Length L-function for network-constrained point data

Zidong Fang¹,² | Ci Song¹,² | Hua Shu¹,² | Jie Chen¹,² | Tianyu Liu¹,² | Xi Wang¹,² | Xiao Chen¹,² | Xiaorui Yan¹,² | Tao Pei¹,²,³

¹State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
²University of Chinese Academy of Sciences, Beijing, China
³Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China

Abstract

Network-constrained points are constrained by and distributed on road networks, for example, taxi pick-up and drop-off locations. The aggregation pattern (clustering) of network-constrained points (significantly denser than randomly distributed) along roads may indicate spatial anomalies. For example, detecting and quantifying the aggregation with the highest intensity (i.e., the number of taxi pick-up points per network length) can reveal high taxi demand. The network K-function and its derivative (incremental network K-function) have been utilized to identify point aggregations and measure aggregation scale, yet can only identify radius-based planar-scale results, thereby mis-estimating aggregation patterns owing to the network topology configuration heterogeneity. Specifically, complex road networks (e.g., intersections) may incur aggregations despite their low intensity. This study constructs the length L-function for network-constrained points, using its first derivative to quantify the true-to-life aggregation scale and the local function to extract aggregations. Synthetic and practical data experiments show innovative detection of aggregations at the length-based scale and with high intensity, providing a new approach to point pattern analysis of networks unaffected by topological complexity.
INTRODUCTION

Network-constrained points are those located only over restricted network space, such as pedestrian dwell points in an urban road network. Among point patterns, aggregation (also known as clustering) is the most common and significant form, which represents the most intense gathering of points within a given research area. The point set exhibiting the aggregation pattern is also called aggregation (or cluster) and the range of this point set is the aggregation scale (or cluster size). Aggregation can be identified in planar space if the average number of points within a circular radius-based \( r \)-neighborhood centered on a given point is statistically greater than expected for a distribution that follows complete spatial randomness (CSR) (Kiskowski et al., 2009). Similarly, an aggregation in network space is a point set with a higher intensity (i.e., points number per unit road network length) than would be expected for CSR distribution. Detecting the point aggregation and quantifying its scale helps perceive the range and degree of the spatial anomalies and plays an important role in their application to road networks, for example, detecting spatial aggregations of traffic crashes (Nie et al., 2015), origin and destination points of taxis (Deng et al., 2019), and roadside plant populations (Spooner et al., 2004). The solution for detecting aggregation and measuring its scale in networks has been proposed using the network K-function (Okabe & Yamada, 2001). Yet, existing methods can only detect the radius-based planar scale using an \( r \)-neighborhood and thereby may overestimate or underestimate the aggregation in networks. Specifically, points with high intensity along a straight road (see Figure 1a) are not treated as aggregation, while the points with low intensity in dense roads (see Figure 1b) will be seen as aggregation because the \( r \)-neighborhood possesses more points in dense roads. It is thereby needed to judge aggregation from a different scale to solve issues related to the biased estimation of the aggregation pattern in the network K-function method.

Identifying and quantifying aggregations is important for understanding point patterns in road networks, and some related methods in the context of urban networks have been proposed (Baddeley et al., 2021; Okabe et al., 2006; Okabe & Sugihara, 2012). The first group of methods is mainly based on measuring the spatial association of points, which are extended from some classic spatial statistical methods. For instance, Moran’s I statistic (Moran, 1950) has been applied to measure the spatial autocorrelation of network-constrained points based on proximity relationships (Black, 1992). Nevertheless, this group of methods focuses on the spatial association of proximate network points and thus the distribution characteristics of aggregation, such as scale, are not able to be well quantified.

The second group of methods are based on point pattern analysis in networks. There are two correlative approaches related to the two distinct aspects of spatial point patterns: first- and second-order effects (O’Sullivan & Unwin, 2003). The first class is based on point density and focuses on the underlying properties of point events and measures the variation in the mean value of the process. It includes methods such as network-based quadrat count analysis (Shiode, 2008), kernel density estimation for network-constrained points (Okabe et al., 2009; Xie & Yan, 2008, 2013), and others. The second class is based on point separation and mainly examines the spatial interaction of points for spatial patterns, including methods extended from Ripley’s K-function (Ripley, 1976, 1977). These

![Figure 1](image-url)

**Figure 1** \( r \)-neighborhood in the network K-function. (a) and (b) have the same radius as \( r \) but different lengths of networks: (a) the length of the \( r \)-neighborhood is \( 2r \), and (b) the length of the \( r \)-neighborhood is \( 6r \).
two classes of approach focus on different properties of the point process; thus, the analysis results also have different emphases. For instance, the kernel-density estimated surface derived from a point pattern is helpful in identifying the regions of aggregation of a phenomenon of interest, while the distance-based K-function method may not only identify the aggregation but also help to quantify characteristic distances within a pattern.

