INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
VOL. 30, NO. 9, 1717–1737
http://dx.doi.org/10.1080/13658816.2015.1119279

A network Kernel Density Estimation for linear features in space–time analysis of big trace data

Luliang Tang[a], Zihan Kan[a], Xia Zhang[b], Fei Sun[a], Xue Yang[a] and Qingquan Li[a]

[a]State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China; [b]School of Urban Design, Wuhan University, Wuhan, China

ABSTRACT

Kernel Density Estimation (KDE) is an important approach to analyse spatial distribution of point features and linear features over 2-D planar space. Some network-based KDE methods have been developed in recent years, which focus on estimating density distribution of point events over 1-D network space. However, the existing KDE methods are not appropriate for analysing the distribution characteristics of certain kind of features or events, such as traffic jams, queue at intersections and taxi carrying passenger events. These events occur and distribute in 1-D road network space, and present a continuous linear distribution along network. This paper presents a novel Network Kernel Density Estimation method for Linear features (NKDE-L) to analyse the space–time distribution characteristics of linear features over 1-D network space. We first analyse the density distribution of each linear feature along networks, then estimate the density distribution for the whole network space in terms of the network distance and network topology. In the case study, we apply the NKDE-L to analyse the space–time dynamics of taxis' pick-up events, with real road network and taxi trace data in Wuhan. Taxis' pick-up events are defined and extracted as linear events (*LE*) in this paper. We first conduct a space–time statistics of pick-up *LE* in different temporal granularities. Then we analyse the space–time density distribution of the pick-up events in the road network using the NKDE-L, and uncover some dynamic patterns of people's activities and traffic condition. In addition, we compare the NKDE-L with quadrat method and planar KDE. The comparison results prove the advantages of the NKDE-L in analysing spatial distribution patterns of linear features in network space.

ARTICLE HISTORY
Received 30 June 2015
Accepted 4 November 2015

KEYWORDS
Kernel Density Estimation (KDE); network space; linear features; space–time analysis; pick-up events; big trace data

1. Introduction

In the real world, there is a kind of events occurring in road network space, presenting a linear continuous distribution along road network between its start and end points. We name this kind of events – linear events (*LE*), such as traffic congestion events, queuing up at intersections and passenger-carrying events. Studying the distribution characteristics and patterns of these events are beneficial for traffic control optimization and travel efficiency improvement. Existing approaches of analysing linear features aim to

CONTACT Zihan Kan kzh@whu.edu.cn

study the distribution over homogeneous 2-D space, ignoring the truth that the occurring context and distributing space of some *LE* are the inhomogeneous 1-D network space. So developing a spatial analysis method to analyse the distribution of *LE* in network space is important and necessary.

Many approaches for analysing spatial distribution patterns for point events have been presented and developed in the past decades. There are mainly two categories of Point Pattern Analysis (PPA) methods. The first one is first-order effect based, analysing point features' aggregation characteristics, such as quadrat method, and Kernel Density Estimation (KDE). The other is second-order effected based, aiming at examining the correlation and independency of point features of the same or different kinds, such as *K*-function and the nearest neighbour methods. Among the first-order methods, quadrat method divides the study area into samples of the same size, and describes spatial distribution patterns with the density of each sample. KDE is a well-known non-parameter method for analysing underlying aggregation effects of point events, proposed by Parsen in 1956 and 1962 (Silverman 1986, Bailey and Gatrell 1995), respectively. Based on Tobler's First Law of Geography, which is 'all attribute values on a geographic surface are related to each other, but closer values are more strongly related than more distant ones', KDE analyses the aggregation properties of point features by producing a smooth density surface. KDE is most developed for detecting hotspots of point events, which is widely applied in criminology (Anselin *et al*. 2000), economics (Lahr *et al*. 2014), traffic accidents detections (Erdogan *et al*. 2008), traffic hazard intensity analysis (Ha and Thill 2011) *etc*. Standard KDE takes Euclidean distance as spatial measure, and estimates the spatial distribution of point features over 2-D planar space. However, the occurrences and distributions of many events are constrained by 1-D road network configuration, in which situation the assumption of uniformity 2-D space is too strong (Miller 1999).

For adapting standard KDE to network context, one development is to take network distance (the shortest path distance) instead of Euclidean distance, known as network KDE. The network KDE can be classified into two categories: 2-D methods and 1-D methods. 2-D methods constrain the space of density estimation within a certain range by network distance, but the estimation result is still over 2-D space. For example, Burruso presents a Network Density Estimation (NDE) method, which obtains a polygon search region instead of the whole planar space based on network distance (Borruso 2005, 2008), but the estimation context is still a 2-D region. Different from 2-D methods, 1-D methods constrain the density estimation result into 1-D linear space. For instance, Flahaut *et al*. (2003) identify the concentration of accidents (black zones) along road network using KDE method, but their estimation result is on a single road, which means network topology is not considered. Xie and Yan (2008) take network linear unit (named *lixel* in their paper) as the basic unit to estimate density within the network; Okabe *et al*. (2009) argue that the network KDE proposed by Xie and Yan (2008) overestimates the densities around nodes, then they present two network kernel functions, named discontinuous kernel function and continuous kernel function, and prove the functions' unbiasedness around nodes in network space; Li *et al*. (2011b) analyse the accessibility of POIs in urban road network using Borruso's 2-D method (2008) and Xie and Yan's 1-D method (2008). The results of their paper show that the two methods perform much differently, and the 1-D method is more accurate in representing the network accessibility while the calculation cost is high.

The KDE methods have been widely developed and examined in the studies of point features' spatial distribution both in planar space and network space. However, the methods for analysing spatial distribution patterns of linear features remain few. So far the linear feature pattern analysis methods focus on the linear features' distribution over 2-D homogeneous space. The existing approach to analyse the distribution pattern of linear features mainly fall into two categories: one is to transform spatial distribution of lines into that of points. For example, Borruso (2003) approximates the density of road network with the density of cross nodes in road network; Worton (1989) studies the wild animals' tracks and home ranges by producing a density surface for their track points; Downs (2010) integrates the traditional KDE and a geo-ellipse to estimate spatial density of adjacent control points in a moving object's path. Later, the authors use a potential path tree to estimate space–time density of GPS trace data, by calculating the space–time potential of a moving vehicle between each two control points (Downs and Horner 2012). Recently, they propose probabilistic space–time prisms that are used to analyse animals' movement trace (Downs *et al*. 2014). Timothée *et al*. (2010) calculate the network density of street centrality by deducing each network edge to its midpoint. The other kind of linear feature pattern analysis methods is to simply extend the planar PPA methods to lines, such as the planar KDE or *K*-function for linear features, which are available in ESRI's ArcGIS software. Using the planar KDE for lines, Cai *et al*. (2013) and Ying *et al*. (2014) study the distribution of road network over the planar space; Scheepens *et al*. (2011) and Lee and Hahn (2014) extend the planar KDE for lines from 2-D planar space to 3-D stereo space to estimate the space–time density of trajectories (not constrained in road network), where the cross section reflects the spatial density distribution of trajectories at a certain timestamp, and the vertical section reflects the temporal density distribution of trajectories in a certain range.

Some recent developments of KDE are shown in Table 1, which indicates that an ideal method to analyse the spatial distribution characteristic of linear features is still in deficiency. Existing methods for studying the spatial distribution of linear features are all based on homogeneous 2-D or 3-D space, without considering constrains of road network configuration and network direction to some kinds of *LE*.

This paper presents a novel Network KDE for Linear events (NKDE-L) to analyse spatial distribution patterns of linear features in network space. We test the NKDE-L on the dynamics of taxis' pick-up events, and analyse the spatial distribution of them over 1-D road network space. Taxis' pick-up activities reflect people's demands for taxis and hot zones of people's activities in cities. We study the space–time distribution of taxis' pick-

Table 1. Developments of KDE method and research topics in recent years.

