Title
Point Pattern Analysis for Clusters Influenced by Linear Features

Journal
Transactions in GIS, 19(6)

ISSN
1361-1682

Authors
Li, Li
Bian, Ling
Rogerson, Peter
et al.

Publication Date
2015-12-01

DOI
10.1111/tgis.12119

Transactions in GIS

Research Article Transactions in GIS, 2015, 19(6): 835–847

Point Pattern Analysis for Clusters Influenced by Linear Features: An Application for Mosquito Larval Sites

Li Li,* Ling Bian,† Peter Rogerson,† and Guiyun Yan‡

*IBM China Research Lab, Beijing
†Department of Geography, University of Buffalo, New York
‡University of California, Irvine

Abstract

Although many statistical approaches have been developed to quantify and assess spatial point patterns, the challenge to analyze complicated patterns has yet to be met. Statistics that describe the level of clustering usually assume that point events are isotropic. Many point events are influenced by linear features and clusters of these point events often have an elongated shape. Existing statistical cluster detection approaches often ignore these types of processes. This study proposes a new method, termed an *L*-function analysis for clusters influenced by linear features (*L*-Function-*l*) to test anisotropic point patterns with respect to the orientation of nearby linear features. To explicitly account for the influence of the underlying linear features on the point events, a number of ellipses with varying lengths, orientations and eccentricities are used to replace the circles that are drawn in the original *L*-function analysis. A case study of testing anisotropically clustered patterns of mosquito larval sites is used to illustrate the application of this method. The results indicate that the proposed approach provides a more flexible and comprehensive description of point patterns than the original *L*-function analysis.

1 Introduction

Traditional spatial statistical methods designed for point pattern analysis identify whether point events have regular, random, or clustered patterns (Foxall and Baddeley 2002, Munch et al. 2003, Scalon et al. 2003, Fleischer et al. 2006, Mattfeldt et al. 2006. Si et al. 2008). A clustered pattern often leads to the formation of hypotheses, such as the existence of a non-random process that influences these point events (Sterner et al. 1986). The various statistics that have been used to describe the level of clustering are usually based on an assumption that point events are isotropic. That is, the characteristics of point events are homogeneous in all directions. Subsequently formed hypotheses, when built on such an assumption, may not be well suited for investigating environmental processes that are anisotropic, such as those in a linear form.

Many point events are influenced by nearby linear features. For example, the point observations of wildlife, plant species, and water-borne diseases can be influenced by the spatial alignment of nearby streams (Real and Biek 2007, Maheu-Giroux et al. 2007). Clusters of these point events often have an elongated shape (Conley et al. 2005, Duczmal et al. 2008, Curzon and Keeton 2010, Yiannakoulias et al. 2010). The shape of point clusters often has a strong influence on certain spatial measures for point clusters. For example, commonly used home-range estimators have a poor performance on linear point patterns (Blundell et al. 2001,

Address for correspondence: Dr. Li Li, IBM China Research Lab, Software Park, Bld19, A, F2, Beijing, China 100193. E-mail: lili36@ub-alumni.org

 doi: 10.1111/tgis.12119

Downs and Horner 2008). Gatrell et al. (1996) expressed their concerns regarding the separation between cluster detection and process specification. The information of underlying processes should not be ignored, and the incorporation of such contextual information in the analysis can yield more informative results.

The existing approaches that create measures on the level of clustering for point patterns lack the mechanism to take into account the possible underlying processes. Many methods (e.g. *t*-statistics and regression) test the strength of the association between point locations and possible underlying processes. These methods, however, treat individual points as independent events without considering the spatial interactions between them in the context of a cluster. At the time of writing, few approaches have been proposed to evaluate clustered patterns with regard to their possible underlying processes (Veen and Schoenberg 2006).

This study proposes a new method based on a modified *L*-function analysis to analyze the anisotropically clustered patterns of point events with respect to the orientation of possible underlying linear features. The proposed method, named the *L*-function analysis for clusters influenced by linear features (*L*-Function-*l*), is inspired by the *L*-function (Ripley 1976) and the elliptic spatial scan statistic that detects anisotropic clusters at a local level (Kulldorff 1997). To provide background information for the proposed approach, the next section discusses the basic principles of *L*-function analysis. The *L*-Function-*l* approach is discussed in the section that follows. In the last section, a case study for detecting anisotropic clusters of mosquito larval sites with respect to the orientation of nearby streams is presented to illustrate the application of the approach.

2 *L-function* Analysis for Clusters Influenced by Linear Features

2.1 L-*function Analysis*

A variety of statistical methods have been developed to study point patterns (Boots and Getis 1988, Legendre and Fortin 1989, Ripley 1987). Among them, the *L*-function is one of the most popular for three principal reasons (Fleischer et al. 2006, Foxall and Baddeley 2002, Munch et al. 2003): the level of measurement, the scale of analysis, and properties of complete spatial randomness (CSR). First, point events are often limited to binary attributes (e.g. presence and absence). The *L*-function analysis enables cluster analysis to be conducted on these binary attributes. Second, all geographic processes are scale-dependent and their characteristics may change across scales (Guisan and Thuiller 2005). The *L*-function incorporates the concept of scale in the detection of the level of the event clustering. Finally, the *L*-function considers both properties of CSR: (1) the intensity of events does not vary across a region; and (2) events do not interact with each other within the region (Diggle 1983). CSR is one of the most frequently used spatial reference patterns and the *L*-function tests if an observed pattern is consistent with CSR. The *L*-function has been applied in a variety of disciplines, such as criminology, ecology, epidemiology, geography, and geology (Anselin 2004).

