 2009). To answer this call, work on data collection techniques and spatial cluster analysis methods that incorporate a notion of an individual's activity space and mobility (Orellana and Wachowicz, 2011; Zenk et al., 2011), or residential history (Meliker and Sloan, 2011), should be encouraged throughout the disciplines involved with spatial cluster analysis. Building such an awareness throughout this research community will also bring more macroscopic issues to the front, such as privacy issues in health research caused by the distribution of sensitive, location-specific health data (Boulos et al., 2009; Meliker et al., 2009).

Last, selection of appropriate *spatial weights* and, moreover, *selection of a method* suitable for particular datasets largely impacts the results from tests. As our test illustrated with the spatial scan statistic scanning windows, how the spatial neighbours are conceptualized can dramatically impact the location and extent of clusters. Recent papers have addressed both of these issues (Jacquez, 2009; Meliker and Sloan, 2011), and called for the use of tools to aid users in selecting appropriate methods and spatial weights. We would like to echo their recommendations in light of the large disparity between methods employed and methods available to the user. Providing users with more sup-

port around these two objectives will build a greater demand for the use of spatial cluster analysis methods on individual level, address location data.

Limitations

There are inherent limitations to conducting scoping reviews, as there is a subjective nature of setting search terms, and generating thematic categories (Levac et al., 2010). This is a caveat associated with synthesizing data from various disciplines, and multiple methods of different statistical categories. With regard to search term selection, we attempted to prevent this issue by reading review papers and various methodological articles to extract key language identifiers. We felt that this was sufficient for the task, but also acknowledge there may have been articles excluded because of this (e.g. spatial point process methods and Bayesian disease mapping). It is strongly recommended that any researcher aiming to apply a scoping review methodology perform an exhaustive search for language germane to the respected literature before setting search terms. The low ratio of searched to accepted papers could also be explained by the focus of this paper to accept only papers that used x-y coordinate or address location data. In other words, the initial search may have yielded all results that used spatial clustering methods on all data aggregation scheme thus inflating the initial pool of articles. We tried several combinations of search terms to yield only the address level papers, however the searches resulted in significantly low search returns, often leaving out papers that actually used x-y coordinate or address level data.

A second limitation relates to the decision to exclude methodological papers, such as simulation studies, from the review. We sought to include all papers that have applied methods to real-data so as to extract methodological concerns raised by researchers who do not necessarily develop some of the approaches we reviewed. Not only did this specific procedure allow us to conclude which methods are used most readily, but also provided a rich contrast between issues identified in methodology papers and issues in more practical settings. We acknowledge that by excluding methodology articles we not only left out a large portion of spatial cluster analysis methods, but also more in-depth methodological issues. A few articles that were contributed by methodology developers provided us insight to those deeper issues, complementing other articles without that focus. An interesting future study would be a review based on methodological papers, with the end goal of outlining a framework for apply-

ing spatial clustering techniques to individual-level, address location data.

Conclusions

The study of spatially dependent variables in space has a long history that spans numerous disciplines and so communication and knowledge transfer between those academic communities is assumed to be a major enabling factor. The discipline of spatial epidemiology, in itself, is an immensely interdisciplinary field, unifying researchers in statistics, public health, global health, environmental sciences, geography and parasitology, as a small sampling of disciplines. In order to effectively search for, and select an appropriate method, it is therefore important to understand how data-related factors will govern the caveats and strengths of each respected approach. Our application of a scoping review technique allowed us to catalogue the approaches and summarise the issues associated with them. By compiling the literature that has applied these techniques, our results not only speak to the growth and diversity of disciplines that apply to them, but also highlight the potential to communicate various approaches in spatial cluster analysis. A scoping review methodology presents itself as a useful alternative to systematic reviews, as it strives in identifying broad research gaps and qualitative themes across a narrow subset of a field. It is our intention that researchers seeking direction for using spatial cluster analysis methods, consider the caveats and strengths of each approach, but also explore the numerous other methods available for this type of analysis.

Acknowledgements

References

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

Andrade ALSS, Silva SA, Martelli CMT, Oliveira RM, Morais Neto OL, Siqueira Jr. JB, Melo LK, Di Fábio JL, 2004. Population-based surveillance of pediatric pneumonia: use of spatial analysis in an urban area of Central Brazil. Cad Saude Publica 2, 411-421.