	2-D Homogeneous Planar Space	1-D Inhomogeneous Network Space
KDE for Point Features	Anselin *et al*. (2000), criminology Erdogan *et al*. (2008), traffic accidents Downs (2010), space time points Ha and Thill (2011), traffic hazard intensity Lahr *et al*. (2014), economic	Flahaut *et al*. (2003), accidents in one road Borruso (2005), nodes of road network Borruso (2008), economic entity Xie and Yan (2008), traffic accidents Okabe *et al*. (2009), traffic accidents Timothée *et al*. (2010), economic activity
KDE for Linear Features	Cai *et al*. (2013), density of road network Lee and Hahn (2014), density of trajectories	This paper research, such as taxis' pick-up *LE*

up events, which contributes to understanding the urban dynamics, optimizing the urban transportation resources allocation and improving the urban traffic efficiency. In literature, the space–time distribution of taxis' pick-up or drop-off events is widely uncovered mainly for two purposes: (1) to analyse passenger-finding-strategies and make recommendation for empty taxi drivers; (2) to detect hotspots in city. Among the former studies, Lee *et al*. (2008) design a location recommendation service for empty taxis based on pick-up data, with a clustering approach; Li *et al*. (2011a) extract pick-up and drop-off events from each GPS trace, and study the taxi drivers' behaviours before picking up and after dropping off passengers, to provide taxi drivers correct and efficient driving strategies. Liu *et al*. (2010) rank taxi drivers by their daily income, and analyse spatial distribution of pick-up points of top drivers. Among the latter studies, Giannotti *et al*. (2011) cluster the destinations of trips starting from city centre, obtain three origin-destination patterns, and find out the most popular itineraries and destinations. Li *et al*. (2012) define hotspots as areas where pick-up and drop-off events occur frequently, then characterize spatial mobility patterns of city by analysing people's activities in hotspots.

Taxis' pick-up events, which are usually treated as point events, are defined and extracted as the form of *LE* from taxis' GPS traces in this paper (the reasons and details are illustrated in Section 2). We apply the NKDE-L to analyse the space–time dynamic of taxis' pick-up *LE*, with real road network and taxi trace data in Wuhan. We first describe a time tuple for expressing various temporal granularities, and conduct a space–time statistics of pick-up *LE* in different temporal granularities. Then we analyse the space–time distribution of the pick-up events in the road network of study area, using the NKDE-L, and uncover the dynamic patterns of people's travel and traffic condition. Finally, we compare the NKDE-L with quadrat method and planar KDE. The comparison results prove the advantages of the NKDE-L in analysing spatial distribution patterns of linear features in road network.

The remainder of this paper is organized as follows. Section 2 defines *LE* in network space and explains the linear form of taxis' pick-up events. The NKDE-L is described in Section 3, where the details of the proposed method and its algorithm implementation are discussed. Section 4 presents a case study of Wuhan road network and taxis' pick-up *LE*, in which the NKDE-L is applied to analyse the space–time dynamics of taxis' pick-up events. Comparison results of the NKDE-L method with quadrat method and planar KDE are also included. Discussions and conclusion are shown in Section 5.

2. Definition and description of LE

In real world, there is a kind of events which occur and present a linear continuous distribution in network space, with start and end points. We name this kind of events *LE*. *LE* can be expressed by

$$LE = \{ID, S, E, L, T_s, T_e\}$$

where *ID* is the ID number of *LE, S* and *E* are the start and end points of *LE*, respectively. *L* denotes the network space where *LE* occurs and distributes, which is represented by a subset of the network $N : L \in N$. T_s and T_e are the start and end time of *LE*. The representation of *LE* in road network space is shown in Figure 1.

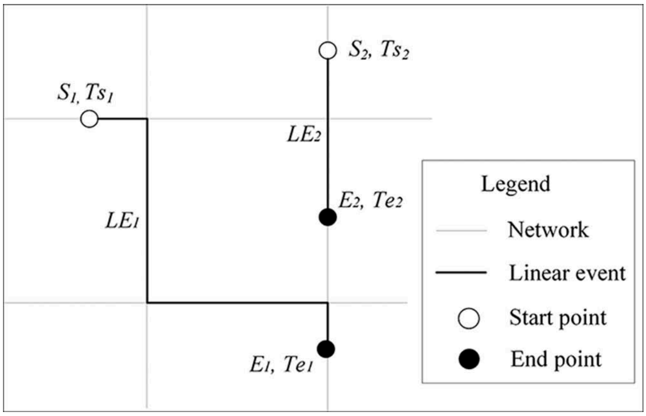


Figure 1. Representation of *LE* in network space.

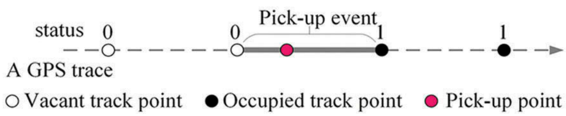


Figure 2. A taxi pick-up *LE*.

For taxis' pick-up events, they are analysed as point events from taxis' GPS traces in most cases. For example, Castro *et al.* (2013) split each taxi's continuous records up into vacant trajectories and occupied trajectories according to the status information (the status is 0 when the taxi is vacant, and the status is 1 when the taxi is occupied). The authors define the first and the last point of each occupied trajectory as pick-up point and drop-off point.

Although a pick-up event takes place at an actual location where the status of taxi changes from '0' to '1', the exact location is not available from the given data, because the value '0' or '1' only reflects the status of a taxi at the moment when data is collected. Taxi trace data can only indicate the location of a passenger instead of the actual place of his/her activity (Gong *et al.* 2015). We can only ensure the pick-up event happens between the locations where the adjacent statuses are '0' and '1', respectively, instead of knowing the exact location where the event takes place. Based on the above analysis, the pick-up events can be defined and extracted as *LE* in road network. The start and end point of a pick-up *LE* are the adjacent points in a GPS trace, with status of '0' and '1', respectively, as Figure 2 shows.

3. Network Kernel Density Estimation for Linear features (NKDE-L)

This paper first abstracts the taxis' pick-up events in reality into linear features in network space, then proposes the NKDE-L by improving and extending the standard planar KDE to network space. The road network in this paper is network topology consists of nodes and links with length attribute. The complex road elements (Li *et al.* 2004) inside a road (such as road centreline, road outline) are not considered in this paper. To introduce the proposed NKDE-L method, this paper first analyses the density distribution of each linear feature along networks, then estimates the density distribution for the whole network space in terms of the network distance and network topology. Considering the inhomogeneous characteristic and topological direction of network, the NKDE-L improves and extends the standard planar KDE in two aspects:

(1) For the density's extension directions, the NKDE-L improves 'homogeneous planar extension' in standard planar KDE to 'inhomogeneous network extension', and improves 'Euclidean distance decay effect' in planar KDE to 'network distance decay effect'.
(2) For the density calculation of linear features, the NKDE-L takes orientation of linear features and the topological connectivity of network into consideration. The NKDE-L also considers the particularity of estimating density at nodes in network, and guarantees the unbiasedness of density estimation at nodes.

3.1 *Density distribution of a single linear feature in network space*

The NKDE-L is based on the standard planar KDE for point features. The planar KDE produces density distribution of each point feature through considering the 'distance decay effect'. It estimates density distribution of a point feature by producing a smooth surface over the planar space, as is shown in Figure 3.