Because of its powerful capability for detecting and quantifying aggregation, scholars have extended Ripley’s K-function to the network space, making it adaptable for urban applications. The network K-function using the shortest-path distance and its derivative, namely the incremental network K-function, has been proposed to analyze the point distribution in networks (Okabe & Yamada, 2001; Tao et al., 2015; Yamada & Thill, 2007). In addition, the planar K-function has also been used, but studies have shown that it tends to overestimate the aggregated tendency based on vehicle crash distribution (Lu & Chen, 2007). Furthermore, the network cross K-function and dual K-function methods were proposed to clearly depict co-locations in networks (Morioka et al., 2022; Okabe & Yamada, 2001). Recently, an adjusted network K-function has been proposed to intrinsically correct for the inhomogeneous geometry of the network (Ang et al., 2012). In these studies, network distances based on the shortest path distance, instead of the Euclidean distance, were used as the distance metric, while linear segments instead of planar circles were utilized as the statistical units (neighborhood), bringing the detection results closer to the actual situation.

The existing network K-functions are still based on the radius-based planar scale, leading to two challenges in detecting aggregation with high intensity in networks. First, the radius-based planar scale may affect the evaluation of point intensity. As shown in Figure 1, the total network lengths corresponding to the planar scale vary with the locations. The road network covered by the r-neighborhood is longer and has more points at the intersection (see Figure 1b) than the one at the straight road (see Figure 1a) at the planar scale of r. Thus, when using the network K-function, the points at this intersection will be identified as the aggregation at this scale, while the ones along the straight road will not have this “privilege,” although that aggregation has the higher intensity. Second, the L-function, which is extended from the K-function by normalizing with its expectation value under CSR, can significantly amplify the difference between observed data and CSR data, and is thus more sensitive to aggregation patterns and able to perform a more precise identification of aggregation size (Besag, 1977; Ehrlich et al., 2004; Kiskowski et al., 2009). However, the crucial expectation value of network K-function cannot be derived under the planar scale.

In this study, we proposed the h-neighborhood and corresponding K-function, called the length K-function, from the perspective of length-based network scale, which is more suitable for network space, thus overcoming the weakness of the existing network K-function. In addition, we extended the length K-function to the length L-function with a derived theoretical value (expectation value) of the K-function under CSR. The first derivative of the length L-function quantifies the aggregation scale and the local length L-function helps extract aggregations. We verified our method using synthetic case data simulating various aggregation types in road networks, thereby demonstrating the method’s superiority in identifying aggregations of high intensity as compared to the existing network K-function. We then validated our method by detecting the aggregation of taxi-trip pick-up point data in Beijing to examine taxi demand within the research area and understand residents’ demand for taxis.

2 | BASIC CONCEPTS

2.1 | Definition 1: Aggregation scale

According to its type, space can be measured by the planar scale and network scale. For point aggregation, the planar scale of space depicts its radius, which roughly signifies the average length from the ends to the center in networks. The network scale is the coverage length, namely the total length of the networks the aggregation occupies. The planar and network scales describe the range of the aggregation in two-dimensional planar space and one-dimensional network space, respectively. For example, for the aggregations shown in Figures 1a,b, both of the planar scales are r while the network scales are 2r and 6r, respectively.
2.2 | Definition 2: Road network distance

This refers to the shortest path length between two points in the network. Calculating this shortest path on a road map can be modeled as a special case of the shortest-path problem in undirected graphs, using Dijkstra's algorithm (Dijkstra, 1959), wherein vertices correspond to intersections and edges correspond to road segments, each weighted by the length of the road segment.

2.3 | Definition 3: The $h$-neighborhood

The total network length of a subset of network space centered on a certain point is $h$ (see Figure 2). It has two distinctive properties: first, the road network length of each branch in the $h$-neighborhood, namely a possible path from the center point to the end, is the same, and second, the total length of networks occupied by all the branches is exactly $h$. It should be emphasized here that, although individual branches of the $h$-neighborhood may overlap with each other owing to the configuration of network(s), the length of the overlapping part can only be counted once in the total length of $h$-neighborhood. Figure 2 shows examples of an $h$-neighborhood placed in four positions. Note that in Figure 2c, there are four branches in the neighborhood; however, three of the branches passing through the node have overlapping parts. Therefore, when calculating the total length of this neighborhood, the overlapping part, the length of which is $1/3h$, will only be calculated once.

FIGURE 2 Illustration of the $h$-neighborhood. (a–d) shows the $h$-neighborhood areas centered at different places in road networks; the neighborhood branches are depicted in red lines starting from the center: (a) an edge without a node in it, (b) a node without other nodes in it, (c) an edge with a node in it, and (d) an edge with a closed loop in it.
The key step in the L-function process is to calculate the number of points within an \( h \)-neighborhood (hereinafter also referred to as "neighborhood") centered on a certain point \( p_i \). However, it is difficult to determine the range of the road network covered by the neighborhood because, given a center point and the topology of the network, it is not easy to directly know which parts of branches will be reduplicated and thus, how to allocate the exact length of branches in the neighborhood, especially when the road network structure is complex. Using the \( k \)th-nearest neighbor distance of \( p_i \) (denoted as \( d_{i,k} \)), it is much easier to count target points without need of getting the exact range of the neighborhood. For each \( d_{i,k} \) (from \( k = 1 \) to \( k = n \), \( n \) is the number of points in the research area), we calculate the length of a neighborhood with branches as \( b = d_{i,k} \), denoted as \( L_{i,k} \). Once \( L_{i,k} \) exceeds \( h \), the number of points within the \( h \)-neighborhood (light blue area) is \( k \).