Anselin L, 1988. Spatial econometrics: methods and models. Kluwer Academic Publishers 16, 284 pp.

Anselin L, 1995. Local indicators of spatial association-LISA. Geogr Anal 2, 93-115.

Arksey H, O'Malley L, 2005. Scoping studies: towards a methodological framework. Int J Soc Res Meth 8, 19-32.

Auchincloss AH, 2012. The use of spatial methods in epidemiology. Annu Rev Publ Health 1.

Austin SB, Melly SJ, Sanchez BN, Patel A, Buka S, Gortmaker SL, 2005. Clustering of fast-food restaurants around schools: a novel application of spatial statistics to the study of food environments. Am J Public Health 9, 1575-1581.

Baddeley A, Turner R, 2005. Spatstat: an R package for analyzing spatial point patterns. J Stat Softw 6, 1-42.

Bautista CT, Chan AST, Ryan JR, Calampa C, Roper MH, Hightower AW, Magill AJ, 2006. Epidemiology and spatial analysis of malaria in the Northern Peruvian Amazon. Am J Trop Med Hyg 6, 1216-1222.

Beale L, Abellan JJ, Hodgson S, Jarup L, 2008. Methodologic issues and approaches to spatial epidemiology. Environ Health Persp 8, 1105-1110.

Berke O, 2005. Exploratory spatial relative risk mapping. Prev Vet Med 3-4, 173-182.

Besag J, Newell J, 1991. The detection of clusters in rare diseases. J R Stat Soc Ser A Stat Soc 154, 143-155.

Boulos MNK, Curtis AJ, AbdelMalik P, 2009. Musings on privacy issues in health research involving disaggregate geographic data about individuals. Int J Health Geogr 1, 46.

Broman AT, Shum K, Munoz B, Duncan DD, West SK, 2006. Spatial clustering of ocular chlamydial infection over time following treatment, among households in a village in Tanzania. Invest Opht Vis Sci 1, 99-104.

Brooker S, Clarke S, Njagi JK, Polack S, Mugo B, Estambale B, Muchiri E, Magnussen P, Cox J, 2004. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. Trop Med Int Health 7, 757-766.

Brown MA, 1982. Modelling the spatial distribution of suburban crime. Econ Geogr 3, 247-261.

Cançado ALF, Duarte AR, Duczmal LH, Ferreira SJ, Fonseca CM, Gontijo ECDM, 2010. Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. Int J Health Geogr 9, 55.

Chaix B, Leyland AH, Sabel CE, Chauvin P, Råstam L, Kristersson H, Merlo J, 2006. Spatial clustering of mental disorders and associated characteristics of the neighbourhood context in Malmö, Sweden, in 2001. J Epidemiol Commun Health 5, 427-435.

Chen J, Roth RE, Naito AT, Lengerich EJ, Maceachren AM, 2008. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. Int J Health Geogr 7, 57.

Chung K, Yang DH, Bell R, 2004. Health and GIS: toward spatial statistical analyses. J Med Syst 4, 349-360.

Clark PJ, Evans FC, 1954. Distance to nearest neighbour as a measure of spatial relationships in populations. Ecology 4, 445-453.

Costa MA, Assunção RM, 2005. A fair comparison between the spatial scan and the Besag-Newell disease clustering tests. Environ Ecol Stat 3, 301-319.

Costa MA, Kulldorff M, 2009. Applications of spatial scan statistics: a review of scan statistics, Birkhäuser Boston, 129-152 pp.

Cuthbert AL, Anderson WP, 2002. Using spatial statistics to examine the pattern of urban land development in Halifax-Dartmouth. Prof Geogr 4, 521-532.

Cuzick J, Edwards R, 1990. Spatial clustering for inhomogeneous populations. J Roy Stat Soc B Met 1, 73-104.

Davies TM, Hazelton ML, Marshall JC, 2011. sparr: analyzing spatial relative risk using fixed and adaptive Kernel density estimation in R. J Stat Softw 1, 1-14.

Day PL, Pearce J, 2011. Obesity-promoting food environments and the spatial clustering of food outlets around schools. Am J Prev Med 2, 113-121.

Duczmal L, Moreira G, Burgarelli D, Takahashi R, Magalhaes F, Bodevan E, 2011. Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. Int J Health Geogr 1, 29.