Here $f(x)$ is the density distribution function of a single point i, which denotes the density influence of point i at location x within the range of distance threshold r. Figure 3 shows that the value of $f(x)$ decreases with the increase of distance from i to x, and it comes to zero where distance between i and x equals or more than r. Function $f(x)$ is determined by Equation (1).

$$f(x) = k\left(\frac{x - s_i}{r}\right) \tag{1}$$

where s_i is location of point i, and $x - s_i$ is the Euclidean distance between x and i, r is the distance threshold, which is also called search bandwidth, and k is the distance decay function, which is also called kernel function. The form of function k includes Gaussian function, quadratic function, quartic function *etc*. Many studies show that the effect of kernel function's form on the density estimation is not significant when the

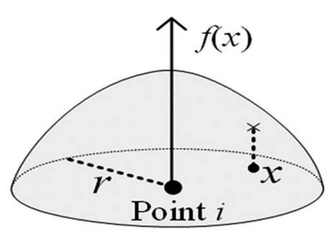


Figure 3. Standard planar KDE for point features.

search bandwidth r is fixed (Silverman 1986, Bailey and Gatrell 1995, O'Sullivan and Wong 2007).

Similar to planar KDE for point features, planar KDE for linear features covers a kernel surface over the linear feature within the Euclidean distance threshold r, representing the 'distance decay effect' of linear features' density distributions over the homogeneous plane, as is shown in Figure 4(a). The planar KDE for point or linear features are both based on homogeneous distance, producing a homogeneous density surface over the point or linear features. However, the distribution of linear features in network space is constrained by the network configuration and direction, so effective range of density estimation for linear features in network space should be limited in 1-D network space.

Based on the aforementioned analysis, NKDE-L takes network distance as spatial measure to estimate the density distribution of linear features in network space. The NKDE-L covers a smooth curve over each linear feature according to the 'network distance decay effect', as shown in Figure 4(b), where $f(x)$ is the density distribution function of a single linear feature l in network space.

The density distribution function $f(x)$ for linear features can be deduced from network density distribution function of point features. First, we divide the linear feature l into infinite small segments dl, which can be regarded as a point in l. Then the density distribution of a linear feature l can be considered as the integral of density distributions of all points composing l. From Equation (1), we know that the density distribution of a single point s is determined by the kernel function $k(s - x/r)$, so the density distribution of l in network space can be expressed mathematically as the integral of kernel function $k(s - x/r)$ on the linear feature l, by moving $k(s - x/r)$ from the start point S to end point E of l. As a result, the value of $f(x)$ at location x is the area enclosed by $k(s - x/r)$ and l. In Figure 5, the bold black line is linear feature l, with its start point S and end point E. The red curve is the resulting density distribution function $f(x)$. For example, the sizes of

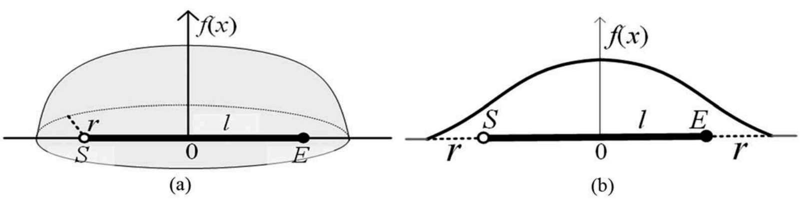


Figure 4. (a) Kernel surface in planar KDE and (b) kernel curve in NKDE-L. (a) Kernel surface of planar KDE and (b) kernel curve of NKDE-L.

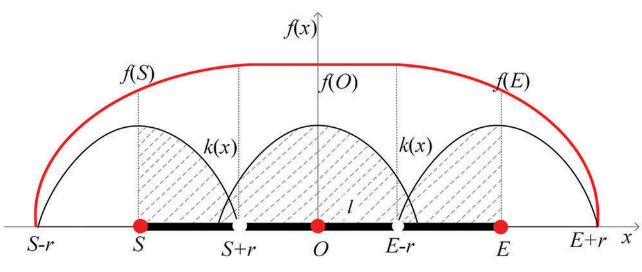


Figure 5. Kernel density distribution for linear feature in network space.

shaded areas are the values of the density distribution functions at locations *S, O* and *E*, respectively. The upper red curve *f*(*x*) is the integral of such little curves of all points in *SE*, from the start point *S* to the end point *E*.

The form of density distribution function *f*(*x*) in Figure 5 indicates that the density of a single linear feature *l* at location *x* comes greater with *x* approaching to the centre point of *l* (point O in Figure 5). The density reaches to the maximum at locations where the distance to *S* or *E* equals to or more than *r* inside *l* (between $S + r$ and $E - r$ in Figure 5). Inversely, the density of *l* decreases with the distance to the centre point increase, and decays to zero where the distance to *S* or *E* equals to or more than *r* outside *l* (less than $S - r$ or greater than $E + r$ in Figure 5).

So the density distribution function *f*(*x*) of a linear feature *l* is as Equation (2) shows:

$$f(x) = \frac{1}{r}\int_{l} k\left(\frac{s-x}{r}\right)\mathrm{d}l \tag{2}$$

This paper chooses quadratic function as kernel function, which is given by Equation (3).

$$k\left(\frac{s-x}{r}\right) = \frac{3}{4}\left[1 - \left(\frac{s-x}{r}\right)^2\right] \tag{3}$$

In Equations (2) and (3), *l* is the linear feature, *s* denotes the location of each single point (d*l*) that composes linear feature *l*, and *x* is the arbitrary location within the range of distance threshold *r*.

3.2 *Density distribution of linear features considering network topological relationship*

The existing spatial distribution pattern analysis methods in road network space assume that every road between intersections is bi-directional (Xie and Yan 2008, Okabe *et al.* 2009). However, this assumption does not hold for real *LE* and road network. First, the *LE* happens and distributes in road network with its own direction. Second, the road network presents some connectivity constrains, such as redirection limit or one-way limit, resulting in various distributions of *LE* in different directions of the same road or different roads. So the constrain of topological direction in road network should be considered when the density distribution of *LE* in road network is estimated. In this section, we first discuss density estimation in network links, then illustrate the difference of density estimation around network nodes.

3.2.1 *Density distribution of linear features in network links*

For estimating the density distribution in the network, this paper searches the linear features within the range of network search bandwidth *r* of location *x*, while considering the network topological relationship. The NKDE-L differs from the planar KDE in two aspects, which is illustrated in Figure 6.

(1) Network space and network distance *r* are used as context and spatial measure, respectively. In Figure 6, the grey round shaped region with Euclidean radius of *r* is the search region in planar KDE, and the linear shaped region with network

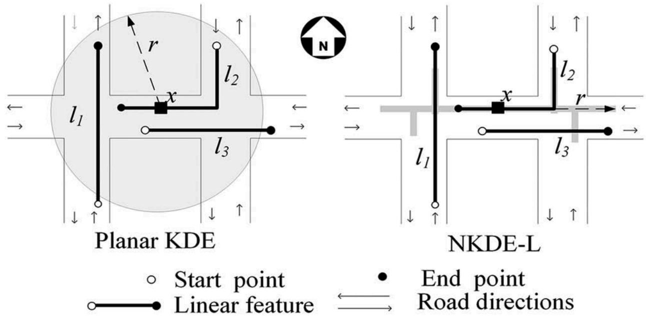


Figure 6. The different search region in Planar KDE and NKDE-L.

radius of r is the search ranges in NKDE-L. Comparing to the network distance and linear search region in NKDE-L, the 2-D Euclidean search bandwidth and round shaped region in the planar KDE lead to more linear features considered when estimating the density at location x.

(2) Connectivity of network and directions of linear features are taken into consideration. In Figure 6, linear feature l_3 is in the opposite direction of x, so it will not be considered in NKDE-L when estimating the density at x. While in planar KDE, all linear features falling into the round shaped search region will be considered.

The density at location x in network links is the accumulation of the density distributions for all linear features within the search bandwidth r of location x. Because NKDE-L divides a linear feature into infinite small segments, it can be a whole or part of linear feature within the search bandwidth r. So the density estimation result $LD(x)$ at location x is

$$LD(x) = \frac{1}{r}\sum_{i=1}^{N}\int_{l_i} k\left(\frac{s-x}{r}\right)\mathrm{d}l_i \tag{4}$$

where N is the number of linear features within the search bandwidth r of location x and l_i is the ith linear feature.