The key step in the L-function process is to calculate the number of points within an \( h \)-neighborhood (hereinafter also referred to as "neighborhood") centered on a certain point \( p_i \). However, it is difficult to determine the range of the road network covered by the neighborhood because, given a center point and the topology of the network, it is not easy to directly know which parts of branches will be reduplicated and thus, how to allocate the exact length of branches in the neighborhood, especially when the road network structure is complex. Using the \( k \)th-nearest neighbor distance of \( p_i \) (denoted as \( d_{i,k} \)), it is much easier to count target points without need of getting the exact range of the neighborhood. For each \( d_{i,k} \) (from \( k = 1 \) to \( k = n \), \( n \) is the number of points in the research area), we calculate the length of a neighborhood with branches as \( b = d_{i,k} \), denoted as \( L_{i,k} \). Once \( L_{i,k} \) exceeds \( h \), the number of points within the \( h \)-neighborhood is \( k \) (see Figure 3), because the nearest neighbor distance after this will bring the neighborhood length longer than \( h \). Although we depict the range of \( h \)-neighborhood in the illustration (light blue area in Figure 3), when implementing the algorithm, we only know the value of \( h \), but the exact range of the neighborhood is unknown.

### 2.4 | Definition 4: Point process intensity

This is the expected number of points per unit length of network, denoted by \( \lambda \), and is usually estimated by \( \hat{\lambda} = \frac{n}{L_A} \), where \( n \) is the number of points in the research area and \( L_A \) is the total length of the research network. When it is difficult to precisely identify the boundary of the research area in the real world with a point process, the intensity parameter can be estimated using the second-order properties of points (Pei et al., 2007, 2015); this is calculated through \( d_{i,1} \) (i.e., 1st-nearest neighbor distance to point \( p_i \)):

\[
\hat{\lambda} = \frac{1}{E(d_{i,1})} = \frac{n}{\sum_{i=1}^{n} d_{i,1}}
\]  

### 2.5 | Definition 5: CSR point processes in networks

During CSR point generation, points occur independently and completely randomly within the research network. CSR points can be represented by a homogeneous spatial Poisson process (Dixon, 2001), which is a random counting measurement method. For a subset \( s \) of the research network space, the number of points within \( s \), namely \( N_s \), follow a Poisson distribution, which is calculated using the following equation:

\[
P(N_s = n) = \frac{(\lambda L_s)^n}{n!} e^{-\lambda L_s}, (n = 0, 1, 2 \ldots)
\]
where \( \lambda \) is the point process intensity in Definition 4 and \( L_s \) is the size of \( s \), that is, the length that \( s \) occupies in the networks.

Our study is performed in the context of bounded networks; however, CSR assumes unbounded space (e.g., the Poisson point process). Therefore, edge effects inevitably appear. Although the edge effect is usually not as extreme in network spatial analysis as it is in planar spatial analysis (Okabe & Sugihara, 2012), in this study, to avoid its negative effects, all subsequent analyses take the measure of including some network segments outside the key networks (i.e., network segments where the experimental points are located) into the research scope. As a result, the edge effect or bias was weakened.

### 3 | LENGTH L-FUNCTION IN NETWORKS

In this section, we first introduce the network K-function using the radius-based \( r \)-neighborhood and its derivatives, and then describe the derivation of the length K-function using the length-based \( h \)-neighborhood, as well as its theoretical value under CSR. Then, the length L-function is developed based on the length K-function, depicting and amplifying the deviation from the CSR distribution. The derivatives of the length L-function are also illustrated.

#### 3.1 | Network K-function and its derivatives

Similar to Ripley’s K-function for point-pattern analysis in planar space, the network K-function (Okabe & Yamada, 2001) is defined as the expected number of additional points within the \( r \)-neighborhood of an arbitrary point (i.e., number of points with a distance \( r \) to an arbitrary point) normalized by the point process intensity:

\[
K(r) = \lambda^{-1} \sum \frac{1}{n} \sum \sigma_{ij}(r), (i, j = 1, 2, \ldots, n; i \neq j)
\]  

(3)

\[
\sigma_{ij}(r) = \begin{cases} 
1, & \text{point } j \text{ is inside } r\text{-neighborhood of point } i \\
0, & \text{point } j \text{ is outside } r\text{-neighborhood of point } i 
\end{cases}
\]  

(4)

where \( r \) is the designated planar scale, \( n \) is the number of points in the research area, and \( \lambda \) is the intensity of the point process.

The local network K-function for each point is defined as:

\[
K_i(r) = \lambda^{-1} \sum \frac{1}{n} \sum \sigma_{ij}(r), (j = 1, 2, \ldots, n; j \neq i)
\]  

(5)

where \( r, \lambda, n, \) and \( \sigma_{ij}(r) \) in Equation (5) have the same meanings as the corresponding expressions in Equation (3).