Elliott P, Wartenberg D, 2004. Spatial epidemiology: current approaches and future challenges. Environ Health Persp 9, 998-1006.

Epp T, Argue C, Waldner C, Berke O, 2010. Spatial analysis of an anthrax outbreak in Saskatchewan, 2006. Can Vet J 7, 743-748.

Ernst K, Adoka S, Kowuor D, Wilson M, John C, 2006. Malaria hotspot areas in a highland Kenya site are consistent in epidemic and non-epidemic years and are associated with ecological factors. Malar J 1, 78.

ESRI 2009. ArcGIS: a complete integrated system Redlands, CA, USA, ESRI.

Fotheringham AS, Wong DWS, 1991. The modifiable areal unit problem in multivariate statistical-analysis. Environ Plann A 7, 1025-1044.

Gatrell AC, Bailey TC, Diggle PJ, Rowlingson BS, 1996. Spatial point pattern analysis and its application in geographical epidemiology. T I Brit Geogr 1, 256-274.

Getis A, 2008. A history of the concept of spatial autocorrelation: a geographer's perspective. Geogr Anal 3, 297-309.

Getis A, Franklin J, 2010. Second-order neighbourhood analysis of mapped point patterns. Riley, Blackwell, 93-100 pp.

Grubesic TH, 2010. Sex offender clusters. Appl Geogr 1, 2-18.

Gruebner O, Khan MMH, Lautenbach S, Muller D, Kraemer A, Lakes T, Hostert P, 2011. A spatial epidemiological analysis of self-rated mental health in the slums of Dhaka. Int J Health Geogr 10, 36.

Haining R, Wise S, Ma JS, 1998. Exploratory spatial data analysis in a geographic information system environment. J Roy Stat Soc D-Sta 47, 457-469.

Han D, Rogerson PA, Nie J, Bonner MR, Vena JE, Vito D, Muti P, Trevisan M, Edge SB, Freudenheim JL, 2004. Geographic clustering of residence in early life and subsequent risk of breast cancer (United States). Cancer Cause Control 9, 921-929.

Hastie T, 2011. gam: Generalized Additive Models. Version 1.08. Available at: http://CRAN.R-project.org/package=gam (Accessed on February 2011).

Hastie T, Tibshirani R, 1987. Generalized additive-models - some applications. J Am Stat Assoc 398, 371-386.

Hillier A, Cole BL, Smith TE, Yancey AK, Williams JD, Grier SA, McCarthy WJ, 2009. Clustering of unhealthy outdoor advertisements around child-serving institutions: a comparison of three cities. Health Place 4, 935-945.

Huang L, Stinchcomb DG, Pickle LW, Dill J, Berrigan D, 2009. Identifying clusters of active transportation using spatial scan statistics. Am J Prev Med 2, 157-166.

Jackson M, Huang L, Luo J, Hachey M, Feuer E, 2009. Comparison of tests for spatial heterogeneity on data with global clustering patterns and outliers. Int J Health Geogr 1, 55.

Jacquez G, Kaufmann A, Meliker J, Goovaerts P, AvRuskin G, Nriagu J, 2005. Global, local and focused geographic clustering for case-control data with residential histories. Environ Health 1, 4.

Jacquez GM, 2009. Cluster morphology analysis. Spat Spattemporal Epidemiol 1, 19-29.

Jacquez GM, Greiling DA, 2003. Local clustering in breast, lung and colorectal cancer in Long Island, New York. Int J Health Geogr 1, 3.

Johnson S, 2006, The ghost map: the story of London's most terrifying epidemic and how it changed science, cities, and the modern world. Riverhead Books, 299 pp.

Kelsall JE, Diggle PJ, 1995. Non-parametric estimation of spatial variation in relative risk. Stat Med 21-22, 2335-2342.

Kelsall JE, Diggle PJ, 2002. Spatial variation in risk of disease: a nonparametric binary regression approach. J Roy Stat Soc C-App 47, 373-373.

Kulldorff M, 1999. An isotonic spatial scan statistic for geographical disease surveillance. J Natl Inst Pub Health 48, 94-101.

Kulldorff M, 2006. Tests of spatial randomness adjusted for an inhomogeneity: a general framework. J Am Stat Assoc 475, 1289-1305.