The *L*-function is derived from the *K*-function. To test if a point pattern is consistent with CSR, the *K*-function (*K(r)*) compares the observed number of points within a distance *r* of a chosen point to the expected number. To calculate the *K*-function at scale *r*, hypothetical circles (with radius *r*, often called a lag distance) are placed around each point location, and the average number of points within those circles is calculated. If the point pattern is consistent with CSR, the number of points within these hypothetical circles should be proportional to the areas of the circles. The following function is a *K*-function corrected by an edge effect.

$$K(r) = A/N^2 \sum_i \sum_j I_{d(ij)}/w_{ij} \quad (1)$$

where A is the area of a study region, N is the number of points in the study region, and i and j denote a pair of points in the region. $I_{d(ij)} = 1$ if the distance between point i and point j is less than r; otherwise $I_{d(ij)} = 0$. The weight w_{ij} is used to account for edge effects (when a point is close to the edge of the region, the hypothetical circle centered on i may occupy an area outside of the study region). A widely accepted algorithm for determining w_{ij} is to equate it with the fraction of the circumference of the circle that falls within the study region. This edge-corrected estimate of the K-function provides an unbiased estimator of the K-function.

This procedure is repeated for a range of values for r. The K-function uses the information on all inter-point distances and provides more information on the spatial scale of the pattern than statistics that only use nearest neighbor distances (Diggle 1983). Furthermore, the K-function is a global statistic; at one particular scale, there is one measurement to quantify the level of clustering for the entire region. A CSR process has an expected $K(r)$ value equal to πr^2; a clustered pattern has a $K(r) > \pi r^2$; and a pattern with points repelling each other has a $K(r) < \pi r^2$. Although the K-function can be used to identify clustering, its results are usually transformed using the L-function, which normalizes the K-function (Besag and Diggle 1977).

$$L(r) = \sqrt{\frac{K(r)}{\pi}} - r \quad (2)$$

This transformed version of the K-function simplifies interpretation. The value range of L-function is $[-\infty, +\infty]$. An L-function value of zero indicates a random pattern. If an L-function value is greater than zero at a certain scale, it indicates the possible existence of a clustered pattern at that scale. If an L-function value is less than zero at a scale, it indicates the possible existence of spatial repulsion. In addition to hypothesis testing, the L-function has also been used to compare the differences between point patterns. A Monte Carlo method is used to test statistical significance by generating an arbitrary number of simulations of a null model. The L-function analysis is applied to the results of these simulations to derive a wide range of L-function values, from which confidence intervals around the observed L-function values can be obtained.

The L-function analysis does not consider any other information about a point event besides its location. Contextual information, such as possible influences of linear features, cannot be easily incorporated within L-function analysis. The proposed L-Function-l detects anisotropically clustered patterns with respect to the orientation of nearby linear features.

2.2 L-*Function*-l

The proposed L-Function-l is derived from the original L-function. For L-Function-l, the null hypothesis assumes that there is no clustered pattern with respect to the orientation of nearby linear features.

If point events are clustered along linear features, they tend to locate within a certain distance of these linear features. These clusters may have an elongated shape whose alignment is likely influenced by the linear features. Elongated clusters have been considered previously in cluster analysis (Ho and Chen 1995, Kulldorff et al. 2006). For example, the elliptic spatial scan statistic uses an elliptical scan window, instead of the circular scan window used in the original spatial scan statistic (Kulldorff 1997), to search for elongated clusters. The major advantage of using the elliptic window is that it is more flexible than the circular window in

identifying clusters with a broader range of orientations and shapes. Although ellipses have been used to detect the locations of clusters, they have not been applied in the context of *K*- or *L*-statistics. This study incorporates ellipses to identify anisotropically clustered patterns with respect to the orientation of nearby linear features. To take the orientation of linear features into account, it is necessary to incorporate this attribute into the proposed test.

Linear features are often represented by a set of lines and their associated vertices. If they have a simple structure, such as a straight line, the definition of orientation is straightforward. However, linear features, such as streams, often have a complicated structure with various parts aligned in different orientations. To quantify the orientation of these lines, the most effective way is to break them into 'basic sections', for which an orientation can be easily identified. There are several possible approaches to identifying the basic sections (Li and Openshaw 1992). The three that are discussed in this article are the most straightforward to apply. One is used in the case study presented in a later section.

Suppose a linear feature is considered as consisting of a series of line segments, each of which lies between two consecutive vertices. Each line segment has an orientation, ranging from 0 to 180 degrees clockwise (without differentiating which of the two vertices is the starting point of a line segment), with due North given as 0 degrees. The first approach treats each line segment as a basic section. In the second approach, a number of consecutive line segments are considered to constitute one basic section if they fall between two (not necessarily consecutive) vertices that are a certain distance apart. Using the third approach, a number of consecutive line segments are considered as one basic section if the differences in the orientation of these pairs of consecutive line segments are less than a pre-determined threshold. After a basic section is determined, its orientation can be determined in several ways. In this article, the third approach of identifying a basic section is used, and the locations of the starting and ending vertices of a basic section are used to calculate its orientation. This orientation is subsequently used as the reference orientation to evaluate anisotropic clusters. The choice of an appropriate approach to identify a basic section depends on the goal of the study, the scale of analysis, and the characteristics of the linear features.