Due to the cumulative effect, the network K-function alone is unable to quantify the scale of aggregation, meaning that the values at large-scale aggregations are strongly affected by small-scale ones as they count all points at certain planar scales. To obtain a more appropriate determination of the scale of aggregation, the incremental network K-function was proposed, which examines the number of points by unit increments of planar scale \( r \) rather than the total number of points within \( r \)-neighborhood. The incremental network K-function \( K_{\text{incr}}(r_t) \) can be derived as the difference between the network K-function value at two consecutive scales (Yamada & Thill, 2007):

\[
K_{\text{incr}}(r_t) = K(r_t) - K(r_{t-1}) \quad t = 2, 3, \ldots,
\]

(6)

\[
K_{\text{incr}}(r_1) = K(r_1) \quad t = 1
\]

where \( K(r) \) is calculated as Equation (3) and \( r_t \) represents the \( t \)-th scale in ordered scales.
3.2 | Length L-function and its derivatives

The length K-function is defined as the expected number of additional points within the \( h \)-neighborhood of an arbitrary point normalized by the point process intensity. The derived formula for \( K(h) \) is as follows:

\[
K(h) = \lambda^{-1} \sum_{i} \sum_{j} \sigma_{ij}(h) \frac{1}{n}, (i,j = 1, 2, \ldots, n; i \neq j)
\]

(7)

\[
\sigma_{ij}(h) = \begin{cases} 
1, & \text{point } j \text{ is inside } h\text{-neighborhood of point } i \\
0, & \text{point } j \text{ is outside } h\text{-neighborhood of point } i 
\end{cases}
\]

(8)

where \( h \) is the designated network length, \( n \) is the number of points in the research area, and \( \lambda \) is the intensity of the point process.

In point-pattern analysis, CSR describes a point process whereby point events occur within a given study area in a completely random distribution. CSR is often applied as a standard or benchmark against which datasets are tested. In this study, inferences about CSR points assist in deriving the length L-function from the length K-function.

Using the aforementioned definition of the length K-function, the expected number of additional points within an \( h \)-neighborhood is \( \lambda K(h) \). Combining this with the properties of CSR points, namely that the expected number of additional points within an \( h \)-neighborhood is \( \lambda h \), we can conclude that \( \lambda K(h) = \lambda h \). Thus, the expectation for the length K-function (i.e., the theoretical value) with a homogeneous Poisson point process is:

\[
E(K(h)) = \frac{\lambda h}{\lambda} = h
\]

(9)

By normalizing the length K-function with its expected value, we can obtain the length L-function, the expectation of which is zero at any length (i.e., network scale) for CSR points. In this way, we can amplify and analyze the deviation from the CSR point using the following equation:

\[
L(h) = K(h) - h = \lambda^{-1} \sum_{i} \sum_{j} \sigma_{ij}(h) \frac{1}{n} - h, (i,j = 1, 2, \ldots, n; i \neq j)
\]

(10)

The local length L-function for each point is defined as:

\[
L_i(h) = K_i(h) - h = \lambda^{-1} \sum_{j} \sigma_{ij}(h) \frac{1}{n} - h, (j = 1, 2, \ldots, n; j \neq i)
\]

(11)

where \( h, \lambda, n, \) and \( \sigma_{ij}(h) \) in Equations (9)–(11) have the same meanings as the corresponding expressions in Equation (7).

To overcome the cumulative effect in the length L-function, the first derivative of the L-function, \( L'(h) \), is calculated to obtain more precise the network scale of aggregation (Kiskowski et al., 2009). Usually, \( L'(h) \) is obtained by: first, calculating the value of L-function at series of scales, second, fitting the curve of L-function and deriving it. In our work, we used the SciPy library to perform this step, which contains fundamental algorithms for scientific computing in Python.

4 | EXPERIMENT WITH SYNTHETIC DATA

This section describes the experiment aimed at testing the correctness and superiority of the length L-function. First, the theoretical values of the length K-function and length L-function under CSR were tested using Monte Carlo simulations. The detailed steps involved in detecting, quantifying, and extracting the aggregation are described. Then, we generated synthetic point datasets with different aggregations to test the capabilities of identifying aggregations. We compared the identification results of the existing network K-function and the proposed length L-function.
4.1 Monte Carlo tests of null models

Under the null hypothesis of CSR, the expectation of $K(h)$ is $h$, and the expectation of $L(h)$ is zero, which are theoretical value curves for each function. Monte Carlo simulations were used to test our hypothesis, wherein we used three types of datasets to simulate classical road network patterns: a grid pattern with 500 CSR points, a radial pattern with 300 CSR points, and a hybrid pattern with 300 CSR points; each dataset has different point process intensities in the network space. For each road network pattern, 99 simulations were performed to obtain the average value.

Figure 4 shows the data and results of the CSR points for each row corresponding to a network pattern. The dark blue line and dark purple lines with stars correspond to the theoretical $K(h)$ and $L(h)$ curves (expectation under CSR, i.e., $K(h) = h$ and $L(h) = 0$, respectively); the light blue and light purple lines correspond to the simulated means of $K(h)$ and $L(h)$ curves with 99 runs of CSR points. The blue and purple belts are the 95% confidence bands based on the Monte Carlo simulations. From Figure 4, we can conclude that both theoretical $K(h)$ and $L(h)$ curves fall within the 95% confidence intervals, and almost match the simulated means. Therefore, the theoretical null models of $K(h)$ and $L(h)$ are thus correct and can be used as benchmarks for determining point aggregation patterns.