3.2.2. *Density distribution of linear features at network nodes*

Network topology changes at nodes. For example, the number of links often increases at nodes, causing the extending of search ranges around nodes. The extension of search range makes it difficult to ensure the correctness of density estimation at nodes. This paper estimates density around nodes unbiasedly based on equal-split kernel functions, proposed by Okabe *et al.* (2009). The equal-split kernel function is illustrated in Figure 7; I is a node in network space. The degree of node I is the number of links which take the node I as an endpoint, so the degree of node I is four. P, P_1, P_2 and P_3 are points in different links which aim to denote the links PI, IP_1, IP_2 and IP_3. For the location x, the form of kernel function $k(x)$ is kept unchanged on the same link ($P \rightarrow I$), and the form of kernel function is changed $k(x)$ into 1/3 $k(x)$ on the adjacent links ($I \rightarrow P_1$, $I \rightarrow P_2$ and $I \rightarrow P_3$). If there is another node within the search bandwidth r with degree of n_i, the kernel function $k(x)$ will be changed to $1/(n_i - 1)$ $k(x)$ on the next links, and so on. As a

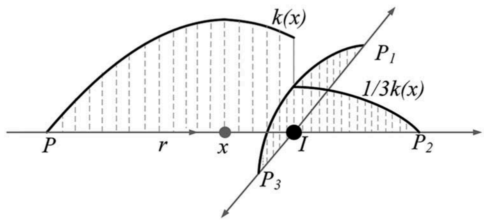


Figure 7. The equal-split kernel function.

result, no matter how many nodes or links within the search range of location x, the maximum of density is fixed for each linear feature. The equal-split kernel function method is able to normalize the density estimation results, so as to avoid density overestimation around nodes and ensure the estimate result is authentic. The form of equal-split kernel function (Okabe *et al*. 2009) is given by the following equation:

$$k\left(\frac{s-x}{r}\right) = \begin{cases} \frac{k((s-x)/r)}{(n_1-1)(n_2-1)\ldots(n_s-1)} & s-x <= r \\ 0 & s-x > r \end{cases} \tag{5}$$

Finally, density distribution in network space can be estimated according to Equations (3)–(5).

3.3. *Algorithm and its implementation of the NKDE-L*

The basic algorithm and implementation process of the NKDE-L is divided into three parts: road network segmentation, linear features processing and density calculation, as presented in Figure 8.

(a) Road network segmentation. First, the road network is broken into road segments at nodes. The road segments are the parts of roads between adjacent intersections. Second, a defined length is used to divide each road segment into Basic Segment Unit (*BSU*). If there is a residual when dividing the road segments into *BSU*s, we take the residual as a *BSU*. Finally, a *BSU*-based network topology is established, which is the density estimation context in the NKDE-L algorithm.

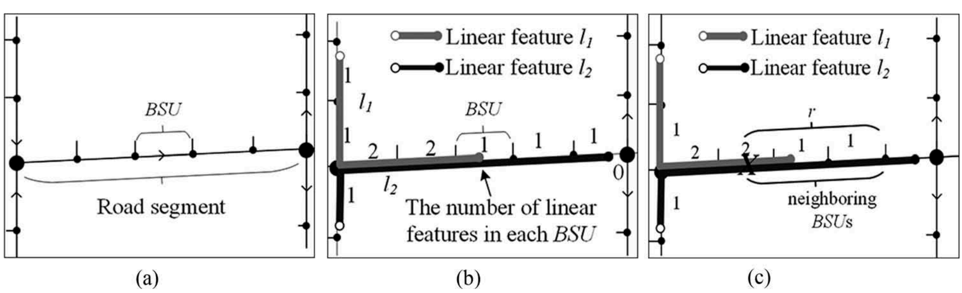


Figure 8. Algorithm and implementation of NKDE-L. (a) Road network segmentation, (b) linear features processing and (c) density calculation.

(b) Linear features processing. Once a *BSU*-based network topology has been established, the endpoints of each *BSU* and connectivity relationship between *BSU*s are known. For each linear feature, first match its start and end points to the endpoints (the intersection points of two adjacent *BSU*s) of its nearest *BSU* in the same direction. Each linear feature may traverse one or many *BSU*s. So given all the linear features and the *BSU*-based network, we know the number of linear features on each *BSU*. Then we get the traversed *BSU*s for each linear feature. Finally, the total number of linear features on each *BSU* is counted as a property of each *BSU*.

(c) Density calculation. For each *BSU*, the network distance from its centre point to the centre points of all its neighbouring *BSU*s within a given search bandwidth r is first calculated (including the *BSU* itself, in which case the network distance is 0). Then for the *BSU* and each of its neighbouring *BSU*, a density value based on the kernel function k and the network distance between them is calculated. In this step, the number of linear features on the neighbouring *BSU* is taken as a multiplier of the density value (the multiplier is zero if there is no linear features traversing on the neighbouring *BSU*). Finally, all density values of the *BSU* and each of its neighbouring *BSU* are cumulated as the final density of the *BSU*, as Equation (6) shows.

$$LD(\mathrm{s}) = \frac{1}{r}\sum_{i=1}^{M} N_i \int_{l_i} k\left(\frac{s-x}{r}\right)\mathrm{d}l_i \tag{6}$$

In Equation (6), $LD(\mathrm{s})$ is the final density value of a *BSU*, r is the search bandwidth. M is the number of neighbouring *BSU*s within search bandwidth r of the *BSU*, and $s - x$ is the network distance between the centre point of the *BSU* to its neighbouring *BSU* (say *BSU* i). N_i is the number of linear features on *BSU* i. So $N_i k\left(\frac{s-x}{r}\right)$ is the density value of the neighbouring *BSU* i. As a result, the final density value of the *BSU* is the accumulation of all density values of its neighbouring *BSU* i from 1 to M, which is $\sum_{i=1}^{M} N_i k\left(\frac{s-x}{r}\right)$.

4. Case study: analysing space–time dynamic of taxis' pick-up events in Wuhan

In this section, we apply the NKDE-L to analyse the space–time dynamic of taxis' pick-up events, with real road network and taxi trace data in Wuhan. The taxi trace data is collected from 10,614 taxis operating in the urban area of Wuhan, from 8 March (Sunday) to 14 September (Saturday) 2009. There are 2000 taxis operating in the study area. We choose trace data from all these taxis in the experiment. The taxis' GPS traces are sampled at a fixed time interval of 40 s, with a position accuracy of approximately 15 m. Each record contains information of taxi ID, position (longitude, latitude), velocity, orientation, timestamp and status (the status is 0 when the taxi is vacant and 1 when the

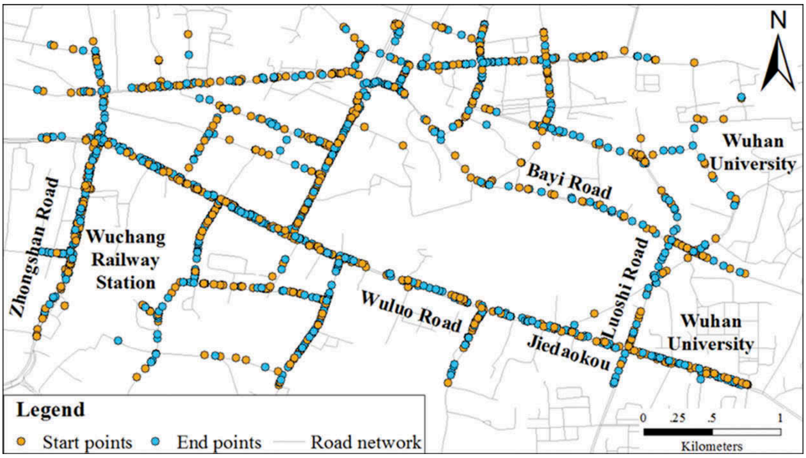


Figure 9. Road network in the study area and a sample of start and end points of pick-up *LE*.

taxi is occupied). The road network in the study area and a sample of start and end points of pick-up *LE* are shown in Figure 9.