Kulldorff M, Huang L, Konty K, 2009. A scan statistic for continuous data based on the normal probability model. Int J Health Geogr 8, 58.

Kulldorff M, Huang L, Pickle L, Duczmal L, 2006. An elliptic spatial scan statistic. Stat Med 22, 3929-3943.

Kulldorff M, Tango T, Park PJ, 2003. Power comparisons for disease clustering tests. Comput Stat Data An 4, 665-684.

Lai P, Low C, Wong M, Wong W, Chan M, 2009. Spatial analysis of falls in an urban community of Hong Kong. Int J Health Geogr 1, 14.

Levac D, Colquhoun H, O'Brien KK, 2010. Scoping studies: advancing the methodology. Implement Sci 5, 69.

Levine N, 2006. Crime mapping and the CrimeStat program. Geogr Anal 1, 41-56.

Lu Y, 2006. Spatial choice of auto thefts in an urban environment. Secur Regul Law J 3, 143-166.

Meliker JR, Jacquez GM, Goovaerts P, Copeland G, Yassine M, 2009. Spatial cluster analysis of early stage breast cancer: a method for public health practice using cancer registry data. Cancer Cause Control 7, 1061-1069.

Meliker JR, Sloan CD, 2011. Spatio-temporal epidemiology: principles and opportunities. Spat Spattemporal Epidemiol 1, 1-9.

Moore DA, Carpenter TE, 1999. Spatial analytical methods and geographic information systems: use in health research and epidemiology. Epidemiol Rev 2, 143-161.

Mostashari F, Kulldorff M, Hartman JJ, Miller JR, Kulasekera V, 2003. Dead bird clusters as an early warning system for West Nile virus activity. Emerg Infect Dis 6, 641.

Neill DB, 2009. An empirical comparison of spatial scan statistics for outbreak detection. Int J Health Geogr 1, 20.

Ngowi HA, Kassuku AA, Carabin H, Mlangwa JED, Mlozi MRS, Mbilinyi BP, Willingham AL, 2010. Spatial clustering of porcine cysticercosis in Mbulu district, northern Tanzania. PLoS Negl Trop Dis 4, e652.

Openshaw S, 1984. Ecological fallacies and the analysis of areal census-data. Environ Plann A 1, 17-31.

Openshaw S, Charlton M, Wymer C, Craft A, 1987. A mark 1 geographical analysis machine for the automated analysis of point data sets. Int J Geogr Inf Sys 4, 335-358.

Orellana D, Wachowicz M, 2011. Exploring patterns of movement suspension in pedestrian mobility. Geogr Anal 3, 241-260.

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.

Ozonoff A, Webster T, Vieira V, Weinberg J, Ozonoff D, Aschengrau A, 2005. Cluster detection methods applied to the Upper Cape Cod cancer data. Environ Health 4, 19.

Paez A, Scott DM, 2004. Spatial statistics for urban analysis: a review of techniques with examples. GeoJ Lib 1, 53-67.

Pasma T, 2008. Spatial epidemiology of an H3N2 swine influenza outbreak. Can Vet J 2, 167-176.

Polack SR, Solomon AW, Alexander NDE, Massae PA, Safari S, Shao JF, Foster A, Mabey DC, 2005. The household distribution of trachoma in a Tanzanian village: an application of GIS to the study of trachoma. Trans R Soc Trop Med Hyg 99, 218-225.

Poljak Z, Dewey CE, Rosendal T, Friendship RM, Young B, Berke O, 2010. Spread of porcine circovirus associated disease (PCVAD) in Ontario (Canada) swine herds: part I. Exploratory spatial analysis. BMC Vet Res 6, 58.

Pollack LA, Gotway CA, Bates JH, Parikh-Patel A, Richards TB, Seeff LC, Hodges H, Kassim S, 2006. Use of the spatial scan statistic to identify geographic variations in late stage colorec-

tal cancer in California (United States). Cancer Cause Control 4, 449-457.

R Development Core Team 2012. R: a language and Environment for Statistical Computing Computing, RFfS ed., Vienna, Austria.

Ripley BD, 1976. The second-order analysis of stationary point processes. J Appl Probab 13, 255-266.

Rothman KJ, 1990. A sobering start for the cluster busters'conference. Am J Epidemiol supp 1, 6-13.