The *L*-Function-*l* approach begins by drawing hypothetical ellipses around points in a point pattern. Only those points that are within a specified distance of linear features are considered, since linear features may not have influence on points that are far away. This distance is determined based on the empirical relationship between the point events and the linear features. While a hypothetical circle used in the original *K*-function is defined by a single parameter, i.e. the radius, an ellipse is defined by three parameters: the length of its semi major axis (a), the orientation of its major axis (θ), and its eccentricity (ε) (i.e. the ratio of c to a, with c being the distance between a foci point and the center of the ellipse). These three parameters represent two additional characteristics to those possessed by a circle, orientation (θ) and eccentricity (ε). Similar to the original *K*-function, the size of ellipse varies within a predefined range by varying a to determine the scale of the clusters. The orientation and eccentricity of ellipses also change by varying θ and ε, respectively. To consider the influence of linear features, the orientation (θ) of ellipses is defined based on the orientation of the nearby linear features. Similar to the original *K*-function analysis, the number of points within the ellipses is counted at each selected scale, orientation, and eccentricity and the level of clustering is calculated using a modified *K*-function, termed the *K*-Function-*l* (t, θ, ε) that is corrected for edge effects:

$$K - \text{Function} - l(t, \theta, \varepsilon) = A/N^2 \sum_i \sum_j I_{d(ij)}/w_{ij} \tag{3}$$

where $I_{d(ij)} = 1$ if the sum of distances from point j to the two focal points of an ellipse with center i is shorter than t ($t = 2a$), otherwise $I_{d(ij)} = 0$. $I_{d(ij)} = 0$ if $i = j$. The weight w_{ij} is used to account for edge effects when an ellipse partially occupies an area outside of the study region. For the original K-function, there are several methods to calculate this weight and the differences between several important methods have been compared (Yamada and Rogerson 2003). Among these methods, Ripley's (1976) correction is the most widely used and it can be easily adapted to the calculation of the K-Function-l. In Ripley's correction, w_{ij} is equal to the ratio between the area of the circle (ellipse) that is inside of the study region and that of the entire circle (ellipse).

K-Function-l (t, θ, ε) is the expected number of points in an ellipse with a major axis of t, orientation of θ, and eccentricity of ε, centered at a point in a point pattern, divided by the intensity of the pattern. A random process has an expected value of K-Function-l $(t, \theta, \varepsilon) = \pi ab$; K-Function-l $(t, \theta, \varepsilon) > \pi ab$ implies clustering, and K-Function-l $(t, \theta, \varepsilon) < \pi ab$ indicates a pattern where points are more dispersed than random. The L-Function-l (t, θ, ε) is scale, orientation, and eccentricity dependent. Similar to the original L-function, K-Function-l (t, θ, ε) is then normalized and transformed into an L-Function-l as follows:

$$L - Function - l(t, \theta, \varepsilon) = \sqrt{\frac{K - Function - l(t, \theta, \varepsilon)}{\pi\left(\sqrt{1-\varepsilon^2}\right)}} - t \tag{4}$$

The value range of the L-Function-l is $[-\infty, +\infty]$. An L-Function-l value of zero indicates a random pattern. If a value is greater than zero, it indicates the possible existence of clustering at that given scale, orientation, and shape, while a value less than zero indicates the possible existence of spatial repulsion.

The L-Function-l is designed to identify clustered patterns with respect to the orientation of nearby linear features. It can be used to summarize a point pattern, test hypotheses and estimate parameters, and fit models. When the L-Function-l is used for hypothesis testing, a significance test is needed. Since the context of the L-Function-l is different from the original L-function, an alternative significance test must be devised. The following section introduces a significance test that takes into account the orientation of linear features for the point pattern analysis.

2.3 *Significance test for* L-*Function*-l *values*

The null model commonly used for testing the original L-function is CSR. The drawback of this null model for the intended study is that it does not consider any contextual information. Since L-Function-l is intended to test clustered patterns with respect to the orientation of nearby linear features, a more informative null model is needed. Various models representing clustering point processes have been developed for this purpose. A Matern cluster point pattern is created through a two-step process (Baddeley et al. 1996). In the first step, a homogeneous Poisson point pattern with intensity λ_1 is created. In this step, the generated points are referred to as mother points. In the second step, each mother point is replaced by a random cluster of points with a predefined sphere, which is often referred to as a cluster of child points. These random clusters of points also have a Poisson distribution. Because the child points created in the second step are based on the locations of mother points, the Matern cluster point pattern is not completely random. Instead, it incorporates the contextual information in the point pattern, i.e. the locations of mother points that influence the locations of child points, making it advantageous for use in the proposed L-Function-l.

Another point process that also incorporates contextual information is the Segment Cox process (Pandey 2010). In this process, line segments are first randomly located in space; then each is populated with random points on them.

In principle, both the Matern cluster point process and the Segment Cox process meet the purpose of *L*-Function-*l* better than the null models based on CSR. Specifically, the two-step Matern process allows for a mother pattern to influence a child pattern, but the process does not involve linear features. The Segment Cox process incorporates linear features in the mother pattern but requires child points to be located on linear features. Therefore, modifications to the Matern cluster point process and the Segment Cox process are necessary here.

To fulfill the needs of this study, we combine elements of the Matern cluster point process and the Segment Cox process. This consists of two steps. In the first step, a buffer with a certain width is created around the linear features. This width can be determined based on the empirical relationship between the point events and the linear features. In the second step, this buffer is then populated with randomly distributed points. The number of randomly distributed points is equal to the number of points within the buffer in the observed point pattern.

