Journal of Transport Geography 31 (2013) 64–71

Contents lists available at SciVerse ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach

Zhixiao Xie [a,*], Jun Yan [b]

[a] Department of Geosciences, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA
[b] Department of Geography and Geology, Western Kentucky University, Bowling Green, KY 42101, USA

ARTICLE INFO

Keywords:
Network KDE
Local Moran's I
Hot spots
Traffic accidents
Monte Carlo simulation

ABSTRACT

Kernel density estimation (KDE) has long been used for detecting traffic accident hot spots and network kernel density estimation (NetKDE) has proven to be useful in accident analysis over a network space. Yet, both planar KDE and NetKDE are still used largely as a visualization tool, due to the missing of quantitative statistical inference assessment. This paper integrates NetKDE with local Moran'I for hot spot detection of traffic accidents. After density is computed for road segments through NetKDE, it is then used as the attribute for computing local Moran's I. With an NetKDE-based approach, conditional permutation, combined with a 100-m neighbor for Moran's I computation, leads to fewer statistically significant "high-high" (HH) segments and hot spot clusters. By conducting a statistical significance analysis of density values, it is now possible to evaluate formally the statistical significance of the extensiveness of locations with high density values in order to allocate limited resources for accident prevention and safety improvement effectively.

Published by Elsevier Ltd.

1. Introduction

Since the invention of gasoline-fueled automobiles in the late 19th century, there have been well over three million people killed in traffic accidents in the US alone (Evans, 2004). In fact, motor-vehicle crashes were the leading cause of death in the US for persons of every age from 3 through 33 in 2003, and carried a tremendous estimated economic cost of $230.6 billion in 2000 (NHTSA, 2006). Globally, the World Health Organization (WHO, 2004) reported that road-traffic accidents are the leading injury related cause of death among people aged 10–24, and it is estimated that more than 1.2 million people are killed each year in automobile crashes in addition to 50 million injuries. With such enormous human and economic costs resulting from traffic accidents, understanding where, why, what, and when automobile crashes occur is crucial to creating a safe driving environment. Traffic accidents are rarely random in space and time due to the fact that the underlying environment of traffic accidents, such as roadway networks, traffic volumes and, ultimately, human activities, often exhibits discernible spatial and temporal patterns. In general, the dynamics of urban transportation systems are largely the product of interactions among various components of the urban system and of human activities in space and time, particularly transportation and urban land use (Knox and McCarthy, 2005; Rodrigue et al., 2009). In most cases, traffic accidents form clusters (known as "hot spots") in geographic space (Xie and Yan, 2008). Hot-spot analysis is deemed a key issue in traffic safety strategies, with four phases involved: identification, ranking, profiling, and treatment (Moons et al., 2009). The identification of traffic accident hot spots is the first essential step for the appropriate allocation of resources for safety improvements (Anderson, 2009).

There are many relevant studies in the literature on traffic accident hot-spot analysis and detection (see Black and Thomas, 1998; Li et al., 2007b), among which the kernel density estimation approach is a widely adopted method. Density estimation is a procedure to construct an estimate of the density function from the observed data (Silverman, 1986). For spatial point data, it is similar to estimating a bivariate probability density (Bailey and Gatrell, 1995). Kernel Density Estimation (KDE) is one of the most popular methods for analyzing the first order properties of a point event distribution (Silverman, 1986; Bailey and Gatrell, 1995) and it has been widely used for traffic accident hot-spot analysis. For example, the planar KDE was used in studies of urban cyclists' traffic hazard intensity (Delmelle and Thill, 2008), pedestrian crash zones detection (Pulugurtha et al., 2007), wildlife-vehicle accident analysis (Krisp and Durot, 2007), highway accident hot-spot analysis (Erdogan et al., 2008), road accident hot-spot classification (Anderson, 2009), and fatal crash analysis (Oris, 2011). The use of planar KDE over a 2-D Euclidean space has its limitations in analyzing traffic accidents because the accidents are often constrained

* Corresponding author. Tel.: +1 561 2972852; fax: +1 561 2972745.
E-mail addresses: xie@fau.edu (Z. Xie), jun.yan@wku.edu (J. Yan).

0966-6923/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.jtrangeo.2013.05.009

only to the network portion of the 2-D space, the so-called network space (Yamada and Thill, 2007). Recently, a network-based kernel density estimation (NetKDE) was developed by Xie and Yan (2008) to estimate the density of traffic accidents strictly over a network space. NetKDE was later included in a free ArcGIS-based software tool, Spatial Analysis on a NETwork (SANET), developed by a group of researchers at the University of Tokyo (Okabe et al., 2009; Okabe and Sugihara, 2012). From a broad perspective, NetKDE represents an emerging effort to extend the applications of standard spatial statistical methods to analyzing spatial point events in a network space (Okabe et al., 1995, 2009; Okunuki and Okabe, 1999; Okabe and Yamada, 2001; Yamada and Thill, 2004, 2007; Lu and Chen, 2007; Xie and Yan, 2008; Okabe and Sugihara, 2012).