Rowlingson BS, Diggle PJ, 1993. Splancs: spatial point pattern analysis code in S-Plus. Comput Geo Sci 5, 627-655.

Sadahiro Y, 2003. Cluster detection in uncertain point distributions: a comparison of four methods. Comput Environ Urban 1, 33.

Sarkar R, Prabhakar AT, Manickam S, Selvapandian D, Raghava MV, Kang G, Balraj V, 2007. Epidemiological investigation of an outbreak of acute diarrhoeal disease using geographic information systems. Trans R Soc Trop Med Hyg 101, 587-593.

Schuurman N, Cinnamon J, Crooks VA, Hameed SM, 2009a. Pedestrian injury and the built environment: an environmental scan of hotspots. BMC Public Health 9, 233.

Schuurman N, Peters PA, Oliver LN, 2009b. Are obesity and physical activity clustered? A spatial analysis linked to residential density. Obesity 12, 2202-2209.

Selvin HC, Stuart A, 1966. Data-dredging procedures in survey analysis. Am Stat 3, 20-23.

Signorino G, Pasetto R, Gatto E, Mucciardi M, La Rocca M, Mudu P, 2011. Gravity models to classify commuting vs. resident workers. An application to the analysis of residential risk in a contaminated area. Int J Health Geogr 1, 11.

Silverman BW, 1986. Density estimation for statistics and data analysis. Chapman and Hall, 175 pp.

Siqueira-Junior JB, Maciel IJ, Barcellos C, Souza WV, Carvalho MS, Nascimento NE, Oliveira RM, Morais-Neto O, Martelli CM, 2008. Spatial point analysis based on dengue surveys at household level in central Brazil. BMC Public Health 20, 411-421.

Sullivan R, Timmermann A, White H, 1999. Data snooping, technical trading rule performance, and the bootstrap. J Financ 5, 1647-1691.

Tango T, Takahashi K, 2005. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 4, 11.

Tanser F, Bärnighausen T, Cooke GS, Newell M-L, 2009. Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. Int J Epidemiol 4, 1008-1016.

Vieira V, Webster T, Weinberg J, Aschengrau A, 2009. Spatial analysis of bladder, kidney, and pancreatic cancer on upper Cape Cod: an application of generalized additive models to case-control data. Environ Health 8, 3.

Vieira VM, Hart JE, Webster TF, Weinberg J, Puett R, Laden F, Costenbader KH, Karlson EW, 2010. Association between residences in U.S. northern latitudes and rheumatoid arthritis: a spatial analysis of the Nurses' Health Study. Environ Health Persp 7, 957-961.

Vieira VM, Webster TF, Weinberg JM, Aschengrau A, 2008. Spatial-temporal analysis of breast cancer in upper Cape Cod, Massachusetts. Int J Health Geogr 7, 46.

Wang J, McMichael AJ, Meng B, Becker NG, Han W, Glass K, Wu J, Liu X, Liu J, Li X, Zheng X, 2006. Spatial dynamics of an epidemic of severe acute respiratory syndrome in an urban area. Bull World Health Organ 12, 965-968.

Warden CR, 2008. Comparison of Poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. Int J Health Geogr 7, 51.

Westercamp N, Moses S, Agot K, Ndinya-Achola JO, Parker C, Amolloh KO, Bailey RC, 2010. Spatial distribution and cluster analysis of sexual risk behaviors reported by young men in Kisumu, Kenya. Int J Health Geogr 9, 24.

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.

Winskill P, Rowland M, Mtove G, Malima R, Kirby M, 2011. Malaria risk factors in north-east Tanzania. Malar J 1, 98.

Yao Z, Tang J, Zhan FB, 2011. Detection of arbitrarily-shaped clusters using a neighbour-expanding approach: a case study on murine typhus in South Texas. Int J Health Geogr 10, 23.

Zenk SN, Schulz AJ, Matthews SA, Odoms-Young A, Wilbur JE, Wegrzyn L, Gibbs K. Braunschweig C, Stokes C, 2011. Activity space environment and dietary and physical activity behaviors: a pilot study. Health Place 5, 1150-1161.

Zheng P, Diggle P, 2009. spatialkernel: Non-parametric estimation of spatial segregation in a multivariate point process. Version 0.4-19. Available at: http://cran.r-project.org/web/packages/spatialkernel/index.html (Accessed on December 2009).