To obtain a reliable estimate of the confidence interval in order to test the statistical significance of observed *L*-Function-*l* values, a large number of simulations are required for this combined Matern-Cox process. The simulated *L*-Function-*l* values for a given scale, orientation, and eccentricity are used to construct 90% confidence intervals and to indicate the significance level of observed values. Unlike the original *L*-function that considers only one parameter (scale), *L*-Function-*l* involves three parameters and each may take various values. This unavoidably involves testing multiple parameters simultaneously and inflates the number of tests. This multiple-testing problem can be corrected by using, for example, a Bonferroni adjustment that adjusts the level of significance for the tests by the number of parameter combinations that are tested.

To determine whether a point pattern is clustered, its *L*-Function-*l* values are compared with the confidence interval (with the multi-testing problem corrected by the Bonferroni adjustment). If the *L*-Function-*l* values are within the confidence interval, the points are considered to be randomly distributed for that particular combination of parameters. If *L*-Function-*l* values are above the confidence interval, this indicates the existence of more points than expected for that particular type of clustered pattern, and if the observed value is below the confidence interval, there are fewer points than expected for the given scale, orientation, and eccentricity.

3 An Application of *L*-Function-*l* to Analysis of Mosquito Larvae Sites

An analysis of the point distribution of mosquito larval sites is used to illustrate the proposed *L*-Function-*l* approach. Malaria is a vector-borne disease and mosquitoes are the vector that transmits the disease to human populations. For the last two decades, malaria control has been the mission of many health organizations worldwide because malaria affects up to a half billion people in Africa each year (Guinovart et al. 2006). The spatial distribution of larvae is an important determinant of the distribution of adult mosquitoes, which in turn determines the areas where the malarial risks are the greatest. Understanding the spatial distribution of mosquito larvae sites is a prerequisite for the design of effective mosquito control strategies. It has been observed that the survival of larval is dependent on aquatic environments, such as areas along streams (Bian et al. 2006, Mushinzimana et al. 2006, Li et al. 2008, 2009). The proposed *L*-Function-*l* is used to detect clustered patterns of mosquito larvae sites with respect to the orientation of nearby streams.

3.1 Study area and Data

The study area is a 4×4 km^2 area in the Kakamega district of western Kenya where malaria epidemics are prevalent. The primary mosquito species in this region is *An. gambiae* (Munga et al. 2009). The region has a lengthy rainy season extending from April through June and most malaria cases occur during this season.

Data on the locations of *An. gambiae* larvae in the study area were collected by a group of biologists in May 2003 (a detailed description of the data can be found in Mushinzimana et al. 2006). The dataset contains 721 sites where larvae were observed with each site represented as a point. In addition to the larval sites, GIS data on streams in the study region were also obtained. The larval sites are mainly distributed along streams. The distance between the sites and the streams ranges from 0–510 m, with an average of 63 m. Of the 721 sites, 680 locations are lying within 480 m of a stream and encompass 95% of the total larval sites.

3.2 L-*Function*-l *Analysis*

The 680 points that are within 480 m of a stream are used in the *L*-Function-*l* analysis. A buffer width of 480 m is selected accordingly. Each point is used as the center for an ellipse. The closest basic section of a nearby stream is identified using the third approach discussed previously (see Section 2.2). Consecutive line segments with an angular difference of less than 10 degrees form a basic section. In this study, the length of the shortest basic segment is 50 m, while the length of the longest basic section is 710 m. Because changes in the orientation of streams in the study area are minor, this approach is considered the most appropriate. If the length of a basic section is shorter than the length of an ellipse, more than one basic section may have an influence on the clusters. In this case, the basic section is extended by merging it with an adjacent basic section. If multiple adjacent sections are present, the one that is closer to the center of the ellipse is chosen first (as shown in Figure 1) and the process continues with other adjacent basic sections until the total length of the located basic sections is equal to or larger than the length of the ellipse. The major axis of the smallest rectangle that encloses the merged basic sections is used to determine their orientation. This axis is used as the reference axis to evaluate the orientation of ellipses (see Figure 1).