In the experiment, we first extract pick-up *LE* from the taxis' GPS trace data in the study area. Second, we divide the road network in the study area into *BSU*s and match each pick-up *LE* to the traversed *BSU*s. Then the number of *LE* on each *BSU* can be obtained, based on which we calculate the density of each *BSU* with NKDE-L, as described in Section 3.3. The algorithm of NKDE-L is implemented using Microsoft Visual C# 2010, and the result is visualized in ESRI ArcGIS 10.1 environment.

In Section 4.1, we illustrate how to get the optimal bandwidth in NKDE-L, and give a comparison of the appearances of different bandwidths on the density results in NKDE-L. In Section 4.2, we discuss temporal and spatial distribution of pick-up *LE*, using the NKDE-L. In Section 4.3, we compare the results of the NKDE-L with that of quadrat method and planar KDE.

4.1. *Determination of optimal bandwidth in NKDE-L*

Search bandwidth is a critical parameter in NKDE-L. It determines the smoothness of the density result, which could reveal hotspots in different spatial scale. Methods for selecting optimal bandwidths in 2-D homogeneous context have been proposed (Silverman 1986, Chiu 1992, Gangopadhyay and Cheung 2002). However, they may not be suitable for density estimating in 1-D inhomogeneous network space. An optimal search bandwidth in NKDE-L should consider both the distribution characteristic of *LE* in the study area and the linear nature of network space. In this section, density results of pick-up *LE* are calculated by NKDE-L, with four versions of bandwidth: 30, 50, 100 and 200 m, respectively, with *BSU* length of 20 m. The road network in the study area and pick-up *LE* is shown in Figure 9. An optimal bandwidth is selected based on visual inspection of density distribution characteristics that different bandwidths reveal. Density results of the entire study area and a local part of the study area under the four bandwidths are shown in Figures 10 and 11.

Both Figures 10 and 11 show that the density results get smoother with search bandwidth increasing, at a fixed *BSU* length (20 m) and kernel function (Quartic

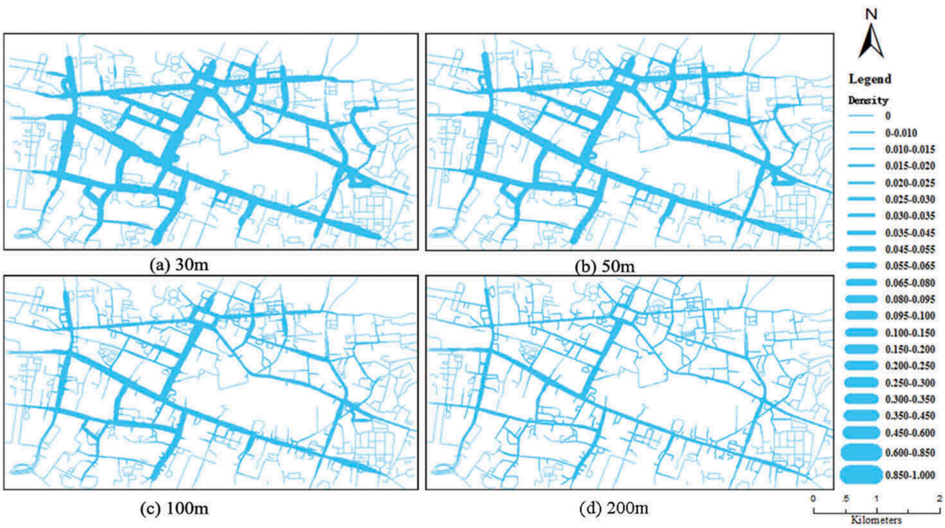


Figure 10. Density results in the study area produced by NKDE-L under four bandwidths (30, 50, 100 and 200 m) with 20 m of *BSU* length.

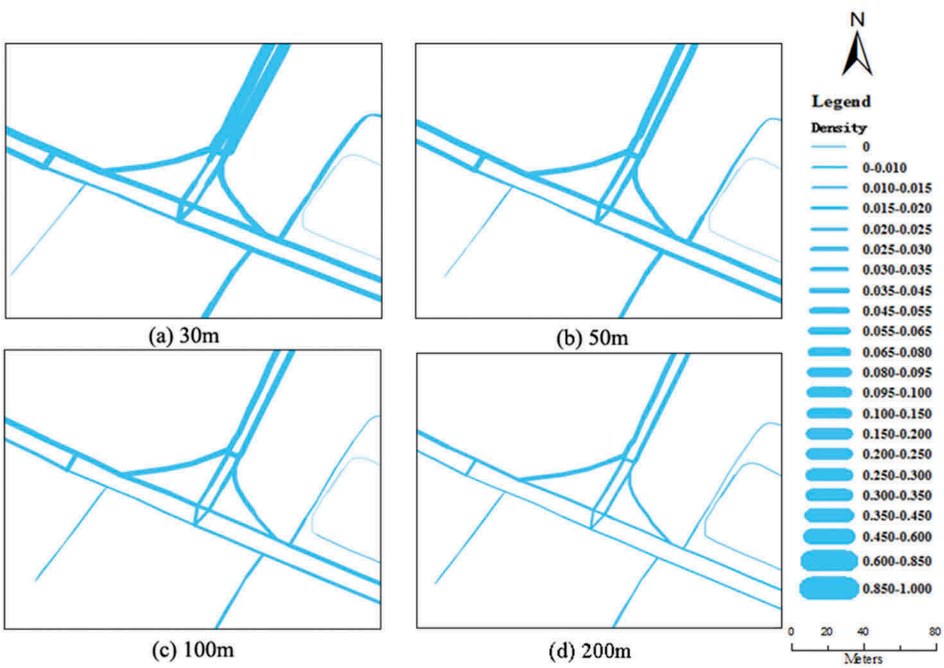


Figure 11. Density results in local scale produced by NKDE-L under four bandwidths (30, 50, 100 and 200 m) with 20 m of *BSU* length.

function). As is shown in Figure 10, the resulted density values at 30 m bandwidth are the maximum. The densities vary a lot on different roads, and distribute unevenly on the same road, presenting an unbalanced distribution in the entire study area. With the bandwidth increasing, the resultant density values decrease (because the denominator r in Equation 6 increases). The densities change more gently both on different roads and on the same road, and the overall density results get smoother. However, the density values change very little with a 200 m bandwidth, which is also undesirable in NKDE-L. So in order to reflect the density distribution variation both on different roads and on the same road, and to keep a balanced distribution at the same time, 100 m of bandwidth length is considered optimal, with the study area and 20 m of *BSU* length. In the local part of the study area as Figure 11 shows, detailed density variation needs to be unrevealed, in which case a narrower bandwidth of 50 m gives a better sense of density distribution. So, for the scale of the whole study area in this study, 100 m of bandwidth is considered optimal.

4.2. *Space–time dynamic analysis of pick-up LE with NKDE-L*

In this section, we first describe a time tuple for expressing various temporal granularities. Then we apply the NKDE-L to characterize the space–time dynamic of pick-up *LE* in road network. Both temporal distribution of the pick-up event volume and spatial distribution of pick-up event density are presented in this section.