4.2 Detecting and quantifying aggregation with the length $L$-function

To verify the superiority of our method, we compare the existing network K-function with the proposed length $L$-function. In this section, we first introduce the method of detecting the planar scale of the aggregation with the network K-function (hereinafter, NK) and the incremental NK, as well as detecting the network scale of the aggregation using the length $L$-function (hereinafter, LL) and the first derivative of LL. Then, the specific steps of extracting aggregations with the aid of local NK and local LL at detected scales are described. We designed two cases of synthetic data in road networks with a total length of 35.97 in hybrid pattern to test the methods.

4.2.1 Methodology

Both NK $K(r)$ and LL $L(h)$ indicate whether the observed pattern is aggregated or not at the scale of $r$ or $h$. For NK, if $K(r)$ for the observed data is greater than or equal to the upper benchmark of CSR points, the observed pattern is considered to exhibit significant aggregation at the scale of $r$. Note that the benchmark for NK should be the envelope stemmed from the Monte Carlo simulation with CSR points in the same networks as observed pattern. For LL, if $L(h)$ for the observed data is greater than or equal to the upper benchmark of CSR points, the observed pattern is considered to exhibit significant aggregation at the scale of $h$. The benchmark for LL should also be the envelope stemmed from the Monte Carlo simulation which has a theoretical value as $L(h) = 0$.

The derivatives of NK and LL are utilized to quantify the precise aggregation scales. For the incremental network K-function, that is, $K_{\text{incr}}(r)$, if a local maximum is attained at scale $[K_{\text{max}}^{\text{incr}}]$, and the corresponding value $K_{\text{incr}}(r = [K_{\text{max}}^{\text{incr}}])$ is above the upper envelope, which is also stemmed from Monte Carlo simulation with CSR points, the scale $[K_{\text{max}}^{\text{incr}}]$ is regarded as a planar scale of the aggregation. When using the length L-function, that is, $L(h)$, we calculate the first derivative of the L-function, $L'(h)$, and the scale that minimizes $L'(h)$, denoted as $[L'_{\text{min}}]$, thus $[L'_{\text{min}}]/2$ is a network scale of aggregation.

After obtaining the planar and network scales, local NK and local LL are calculated, respectively, to indicate the location and degree of the aggregation at the detected scales, helping identify hierarchical aggregations. First, we extract aggregations at each detected scale. At a certain scale, the points with top local function values (here we select the points with values in the top 25% of the entire value range) are the center points of the aggregations. Thus, the point possessing the top local function values at the detected aggregation scale is identified as the aggregation center and the points within the corresponding $r$-neighborhood or $h$-neighborhood of the center are included as part
FIGURE 4  Monte Carlo simulation of the length K-function and length L-function with road networks in various patterns: (a) Grid pattern; (b) radial pattern; and (c) hybrid pattern. Each row shows the data sample of that pattern, its theoretical value, and the simulated value with 99 runs of the length K-function and length L-function.
of the extracted aggregation. For the length L-function, this step is similar to the method of counting points in the \( h \)-neighborhood of a certain point, as was described earlier in Definition 3. That is, the \( k \)-th-nearest neighbor distance \( d_{c,k} \) of the aggregation center \( p_c \) is traversed, if the length of the neighborhood with the branch as \( d_{c,k} + 1 \) exceeds \( h \), the traversed neighbors, except \( p_{k+1} \), are counted into the aggregation.

Detected aggregations at all scales may have overlapping parts, for example, a subset of an aggregation can also be detected as the aggregation at a smaller scale because its intensity is also higher than average. Thus, considering that the target objective of our work is to extract the aggregation with the highest intensity, the steps for extracting the hierarchical aggregations are: First, include all possible aggregations at all scales into the aggregation set. Then, for each detected aggregation, compare it with those that overlap it, and reserve the one with higher intensity.

### 4.2.2 Case study using synthetic data

The two designed cases are shown in Figure 5. Case 1 is designed to test the ability of the length L-function in detecting a single aggregation, which contains a linear aggregation consisting of 120 points, with the planar scale of 0.6 and length scale of 1.2. Case 2 is designed to detect multiple aggregations, which contains two aggregations: a linear aggregation identical to that in Case 1 and a radial aggregation consisting of 360 points, with the planar scale of 0.6 and length scale of 3.6; the intensities of the two aggregations are the same. In addition, both cases contain 300 random points on the simulated road network.