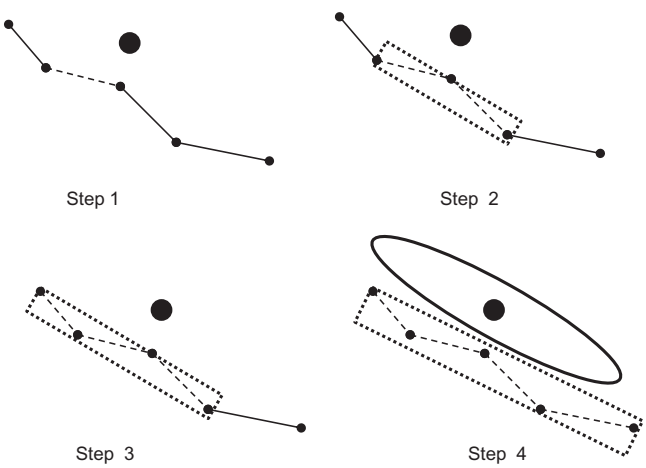

Figure 1 An illustration of how to identify basic sections

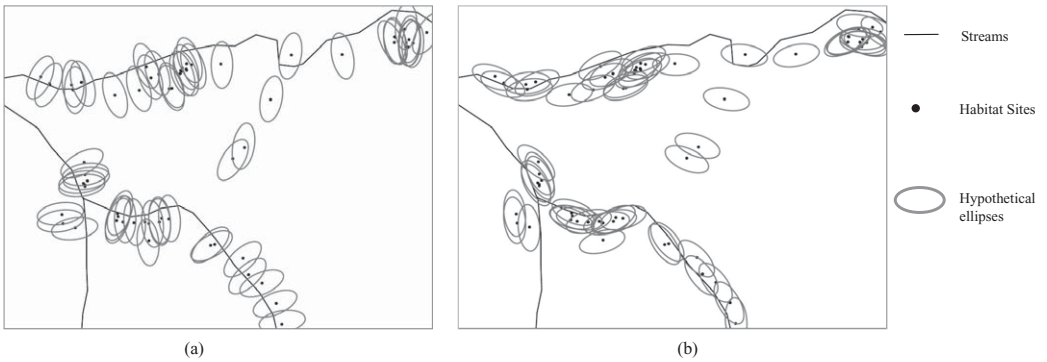

Figure 2 Larval sites surrounded by ellipses with an ε of 0.6 and orientation values of 0 degrees (a) and 90 degrees (b)

The *L*-Function-*l* values (Equation 4) are calculated with varying scales (t), orientations (θ), and eccentricities (ε). The t value ranges from 50–750 m using 50 m increments to detect clusters at the resultant 15 different scales. The use of 50 m increments ensures that a majority of the ellipses touch their nearest basic section. The purpose of selecting 750 m as the maximum t value is because the length of the longest basic segment is 710 m. The chosen maximum value of t is also much smaller than the length of the study area, since if t is too large, edge effects would start to affect the results. In this case, the 750 m scale is 18.75% of the length of the study area. For each given t, the orientation (θ) of the ellipse is varied to detect the clustered pattern with respect to the orientation of nearby streams. Four orientations are evaluated, 0, 45, 90, and 135 degrees from the reference orientation determined by basic sections of nearby streams.

The eccentricity ε of the ellipse is also varied to help detect elongated clusters. The ε value ranges from 0.3 to 0.9, using 0.3 increments, resulting in three eccentricities. A preliminary analysis suggested that using nine eccentricities (0.1 to 0.9 incremented by 0.1) is redundant. Of the three eccentricities, the ellipse with an ε of 0.3 is closer to a circle than the other two values. Using a value 0.9 results in the most elongated ellipse given that an ε equal to 1 is a one-dimension line segment. Figure 2 shows ellipses with an ε of 0.6 centered at larval habitat sitesalong streams. While the 15 scales (i.e. the t values) account for the scale effect that is also considered in the original *L*-function, the four orientations (θ) and the three eccentricities (ε) explicitly account for the anisotropic characteristics of clusters. This combination produces $15 \times 4 \times 3 = 180$ *L*-Function-*l* values in total.

The combined Matern-Cox process is used to test the significance of the *L*-Function-*l* values. In the first step, a buffer with a width of 480 m is created around the streams (see Section 3.1). In the second step, the buffer is populated with 680 random points. To address the multiple testing problem, a Bonferroni adjustment is applied and the significance level for each individual test is $0.05/180 = .0003$. The simulation is conducted 200,000 times in total. The confidence intervals that contain $(1 - .0003) \times 100\% = 99.97\%$ of the simulated *L*-Function-*l* values are obtained. The calculations are implemented in a Matlab environment (www.mathworks.com).

3.3 Results of L-*Function*-l *Analysis*

The 180 *L*-Function-*l* values are displayed in Figure 3 and are organized based on their orientations (Figures 3a–3d) with reference to the basic sections of nearby streams. For a given ori-