This paper uses a time tuple proposed by Tang *et al*. (2013) to express temporal distribution of pick-up events of different time granularities. The time tuple is expressed as: T ⟨Granularity, Day-pattern, Day-from, Day-to, Period-number, Starttime$_1$, Endtime$_1$, Starttime$_2$, Endtime$_2$, Starttime$_3$, Endtime$_3$….⟩. Here Day-pattern is the pattern of dates, such as every day, every week, holiday, workday and weekend. Day-from and Day-to are the start day and end day, respectively. Period-number is the number of time intervals. Starttime and Endtime are the start and the end time of a time interval. For example, 7 a.m.–9 a.m. and 5 p.m.–7 p.m. from 8 March 2009 to 14 March 2009 can be expressed by T ⟨Granularity: day, Day-pattern: everyday, Day-from: 8 March 2009, Day-to: 14 March 2009, Period-number: 2, Startime$_1$: 7 a.m., Endtime$_1$: 9 a.m., Starttime$_2$: 5 p.m., Endtime$_2$: 7 p.m.⟩.

We test the NKDE-L on real road network and pick-up *LE* in a day (9 March 2009). To analyse the space–time dynamic characteristics of the pick-up *LE*, we divide all the pick-up events into 12 groups according to their occurring time, with representation by the proposed time tuple: T ⟨Granularity: day, Day-pattern: every day, Day-from: 9 March 2009, Day-to: 9 March 2009, Period-number: 12, Startime$_1$: 00:00, Endtime$_1$: 2:00, Startime$_2$: 2:00, Endtime$_2$: 4:00, …, Startime$_{12}$: 22:00, Endtime$_{12}$: 24:00⟩. First, density distribution of pick-up events of 12 periods of a day in the study area is calculated with NKDE-L, and displayed in Figure12, with *BSU* length of 20 m and search bandwidth of 100 m. Then, temporal dynamics of pick-up event volume for the entire study area and four main roads (Zhongshan-Road, Zhongnan Road, Wuluo Road and Luoshi Road) is presented in Figure 13(a) and (b).

Figure 12(a)–(l) reflects space–time dynamic of people's commuting and recreational activities as well as traffic condition. In Figures 12(a)–(l), where the width of lines depicts densities of pick-up events in road network, the overall densities of the study area

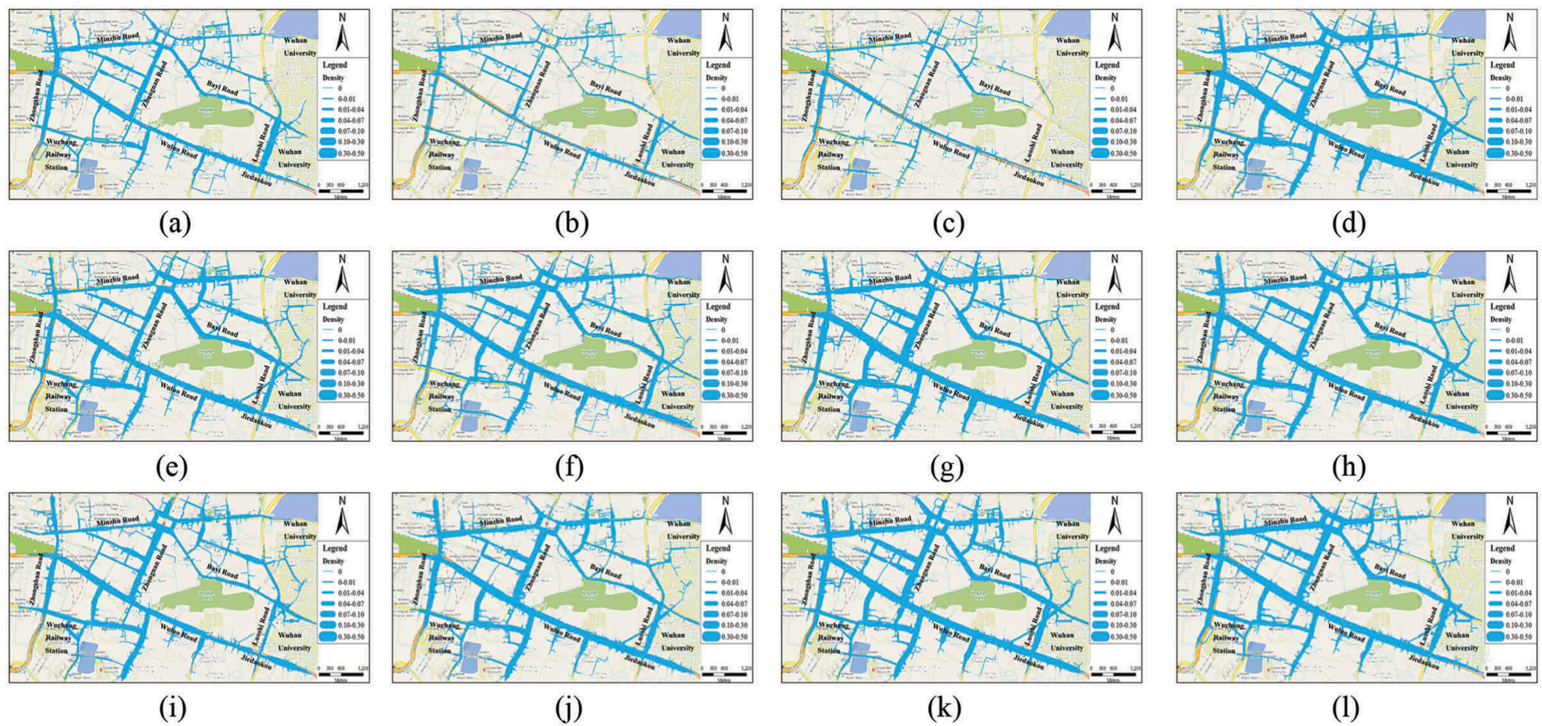


Figure 12. Density distribution dynamic of pick-up *LE* in network space with NKDE-L. (a) 0:00–2:00, (b) 2:00–4:00, (c) 4:00–6:00, (d) 6:00–8:00, (e) 8:00–10:00, (f) 10:00–12:00, (g) 12:00–14:00, (h) 14:00–16:00, (i) 16:00–18:00, (j) 18:00–20:00, (k) 20:00–22:00 and (l) 22:00–24:00.

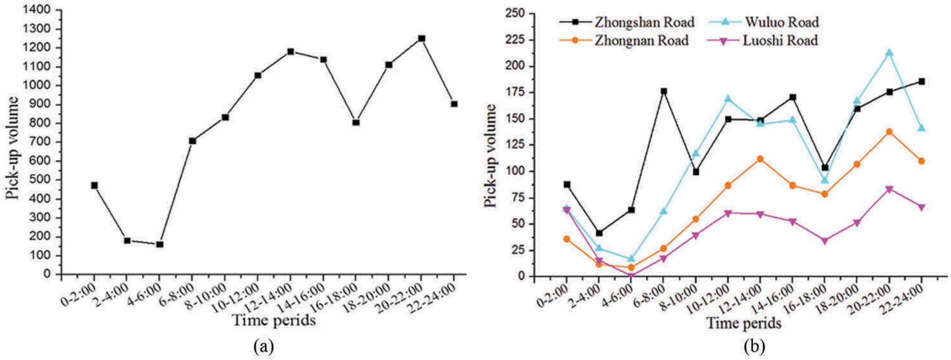


Figure 13. Temporal distribution of pick-up event volume in a day. (a) Pick-up volume of study area and (b) Pick-up volume of four main roads.