The NK and LL results for Cases 1 and 2 are shown in Figure 6, in which we set the step for the scale as 0.1 for both methods. For Case 1, the aggregation pattern is detected by NK at scales less than \( r = 6 \) since the observed K-function is larger than the simulated K-function (stemmed from 99 simulations of 500 random points in the observed network) at these scales. For Case 2, the aggregation pattern is detected by NK at scales less than \( r = 7.3 \). The length L-function for both Cases 1 and 2 is above the benchmark generated from 99 simulated L-function of 500 random points in the observed network, indicating the significant clustering pattern at these scales. The detected aggregation scales using these two methods are represented by gray dotted lines in Figure 6. To extract the aggregations, we first calculate the local function value for all points at detected aggregation scales (i.e., Figure 7 shows the local NK values at \( r = 0.3, 0.6, 0.9 \) and the local LL value at \( h = 1.25, 1.8 \)). The reason that we ignore the planar scale of \( r = 0.1 \) is because: if the detected aggregation scale is too small, the aggregation it represents might be accidental and can be eliminated according to the actual situation. We also ignore the larger scale in both the NK and LL methods since the aggregations they represent may be the union of the independent aggregations. The extraction result is shown in Figure 7; the precisions were 0.94 and 0.93, and both recalls were 1.00, by NK and LL, respectively. In this

![Figure 5](image1.png)  
**Figure 5** Synthetic data in Cases 1 and 2. (a) Case 1 consists of a linear aggregation and 300 CSR points; and (b) Case 2 consists of the linear aggregation in Case 1, a radial aggregation, and 300 CSR points (note that the point process intensity of the linear aggregation is the same as that of the radial aggregation).
case, NK and LL have a similar ability to detect aggregation scales and extract aggregations. For Case 2, Figure 7 visualizes the local function values for all points at detected aggregation scales. NK only detects a planar scale at $r = 0.6$ (Figure 6) and the local NK shows that NK detected the radial aggregation while failed to identify the linear one. However, LL detected both the radial and the linear aggregations. The extraction result is shown in Figure 7. For the radial aggregation, the precisions were 0.61 and 0.58, and the recalls were 0.99 and 0.75, by NK and LL, respectively. For the linear aggregation, the precision was 0.90 and the recall was 0.99 by LL.

The reason why the NK method failed to detect the linear aggregation in Case 2 is that the $r$-neighborhood centered around the crossings possessed more points than around the linear road, thus the points along the straight roads possessed small local NK values, which could be verified from the local NK at $r = 0.6$ for Case 2 (see Figure 7). Thus, the linear aggregation was unable to be detected in the NK method, even if it had the same intensity as the radial one. In addition, we found that the setting of the step in the scale had a great impact on the result outcome using the incremental K-function. For example, if we were to set the step to 0.05, the result will be completely different (i.e., contain too many maximums) from that obtained using the current step of 0.1. However, our LL method only needs to ensure that the step is able to fit the curve of the L-function and facilitate the calculation of the corresponding first derivative. Namely, when using a relatively small step, setting a different step will not affect the detection results in the LL method.

5 | CASE STUDY USING REAL-WORLD DATA

Detecting the aggregations of taxi pick-up locations and quantifying their scales in the road network helps to analyze the range and degree of traffic anomalies. Combined with the built environment around the aggregation, better urban
planning could be made to alleviate traffic congestion. For instance, auxiliary roads can be established around those aggregations with a relatively large scale and which are near to commercial areas, and for those aggregations with relatively small scales and which are close to residential areas, appropriate boarding points could be assigned.

To testify the efficiency of the length L-function in detecting aggregations, we conducted experiments in the Zhongguancun Area of Beijing with corresponding taxi pick-up locations. In the Zhongguancun Area, a prosperous area in the Haidian district, the demand for taxis is strong and scattered. The case study used taxi GPS trajectory data from October 20 to 24 2014, which were workdays in Beijing. After matching the data to the road network, we selected the pick-up locations of each trip as the experimental points and extracted the ones during the morning peak (07:00–10:00) and evening peak (17:00–20:00). The detailed information and distribution of experimental data are shown in Figure 8.

For these two datasets, we set the step for the scale as 50 m for both methods (see Figure 9). It is hard to judge the aggregation pattern at the planar scales using the network K-function since there is no obvious deviation...
between the observed K-function and the simulated K-function (stemmed from 99 simulations of 500 random points in the observed network). However, the length L-function value for both cases is above the benchmark generated from 99 simulations of 500 random points in the observed network, indicating the significant aggregation pattern at these network scales.
The incremental NK and the first derivative of LL detect the aggregation scale at the planar scale and network scale, respectively (Figure 9). The incremental NK detects the aggregation scale at \( r = 50, 1200 \) for the morning peak and \( r = 150, 1200 \) for the evening peak. The first derivative of LL detects the aggregation scale at \( h = 100, 675, 1100 \) for the morning peak and \( h = 50, 275, 600 \) for the evening peak. Figure 10 shows the local function value at these detected scales, indicating the location of the aggregations at corresponding scales. It can be concluded that the NK method is not sensitive to aggregations as it only detects a small aggregation scale for both datasets (note that \( r = 1200 \) for the morning and evening peak is omitted when extracting aggregations since the aggregation at this scale would contain multiple independent aggregations at small scales, as shown in the local NK at \( r = 1200 \) in Figure 10), while the LL method identifies different scales of single independent aggregations, which may result in more precise extraction results.