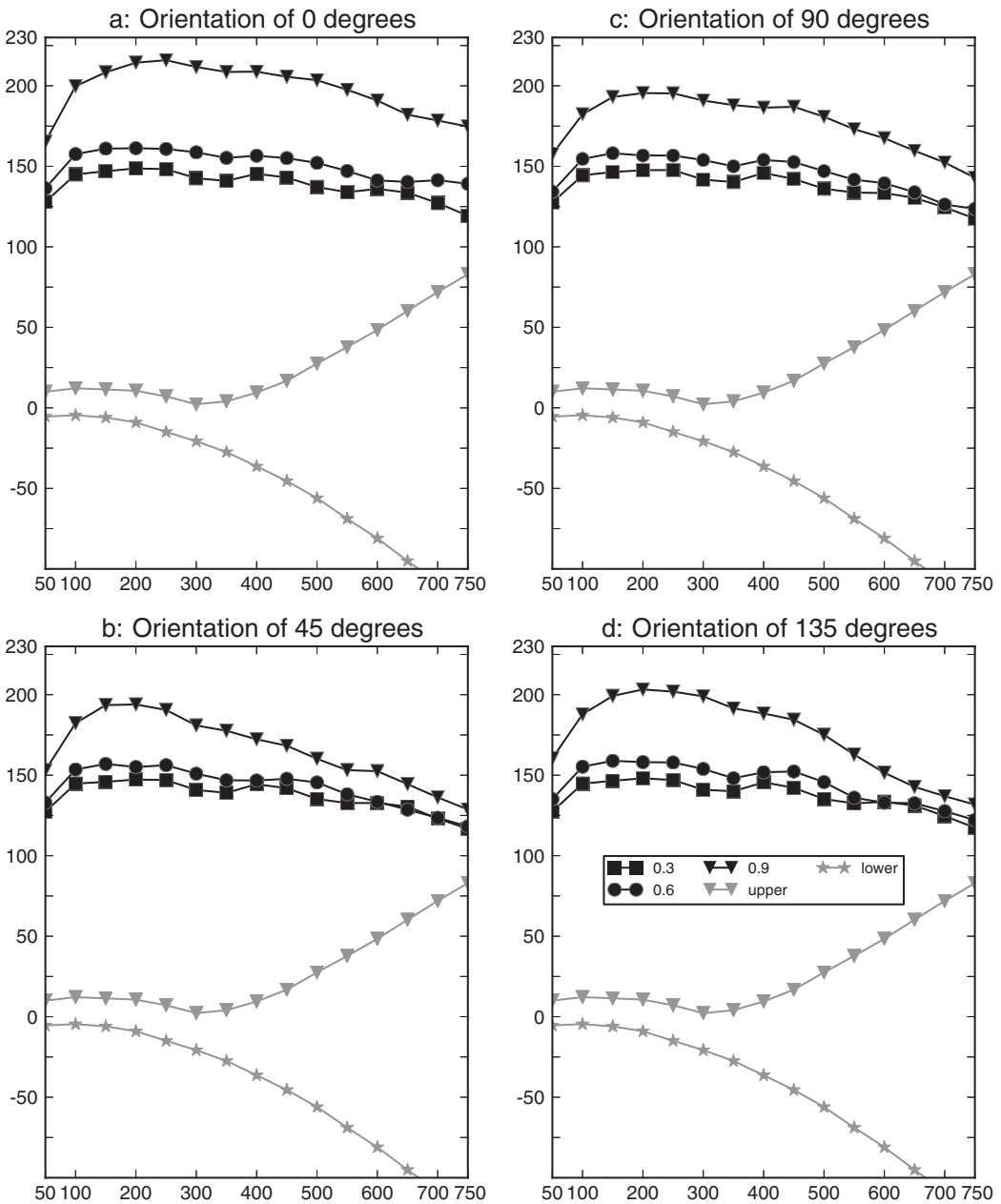

Figure 3 *LACILF* and original *L*-function values organized by their orientation values of 0 degrees (a), 45 degrees (b), 90 degrees (c), and 135 degrees (d) with reference to the basic sections of nearby streams. The horizontal axis represents the 15 scales (50–750 m incremented by 50 m), and the vertical axis represents the *LACILF* values. The two gray lines with inverse triangle and * symbols delimit the upper and lower bound, respectively, of the 90% confidence interval. Black lines with ★, ■, ●, and ▽ symbols are the observed *LACILF* values calculated with the three eccentricities (0.3, 0.6, and 0.9), respectively

entation, a set of three *L*-Function-*l* values corresponding to the three eccentricities ($\varepsilon = 0.3$, 0.6, 0.9) is plotted against the 15 scales. To test their statistical significance, the *L*-Function-*l* values are compared with their upper and lower 99.97% confidence interval values that are calculated using the simulated process associated with the ellipses.

All *L*-Function-*l* values are considerably greater than their upper confidence interval limits as shown in all of the sub-figures. This indicates the existence of clustered patterns in larval sites in all orientations with respect to the orientations of nearby streams. Among the four orientations, the *L*-Function-*l* values corresponding to the orientation of 0 degrees (ellipses parallel to nearby streams) are consistently higher than *L*-Function-*l* values in other orientations. The *L*-Function-*l* values with an orientation of 90 degrees (ellipses perpendicular to the streams) are the lowest among the four. These observations indicate that clustering in the direction parallel to the streams is the most prominent, implying the underlying influence of these streams on the orientation of clusters.

In all orientations, *L*-Function-*l* values increase rapidly with an increasing scale (t) value until the scale reaches around 200–350 m, after which the *L*-Function-*l* values decrease steadily with an increasing scale value. These results suggest that the clustering of larval sites in all orientations are scale dependent, with clusters being most prominent when their sizes (the length of the major axis) are around 200–250 m. For a given orientation and scale, *L*-Function-*l* values corresponding to eccentricity of 0.9 are always considerably higher than other eccentricity values. This suggests that larval sites tend to cluster in a linear shape, further implying the influence of nearby steams on both the orientation and the eccentricity of clusters.

Among all 12 sets of *L*-Function-*l* values (three eccentricity values in four orientations), those based on ellipses with the 0 degree orientation and eccentricity equal to 0.9 are the highest and those based on ellipses with the 90 degree orientation with eccentricity equal to 0.3 are the lowest. This implies that most clusters are aligned with the nearby streams, have a linear shape, and are most prominent when they are 200–250 m in size. The influence of nearby steams is evident on the anisotropy of the clustered patterns.

4 Discussion and Conclusions

In this study, a new method, *L*-Function-*l*, is proposed to test the level of clustering with respect to the orientation of nearby linear features. By incorporating hypothetical ellipses in the analysis, this method offers the flexibility to test various anisotropically clustered patterns. The sensitivity of *L*-Function-*l* analysis to the influences of linear features provides a more realistic and comprehensive description of point patterns. A case study of testing anisotropically clustered patterns in mosquito larval sites is used to illustrate the application of this new method. The proposed approach takes into account the influences of nearly linear features and thus can lead to conclusions that are not typically derived from the traditional approaches. For example, an information-rich conclusion based on our case study is that most clusters of larval sites are aligned with the nearby streams with a size of 200–250 m.