exhibit a similar temporal distribution to the temporal statistics result in Figure 13. The pick-up events' densities keep high in two periods in a day, during 8:00–16:00, and 18:00–22:00. In Figure 13(a) and (b), there is a 'V' shaped distribution of pick-up events around 16:00–18:00, during the rush hour. The reason for the special distribution is that the heavy traffic reduces the operation efficiency of taxis during the rush hour, regardless of the great demand of taxis. The number of pick-up events increases after 18:00 due to alleviation of the congested traffic. Furthermore, some detailed space–time dynamic characteristics can be uncovered from Figures 12 and 13. First, the pick-up events' densities in Zhongshan Road are higher than in other roads. And during 4:00–6:00, in which time the overall density is the minimum in the whole day, Zhongshan Road becomes the only hotspot. Also, after 22:00, pick-up densities in most roads drop down except in Zhongshan Road, where the density keeps increasing. This is because the Wuchang Railway Station locates in Zhongshan Road, which produces high demand for taxis enduringly. Second, density distributions in main roads start to increase after 6:00 with the increase of traffic flow, and stay high during the daytime, such as Wuluo Road and Luoshi Road. However, during the rush hour 6:00–8:00, the density of pick-up events in these roads are lower than other periods of a day. There are two reasons contributing to the special distribution, one is the traffic congestion factor, the other is a subway line set along Wuluo Road. People trend to take subway during rush hour to avoid the heavy traffic. The same distribution appears at another rush hour, 6:00–8:00, indicating that density distribution of pick-up events exhibits a strong periodicity. Third, in the business zone, such as Zhongnan Road, and the segment on Luoyu Road near Jiedaokou, the densities begin to increase after 8:00, and keep stably high during the daytime. After 18:00, the densities go on to increase, and drop down after 22:00. This space–time distribution is relevant to people's shopping and entertainment activities, which occur during the day and become more frequently in the evening. Fourth, there are few density distributions inside the campus area, such as Wuhan University, because taxi drivers are charged if they drive into the campus, and the demand for a taxi in campus is relatively low. Detailed information of distribution dynamic of pick-up events can be uncovered by zooming into a single road in Figure 12, which Figure 13(b) cannot reflect. For example, during the rush hour 6:00–8:00 and 16:00–18:00, traffic jam happens in the segment of Wuluo Road near Jiedaokou, where the density of pick-up *LE* is relatively low. The density of pick-up *LE* increases in the segment of Wuluo Road near Zhongnan Road, where the traffic jam is released. In comparison, the density in the segment of Wuluo Road near Zhongnan Road is higher than that near Jiedaokou during daily time, when the traffic is smoother. In Figure 12, we can clearly get the density space–time distribution both in different roads and zooming in different parts of a single road, by constraining the distribution of pick-up events with the range of road network.

4.3. *Comparison of NKDE-L with other methods*

Given a set of linear features and a *BSU*-based network, an intuitive approach to express the spatial distribution pattern is to count the number of linear features per *BSU*. This approach can be treated as quadrat method because a *BSU* can be regarded as a linear quadrat. This paper compares NKDE-L (20 m of *BSU* length, 100 m of search bandwidth) with quadrat method (20 m of quadrat length), and planar KDE (20 m × 20 m of grid size, 100 m of search bandwidth). The density distribution of pick-up events with NKDE-L, quadrat method and planar KDE are shown in Figure 14(a)–(c). Local detailed distributions with the three methods are presented in Figure 15(a)–(c). The local area in Figure 15 is marked by dashed ellipse A in Figure 14(a)–(c).

Figure 14(a)–(c) shows that all the three methods can reflect the spatial distribution characteristics of *LE* in a degree. Some common hotspots of pick-up *LE* can be identified from the density distribution in Figure 14(a)–(c), where subway stations, business centre and railway station locate, producing a high demand for taxis. However, the three methods characterize density distribution of *LE* in different ways. In Figure 14(a) and (b), density value is calculated for each linear unit (*BSU* or quadrat), and represented by line width. While in Figure 14(c), density value is calculated per raster cell, and described by the colour depth. The density distributions with NKDE-L (Figures 14(a) and 15(a)) show that the density distributions are continuous in the same road, and balanced in

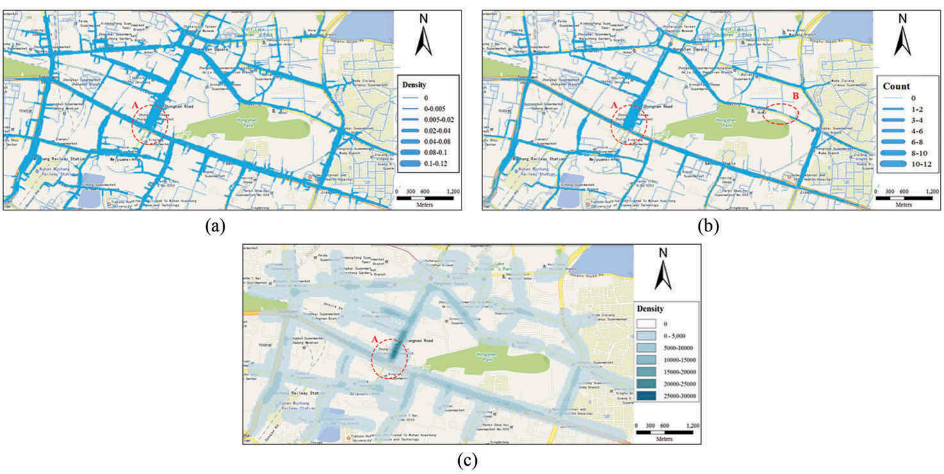
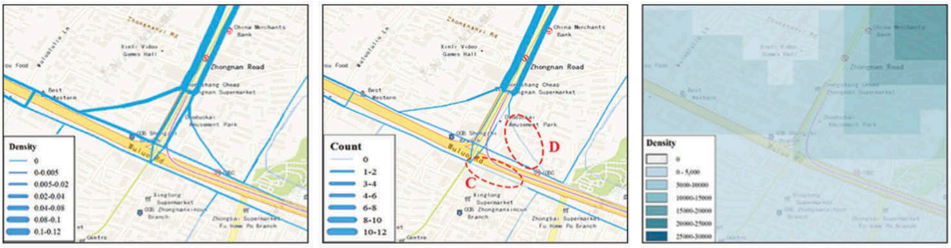


Figure 14. Comparison of NKDE-L with quadrat method and planar KDE in the study area. (a) NKDE-L, (b) Quadrat method and (c) Planar KDE.



(a) (b) (c)

Figure 15. Comparison of NKDE-L with quadrat method and planar KDE in a local scale. (a) NKDE-L, (b) Quadrat method and (c) Planar KDE.

different roads. The density distributions with NKDE-L are various but there is no sharp change along the road network, which reflects the distribution pattern of pick-up events accurately. While in quadrat method (Figures 14(b) and 15(b)), the density of pick-up *LE* presents a discontinuous distribution, and breaks off at some locations along the roads (the location where the *LE* break off is marked in the dashed ellipse B in Figure 14(b) and dashed ellipse C and D in Figure 15(b)), indicating that the likelihood of picking up a passenger changes sharply there, which conflicts with the reality. In planar KDE (Figures 14(c) and 15(c)), the density of *LE* is measured by Euclidean distance, and the whole 2-D plane is taken as context. The search area is larger in planar KDE than in network KDE (NKDE-L), with a same search bandwidth. For the *LE* which happen and distribute in 1-D network space, the 2-D search area in planar KDE makes the effect ranges of these events extend to the location where they do not exist. So the density surface result from planar KDE is smoother than network KDE. In comparison, the NKDE-L restricts the density distributions of pick-up *LE* within road network, where these events really occur and distribute. The density distribution in NKDE-L is spikier than that in planar KDE, but is more reasonable and approximate to the reality.

5. Conclusions and discussions

The existing methods for studying the spatial distribution pattern of linear features are all based on homogeneous 2-D or 3-D space, without considering the constrains of road network configuration and network direction to some kinds of *LE*, such as traffic congestion events, queuing up at intersections and passenger-carrying events. These events occur in 1-D road network space, and present a continuous linear distribution along road network. We call this kind of events as *LE*. Focusing on the spatial distribution of this special kind of *LE*, this paper puts forward a NKDE-L. The NKDE-L takes network distance as spatial measure, considers network topology, and studies the distribution of *LE* in network space. This paper first analyses the density distribution of each linear feature along networks, then estimates the density distribution for the whole network space in terms of the network distance and network topology. The correctness and unbiasedness of density estimation around network nodes are also considered.