Aggregations extracted with the aid of local functions are shown in Figure 11. For the morning peak in the Zhongguancun Area, both NK and LL detected aggregations around the residential locations. The NK method only detected three aggregations near communities, that is, the Feida, Wangquanzhuang, and Zhongguancun communities, while the LL method detected an additional two aggregations around communities, namely the Zhichunli and Shuangyushu communities, which also possess high intensities compared with points in the dataset. For the evening peak in the Zhongguancun Area, all detected aggregations are located around places of work and shopping malls. However, the results obtained by the NK and LL methods differ in that NK detects two aggregations centered at crossings (aggregations ③ and ④ by NK) while LL detects another two aggregations around relatively simple road networks (aggregations ③ and ④ by LL). The aggregations identified by LL are more convincing for two reasons: first, aggregations ③ and ④ by LL possess higher intensities, especially the one near Dangdai Shopping Mall; second,
the identification of aggregations ③ and ④ by NK is the result of road network gathering, rather than the result of higher intensities there, which can be verified from the local LL for evening peak in Figure 10, as local LL values are low at the positions of aggregations ③ and ④ by NK. To sum up, in this case study, the NK method failed to detect aggregations that centered at simple road networks even though they had high intensities, while the LL method could identify those aggregations of high intensities. This demonstrates the ability of the LL method to help reveal important implications of spatial anomalies in road networks.

6 | CONCLUSION

Detecting and quantifying aggregation in networks are significant to urban applications, but methods which are extended from those designed for planar space still use the radius-based planar scale, which is ill-suited to

FIGURE 11 Detected aggregations and the corresponding scales. Aggregations in Zhongguancun Area during the morning peak and evening peak. The aggregations are marked with the serial number and their intensities, which were calculated manually.
network space and leads to the poor estimation of aggregations. To precisely identify the point aggregations of high intensity in various networks, we derived the L-function in the one-dimensional network space, called the length L-function, which adopts the length-based network scale, overcoming the disadvantages of the traditional network K-function.

A synthetic data test and case study assessing Beijing taxi’s pick-up point data were performed, showing the applicability of our method. The proposed length L-function detected aggregations that the traditional network K-function failed to detect even though they had high intensity, demonstrating that the proposed method is more suitable for aggregation detection in the road network space and thus more adaptable for research on urban environments.

The proposed length L-function method contributes to the discipline in that it analyses the point pattern in road networks from the length-based network scale, rather than the conventional radius-based planar scale; therefore, it can detect point aggregations with higher intensity regarding one-dimensional road networks other than the two-dimensional space. It conquers the problem of existing methods in which the detected aggregation is affected by the topological complexity of road networks, thereby eliminating the “false aggregation” caused by the aggregation of road networks rather than point data itself.

This work is a significant advance because it derives the L-function in a network, which is unable to be achieved with the existing network K-function. The proposed length L-function method is more sensitive to aggregation patterns and its first derivative quantifies the aggregation scale automatically (i.e., there is no need to set the step for scales) and more accurately. These qualities are beneficial since it is sometimes difficult to compare the difference between observed and random data in the network K-function; moreover, the detected scale of the incremental K-function largely depends on the manually designated scale step.

Future research directions should include several aspects. First, a study should be conducted to extend other related spatial statistical methods designed for network space with the aid of an h-neighborhood. Second, future studies should focus on point aggregation in road networks as well as point aggregation in other types of networks, for example, flight and social networks. Third, the time dimension should be considered in the existing methods, so as to carry out spatiotemporal clustering in the context of networks.

ACKNOWLEDGMENTS
This work was supported by the National Natural Science Foundation of China (Grant No. 42071436) and the Innovation Project of LREIS (Grant No. KPI002).

CONFLICT OF INTEREST STATEMENT
The authors declare no competing interests.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID
Zidong Fang https://orcid.org/0000-0002-1902-4446
Ci Song https://orcid.org/0000-0003-2146-6259
Hua Shu https://orcid.org/0000-0001-7694-564X
Tianyu Liu https://orcid.org/0000-0001-5045-8261
Xi Wang https://orcid.org/0000-0003-1859-9958
Xiao Chen https://orcid.org/0000-0001-5278-858X
Xiaorui Yan https://orcid.org/0000-0002-4174-0561
Tao Pei https://orcid.org/0000-0002-5311-8761
REFERENCES

Ang, Q. W., Baddeley, A., & Nair, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. Scandinavian Journal of Statistics, 39(4), 591–617. https://doi.org/10.1111/j.1467-9469.2011.00752.x

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., & Davies, T. M. (2021). Analysing point patterns on networks—A review. Spatial Statistics, 42, 100435. https://doi.org/10.1016/j.spa.2020.100435

Besag, J. E. (1977). Comment on ‘modeling spatial patterns’ by BD Ripley. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39, 193–195. https://doi.org/10.1111/j.2517-6161.1977.tb01616.x

Black, W. R. (1992). Network autocorrelation in transport network and flow systems. Geographical Analysis, 24(3), 207–222. https://doi.org/10.1111/j.1538-4632.1992.tb00262.x

Deng, M., Yang, X., Shi, Y., Gong, J., Liu, Y., & Liu, H. (2019). A density-based approach for detecting network-constrained clusters in spatial point events. International Journal of Geographical Information Science, 33(3), 466–488. https://doi.org/10.1080/13658816.2018.1541177

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. Numerische Mathematik, 1(1), 269–271. https://doi.org/10.1007/BF01386390