Examination of clustered patterns in terms of scale, orientation, and eccentricity is important for scientific questions under investigation. For example, a clustered pattern may result from many possible spatial processes. The proposed *L*-Function-*l* method can be used to help determine the dominant underlying processes. In the case of malaria research, this information is quite meaningful. For example, it has been previously suspected that water-related human activities, such as water fetching and brick making in the study area, might have facilitated the

creation of many larval sites (Zhou et al. 2007). Given the observation that the level of clustering is optimal at the orientation parallel to the streams and at the scale of 200–250 m (Figure 3a), it seems less likely that human activities alone are the direct cause of these clusters. This is because human dwellings in the area are located at a higher altitude and are far more than 200–250 m away from stream valleys (Li et al. 2008). Instead, variations in soil, vegetation, landuse, and landforms with a scale of 200–250 m along certain parts of streams might have contributed to the formation of linearly shaped clusters aligned with streams. The reason is that these factors often have a strong influence on the activity range of adult mosquitoes (Zhou et al. 2007).

Local cluster detection approaches which consider the shape of clusters, such as Kulldorff's elliptical scan statistics and Tango and Takahashi's flexible scan statistic, are used in a wide range of applications (Kulldorff 1997, Tango and Takahashi 2005). Note that these approaches are local statistics and the *L*-Function-*l* is a global statistic. While the local statistics attempt to identify where the local clusters are, the global statistics aim to quantify the overall level of clustering for a point pattern. The *L*-Function-*l* method can render critical measures such as the level of clustering and its associated scales. It has greater statistical power for point patterns influenced by linear features. For example, if we create a buffer around a stream and generate some random points within this buffer, elliptical scan statistics can identify local "clusters" from these random points. In comparison, the results of the *L*-Function-*l* will indicate a random point pattern at certain scales since the proposed approach takes into account of the influence of linear features.

Both local and global effects can influence clustered patterns (Diggle 1983). Although the *L*-Function-*l* method is intended to be a global statistic, the simulated null process takes underlying contexts into account, which reflects local effects (e.g. influences of nearby streams on the cluster pattern of larval sites). These characteristics allow the test to consider both global and local influences. Although the initial assumption on the relationship between linear features and point patterns is not required, existing knowledge on such relationship would be helpful for the implementation of the *L*-Function-*l*. For example, in this study, the buffer width of 480 m is selected to include at least 95% of larval sites. A narrower buffer width can be selected to exclude larval sites that are influenced by other water sources (e.g. wells) if prior knowledge or more detailed information on the relationships between larval sites and streams are obtained.

Finally, we conclude with a few remarks on the proposed approach. The *L*-Function-*l* extends the classic circle-based *L*-function into an ellipse-based function. This extension offers greater flexibility than the circular approach in identifying elongated clustered patterns of various orientation and shape. In addition, the test explicitly incorporates nearby linear features into cluster detection. Using the proposed approach, the influence of possible underlying processes that have a linear form can be hypothesized more directly.

There are some limitations of the *L*-Function-*l*. First, the possible influences of the complexity of the linear features on the results of the proposed methods need to be further investigated. For example, if the proposed approach is to be applied to point patterns influenced by road networks densely located in a city, the inherent characteristics of the road networks (e.g. grid like patterns) may have an influence on the measures and testing procedures. Note that the proposed approach is not designed for point patterns constrained by linear features (i.e. points constrained by linear features have to locate on linear features). For point patterns that are constrained by linear features, K-function for Network-constrained Clusters are better suited (Yamada and Thill 2004, 2010). Second, there is room for improvement in the testing procedure. One possible improvement would be to find alternative null models. In this study,

we proposed a simulated process in which we assume that the point pattern is randomly distributed within a buffer of linear features. It is also possible for a point pattern to have a distance decay relationship with linear features.

References

Anselin L 2004 Review of Cluster Analysis Software. Unpublished Report

Baddeley A, Lieshout M, and Moller J 1996 Markov properties of cluster processes. *Advances in Applied Probability* 28: 346–55

Besag J and Diggle P J 1977 Simple Monte Carlo tests for spatial pattern. *Applied Statistics* 2: 327–33

Bian L, Li L, and Yan G 2006 Combining global and local estimates for spatial distribution of mosquito larval habitats. *GIScience and Remote Sensing* 43: 128–41

Blundell M, Maier J, and Debevec M 2001 Linear home ranges: Effects of smoothing, sample size, and autocorrelation on kernel estimates. *Ecological Monographs* 71: 469–89

Boots B N and Getis A 1988 *Point Pattern Analysis*. Thousand Oaks, CA, Sage

Conley J, Gahegan M, and Macgill J 2005 A genetic approach to detecting clusters in point data sets. *Geographical Analysis* 37: 286–314

Curzon M T and Keeton W S 2010 Spatial characteristics of canopy disturbances in riparian old-growth hemlock-northern hardwood forests, Adirondack Mountains, New York, USA. *Canadian Journal of Forest Research* 40: 13–25

Diggle P J 1983 *Statistical Analysis of Spatial Point Patterns*. Thousand Oaks, CA, Sage Publications

Downs J and Horner M W 2008 Effects of point pattern shape on home-range estimates. *Journal of Wildlife Management* 72(8): 13–18

Duczmal L, Cançado A L F, and Takahashi R H C 2008 Geographic delineation of disease clusters through multi-objective optimization. *Journal of Computational and Graphical Statistics* 17: 243–62