In the case study, we apply the NKDE-L to analyse the space–time dynamics of taxis' pick-up events, with real road network and taxi trace data in Wuhan. We first describe a time tuple for expressing various temporal granularities. Then we present a space–time statistics of pick-up *LE* in four different temporal granularities, expressed by the time tuple. People's travel dynamic and traffic dynamic are uncovered through the space–time statistics. We further apply the NKDE-L to characterize the space–time dynamic of pick-up *LE* in the road network. By constraining the distribution of pick-up events with the range of road network, the results can reflect the dynamic of pick-up events accurately. To prove the advantages of the NKDE-L, we compare the NKDE-L with two common methods: quadrat method and the planar KDE. The comparison results show that the NKDE-L reflects the distribution pattern of *LE* in network space more accurately and reasonably.

The limitations of our research include: (1) We divide road network into *BSU*s and establish a BSU-based network. The cost of dividing road network and calculating density is huge when the road network is of high complexity. (2) The NKDE-L method is based on traditional KDE, which has a common limitation: the density estimation result is dependent on the *BSU* length and search bandwidth we choose. So choosing an optimal *BSU* length and search bandwidth is inevitable. Further studies of this paper include improving the theory of line pattern analysis method, improving efficiency of the algorithm as well as extending the applications of the NKDE-L. For example, the NKDE-L may be useful to analyse space–time data, in both spatial and temporal dimension.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported by National Natural Science Foundation of China [grant numbers 41571430, 41271442, 40801155]; Open Research Fund of the Academy of Satellite Application [2014_CXJJ-DSJ_02], and the Fund of Shenzhen Beidou Satellite Technology Engineering Research Center.

References

Anselin, L., *et al*., 2000. Spatial analyses of crime. *Criminal Justice*, 86 (3), 211–223.

Bailey, T.C. and Gatrell, A.C., 1995. *Interactive spatial data analysis*. Essex: Longman.

Borruso, G., 2003. Network density and the delimitation of urban areas. *Transactions in GIS*, 7 (2), 177–191. doi:10.1111/tgis.2003.7.issue-2

Borruso, G., 2005. Network density estimation: analysis of point patterns over a network. *In*: O. Gervasi, *et al*., eds. *Computational science and its applications – ICCSA*. Lecture Notes in Computer Science 3482, Part III. Berlin: Springer-Verlag, 126–132.

Borruso, G., 2008. Network density estimation: A GIS approach for analysing point patterns in a network space. *Transactions in GIS*, 12 (3), 377–402. doi:10.1111/tgis.2008.12.issue-3

Cai, X.J., Wu, Z.F., and Cheng, J., 2013. Using kernel density estimation to assess the spatial pattern of road density and its impact on landscape fragmentation. *International Journal of Geographical Information Science*, 27 (2), 222–230. doi:10.1080/13658816.2012.663918

Castro, P.S., *et al*., 2013. From taxi GPS traces to social and community dynamics. *ACM Computing Surveys*, 46 (2), 1–34. doi:10.1145/2543581

Chiu, S.-T., 1992. An automatic bandwidth selector for Kernel Density Estimation. *Biometrika*, 79 (4), 771–782. doi:10.1093/biomet/79.4.771

Downs, J.A., 2010. Time-geographic density estimation for moving point objects. *In*: S.I. Fabrikant, *et al*., eds. *GIScience*. Lecture Notes in Computer Science 6292, 14–17 September, Zurich. Berlin: Springer-Verlag, 16–26.

Downs, J.A. and Horner, M.W., 2012. Probabilistic potential path trees for visualizing and analyzing vehicle tracking data. *Journal of Transport Geography*, 23, 72–80. doi:10.1016/j.jtrangeo.2012.03.017

Downs, J.A., *et al*., 2014. Voxel-based probabilistic space–time prisms for analysing animal movements and habitat use. *International Journal of Geographical Information Science*, 28 (5), 875–890. doi:10.1080/13658816.2013.850170

Erdogan, S., *et al*., 2008. Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. *Accident Analysis & Prevention*, 40 (1), 174–181. doi:10.1016/j.aap.2007.05.004

Flahaut, B., *et al*., 2003. The local spatial autocorrelation and the kernel method for identifying black zones. A comparative approach. *Accident Analysis & Prevention*, 35 (6), 991–1004. doi:10.1016/S0001-4575(02)00107-0

Gangopadhyay, A. and Cheung, K., 2002. Bayesian approach to the choice of smoothing parameter in Kernel Density Estimation. *Journal of Nonparametric Statistics*, 14 (6), 655–664. doi:10.1080/10485250215320

Giannotti, F., *et al*., 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20 (5), 695–719. doi:10.1007/s00778-011-0244-8

Gong, L., *et al*., 2015. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography & Geographic Information Science*, 1–12. doi:10.1080/15230406.2015.1014424

Ha, H.H. and Thill, J.C., 2011. Analysis of traffic hazard intensity: a spatial epidemiology case study of urban pedestrians. *Computers Environment & Urban Systems*, 35 (3), 230–240. doi:10.1016/j.compenvurbsys.2010.12.004

Lahr, L., *et al*., 2014. An improved test for earnings management using kernel density estimation. *European Journal of Finance*, 23, 559–591.

Lee, D. and Hahn, M.A., 2014. A study on density map based crash analysis. *In*: *2014 international conference on information science and applications* (ICISA), 6–9 May Seoul. New York: IEEE, 1–3.

Lee, J.H., Shin, I., and Park, G.L., 2008. Analysis of the passenger pick up pattern for taxi location recommendation. *In*: *Proceedings of the 4th international conference on networked computing and advanced information management*, 2–4 September Gyeong ju. New York: IEEE, 1:199–204.

Li, B., *et al*., 2011a. Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset. *In*: *8th IEEE international workshop on managing ubiquitous communications and services*, 21–25 March Seattle, WA. New York: IEEE, 63–68.

Li, Q., *et al*., 2004. Transect-based three-dimensional road modeling and visualization. *Geo-spatial Information Science*, 7 (1), 14–17. doi:10.1007/BF02826670

Li, Q., *et al*., 2011b. Dynamic accessibility mapping using floating car data: a network-constrained density estimation approach. *Journal of Transport Geography*, 19 (3), 379–393.

Li, X.L., *et al*., 2012. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science in China*, 6 (1), 111–121.

Liu, L., Andris, C., and Ratti, C., 2010. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers Environment & Urban Systems*, 34, 541–548.

Miller, H.J., 1999. Potential contribution of spatial analysis to geographic information systems for transportation (GIS-T). *Geographical Analysis*, 31 (4), 373–399.

Okabe, A., Satoh, T., and 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.

O'Sullivan, D. and Wong, D.W.S., 2007. A surface-based approach to measuring spatial segregation. *Geographic Analysis*, 39 (2), 147–168.

Scheepens, R., *et al*., 2011. Composite density maps for multivariate trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 7 (12), 2518–2527.

Silverman, B.W., 1986. *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Tang, L., *et al*., 2013. Road network modeling and representation for time-dependent traffic control. *Information*, 16 (9A), 6459–6472.

Timothée, P., *et al*., 2010. A network based kernel density estimator applied to barcelona economic activities. *In*: D. Taniar, *et al*., eds. *Computational science and its applications – ICCSA 2010*. Lecture Notes in Computer Science 6016, 23–26 March, Fukuoka. Berlin: Springer-Verlag, 32–45.

Worton, B.J., 1989. Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70 (1), 164–168.

Xie, Z. and Yan, J., 2008. Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32 (5), 396–406.

Ying, L., *et al*., 2014. Space-time patterns of road network and road development priority in three parallel rivers region in Yunnan, China: an evaluation based on modified kernel distance estimate. *Chinese Geographical Science*, 24 (1), 39–49.