Dixon, P. M. (2001). Ripley’s K function. Encyclopedia of Environmetrics, 3, 1796–1803. https://doi.org/10.1002/9780470057339.var046

Ehrlich, M., Boll, W., van Oijen, A., Hariharan, R., Chandran, K., Nibert, M. L., & Kirchhausen, T. (2004). Endocytosis by random initiation and stabilization of clathrin-coated pits. Cell, 118(5), 591–605. https://doi.org/10.1016/j.cell.2004.08.017

Kiskowski, M. A., Hancock, J. F., & Kenworthy, A. K. (2009). On the use of Ripley’s K-function and its derivatives to analyze domain size. Biophysical Journal, 97(4), 1095–1103. https://doi.org/10.1016/j.bpj.2009.05.039

Lu, Y., & Chen, X. (2007). On the false alarm of planar K-function when analyzing urban crime distributed along streets. Social Science Research, 36(2), 611–632. https://doi.org/10.1016/j.ssresearch.2006.05.003

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17–23. https://doi.org/10.2307/2332142

Morioka, W., Kwan, M.-P., Okabe, A., & McLafferty, S. L. (2022). A statistical method for analyzing agglomeration zones of co-location between diverse facilities on a street network. Transactions in GIS, 26(6), 2536–2557. https://doi.org/10.1111/tgis.12969

Nie, K., Wang, Z., Du, Q., Ren, F., & Tian, Q. (2015). A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: A case study from Wuhan, China. Sustainability, 7(3), 2662–2677. https://doi.org/10.3390/su7032662

Okabe, A., Okunuki, K.-I., & Shiode, S. (2006). SANET: A toolbox for spatial analysis on a network. Geographical Analysis, 38(1), 57–66. https://doi.org/10.1111/j.0016-7363.2005.00674.x

Okabe, A., Satoh, T., & Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. International Journal of Geographical Information Science, 23(1), 7–32. https://doi.org/10.1080/136588110802475491

Okabe, A., & Sugihara, K. (2012). Spatial analysis along networks: Statistical and computational methods. John Wiley & Sons. https://doi.org/10.1002/9781118967101

O’Sullivan, D., & Unwin, D. (2003). Geographic information analysis. John Wiley & Sons. https://doi.org/10.1111/j.1536-5471.1997.tb00448.x

O’Sullivan, D., & Unwin, D. (2003). Geographic information analysis. John Wiley & Sons. https://doi.org/10.1111/j.1536-5471.1997.tb00448.x

Pei, T., Shen, J., & Zhou, C. (2015). Density-based clustering for data containing two types of points. International Journal of Geographical Information Science, 29(2), 175–193. https://doi.org/10.1080/13658816.2014.955027

Pei, T., Zhu, A. X., Zhou, C., Li, B., & Qin, C. (2007). Delineation of support domain of feature in the presence of noise. Computers & Geosciences, 33(7), 952–965. https://doi.org/10.1016/j.cageo.2006.11.010

Ripley, B. D. (1976). The second-order analysis of stationary point processes. Journal of Applied Probability, 13(2), 255–266. https://doi.org/10.2307/3212829

Ripley, B. D. (1977). Modelling spatial patterns. Journal of the Royal Statistical Society: Series B (Methodological), 39(2), 172–192. https://doi.org/10.1111/j.2517-6161.1977.tb01615.x

Shiode, S. (2008). Analysis of a distribution of point events using the network-based quadrat method. Geographical Analysis, 40(4), 380–400. https://doi.org/10.1111/j.0016-7363.2008.00735.x

Spooner, P. G., Lunt, I. D., Okabe, A., & Shiode, S. (2004). Spatial analysis of roadside acacia populations on a road network using the network K-function. Landscape Ecology, 19(5), 491–499. https://doi.org/10.1023/B:LAND.0000036114.32418.d4
Tao, R., Thill, J.-C., & Yamada, I. (2015). Detecting clustering scales with the incremental K-function: Comparison tests on actual and simulated geospatial datasets. In V. Popovich, C. Claramunt, M. Schrenk, K. Korolenko, & J. Gensel (Eds.), Information fusion and geographic information systems (IF&GIS’ 2015): Deep virtualization for mobile GIS (pp. 93–107). Springer. https://doi.org/10.1007/978-3-319-16667-4_6

Xie, Z., & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. Computers, Environment and Urban Systems, 32(5), 396–406. https://doi.org/10.1016/j.compenvurbsys.2008.05.001

Xie, Z., & Yan, J. (2013). Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: An integrated approach. Journal of Transport Geography, 31, 64–71. https://doi.org/10.1016/j.jtrangeo.2013.05.009

Yamada, I., & Thill, J. C. (2007). Local indicators of network-constrained clusters in spatial point patterns. Geographical Analysis, 39(3), 268–292. https://doi.org/10.1111/j.1538-4632.2007.00704.x

How to cite this article: Fang, Z., Song, C., Shu, H., Chen, J., Liu, T., Wang, X., Chen, X., Yan, X., & Pei, T. (2023). Length L-function for network-constrained point data. Transactions in GIS, 27, 476–493. https://doi.org/10.1111/tgis.13035