Fleischer F, Beil M, Kazda M, and Schmidt V 2006 Case studies on spatial point processes models. In Baddeley A, Gregori P, Mateu J, Stoica R, and Stoyan D (eds) *Case Studies in Spatial Point Process Modeling*. Berlin, Springer Lecture Notes in Statistics Vol. 185: 235–60

Foxall R and Baddeley A 2002 Nonparametric measures of association between a spatial point process and a random set, with geological applications. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 51: 165–82

Gatrell A C, Bailey T C, Diggle P J, and Rowlingson B S 1996 Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers* 21: 256–74

Guinovart C, Navia M M, Tanner M, and Alonso P L 2006 Malaria: Burden of disease. *Current Molecular Medicine* 6: 137–40

Guisan A and Thuiller W 2005 Predicting species distribution: Offering more than simple habitat models. *Ecology Letters* 8: 993–1009

Ho C and Chen L 1995 A fast ellipse/circle detector using geometric symmetry. *Pattern Recognition* 28: 117–24

Kulldorff M 1997 A spatial scan statistic. *Communications in Statistics: Theory and Methods* 26: 1481–96

Kulldorff M, Huang L, Pickle L, and Duczmal L 2006 An elliptic spatial scan statistic. *Statistics in Medicine* 25: 3929–43

Legendre P and Fortin M J 1989 Spatial pattern and ecological analysis. *Plant Ecology* 80: 107–38

Li L, Bian L, and Yan G 2008 A study of the distribution and abundance of the adult malaria vector in western Kenya highlands. *International Journal of Health Geographics* 7: 50

Li L, Bian L, Yakob L, Zhou G, and Yan G 2009 Temporal and spatial stability of *Anopheles Gambiae* larval habitat distribution in western Kenya highlands. *International Journal of Health Geographics* 8: 70

Li Z and Openshaw S 1992 Algorithms for automated line generalization based on a natural principle of objective generalization. *International Journal of Geographical Information Systems* 6: 373–89

Maheu-Giroux M and de Blois S 2007 Landscape ecology of *Phragmites australis* invasion in networks of linear wetlands. *Landscape Ecology* 22: 285–301

Mattfeldt T, Eckel S, Fleischer F, and Schmidt V 2006 Statistical analysis of reduced pair correlation functions of capillaries in the prostate gland. *Journal of Microscopy* 223: 107–19

Munch Z, Lill S W V, Booysen C N, Zietsman H L, Enarson D A, and Beyers N 2003 Tuberculosis transmission patterns in a high-incidence area: a spatial analysis. *International Journal of Tuberculosis and Lung Disease* 7: 271–77

Munga S, Yakob L, Mushinzimana E, Zhou G, Ouna T, Minakawa N, Githeko A, and Yan G 2009 Land use and land cover changes and spatiotemporal dynamics of anopheline larval habitats during a four-year

period in a highland community of Africa. *American Journal of Tropical Medicine and Hygiene* 81: 1079–84

Mushinzimana E, Munga S, Minakawa N, Li L, Feng C, Bian L, Kitron U, and Yan G 2006 Landscape determinants and remote sensing of anopheline mosquito larval habitats in the western Kenya highlands. *Malaria Journal* 5: 13

Pandey B 2010 Statistically significant length-scale of filaments as a robust measure of galaxy distribution. *Monthly Notices of the Royal Astronomical Society* 401: 2687–96

Real L A and Biek R 2007 Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. *Journal of the Royal Society Interface* 4: 295–307

Ripley B D 1976 The second-order analysis of stationary point processes. *Journal of Applied Probability* 13: 255–66

Ripley B D 1987 Spatial point pattern analysis in ecology. *Developments in Numerical Ecology* 5: 407–29

Scalon J D, Fieller N R J, Stillman E C, and Atkinson H V 2003 Spatial pattern analysis of second-phase particles in composite materials. *Materials Science and Engineering A* 356: 245–57

Si Y L, Debba P, Skidmore A K, Toxopeus A G, and Li L 2008 Spatial and temporal patterns of global H5N1 outbreaks. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 37: 69–74

Sterner R, Ribic C, and Schatz G 1986 Testing for life historical changes in spatial patterns of four tropical tree species. *Journal of Ecology* 74: 621–33

Tango T and Takahashi K 2005 A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 4: 11

Veen A and Schoenberg F 2006 Assessing spatial point process models using weighted k-functions: Analysis of California earthquakes. In Baddeley A, Gregori P, Mateu J, Stoica R, and Stoyan D (eds) *Case Studies in Spatial Point Process Modeling*. Berlin, Springer Lecture Notes in Statistics Vol. 185: 293–306

Yamada I and Thill J-C 2004 Comparison of planar and network k-functions in traffic accident analysis. *Journal of Transport Geography* 12: 149–58

Yamada I and Thill J-C 2010 Local indicators of network-constrained clusters in spatial patterns represented by a link attribute. *Annals of the Association of American Geographers* 100: 269–85

Yamada I and Rogerson P 2003 An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical Analysis* 35: 97–109

Yiannakoulias N, Wilson S, Kariuki H C, Mwatha J K, Ouma J H, Muchiri E, Kimani G, Vennervald B J, and Dunne D W 2010 Locating irregularly shaped clusters of infection intensity. *Geospatial Health* 4: 191–200

Zhou G, Munga S, Minakawa N, Githeko A K, and Yan G 2007 Spatial relationship between adult malaria vector abundance and environmental factors in Western Kenya Highlands. *American Journal of Tropical Medicine and Hygiene* 77: 29–35