In cases other than spatial point pattern analysis (SPPA), quantitative hot-spot analysis relies almost exclusively on local spatial statistics, another recent significant advance in geographic information science and spatial analysis (O'Sullivan and Unwin, 2010). Local statistics are any descriptive statistic associated with a spatial dataset whose value varies spatially. Many local statistics are available, and a general contextualization of local statistics can be found in Anselin (1995). Two of the most popular local spatial statistics are Getis–Ord G_i^* (Ord and Getis, 1995) and local Moran's I (Anselin, 1995). Both have been used for detecting accident hot spots (Moons et al., 2009; Truong and Somenahalli, 2011; Kuo et al., 2012).

In this paper, we integrate network KDE and a spatial statistics technique, local Moran'I, for hot-spot detection. The KDE belongs to the methods examining the first-order effects of a spatial process, while Moran'I is one of the methods for examining the second-order effects of a spatial process. The combined strength of both should lead to more effective hot-spot detection. In addition, one of the major limitations for KDE and NetKDE is that no formal statistical inference is employed in the process and there is no indication of a density threshold above which a hot spot can be confidently declared (Bailey and Gatrell, 1995; Xie and Yan, 2008; Kuo et al., 2012). Applying a local statistical approach, such as local Moran's I, to density values resulting from NetKDE could provide a useful mechanism for conducting rigorous statistical tests. The rest of the paper is organized as follows: The network KDE and local Moran's I methods are briefly discussed in Section 2, along with a Monte Carlo simulation and significance assessment. In Section 3, the results from six experiments are presented based on a case study with real road network and traffic accident data. A discussion and conclusion follow in the final section.

2. Methods

2.1. Network KDE

As detailed in Xie and Yan (2008), a network KDE is an extension of the standard 2-D KDE and it uses the following form (Eq. (1)) of kernel density estimator for the density estimation of network-constrained point events in a network space.

$$\lambda(s) = \sum_{i=1}^{n} \frac{1}{r} k\left(\frac{d_{is}}{r}\right) \tag{1}$$

where $\lambda(s)$ is the density at location s, r is the search radius (bandwidth) of the KDE (only points within r are used to estimate $\lambda(s)$), $k(\frac{d_{is}}{r})$ is the weight of a point i at distance d_{is} to location s. The so-called kernel function, k, is formed as a function of the ratio between d_{is} and r, so that the "distance decay effect" can be taken into account in density estimation. A number of different kernel functions are available, such as Gaussian, Quartic, Conic, negative exponential, and epanichnekov (Levine, 2004; Gibin et al., 2007). Based on the literature, the choice of kernel function k is less important than the choice of search bandwidth r in planar KDE (Silverman, 1986; Bailey and Gatrell, 1995; Schabenberger and Gotway, 2005; O'Sullivan and Unwin, 2010). Xie and Yan (2008) demonstrated a similar situation in NetKDE. Hence, this paper only tests the Quartic kernel, one of the three most popular kernel functions (Schabenberger and Gotway, 2005). The specific form of the function is:

$$k\left(\frac{d_{is}}{r}\right) = \frac{3}{\pi}\left(1 - \frac{d_{is}^2}{r^2}\right) \text{ when } \quad 0 < d_{is} \leqslant r$$

$$k\left(\frac{d_{is}}{r}\right) = 0 \text{ when } \quad dis > r \tag{2}$$

In implementing NetKDE, Xie and Yan (2008) suggested using a linear segment (called *lixel*) of roads as the basic unit for aggregating accidents, calculating density, and for visualization. They found that segments of shorter length are more capable of showing the local variations of the segments. The linear segments with accidents geocoded are called source segments. For each source segment, the NetKDE density values are computed for the segment and its neighbors. For a segment falling within the search bandwidths of multiple source segments (including functioning as a source segment itself), its density is the cumulative value of the densities computed from all relevant source segments. The result of NetKDE is a linear spatial dataset with density values computed and assigned to each linear segment or *lixel*. For details of the computational implementation of NetKDE, see Xie and Yan (2008). Okabe and Sugihara (2012) also described more treatments of NetKDE implementation in their most recent work.

2.2. Moran's I and spatial autocorrelation of NetKDE density

Moran's I (Moran, 1948) is one of the most commonly used statistics for measuring spatial autocorrelation by translating a non-spatial correlation to a spatial context (O'Sullivan and Unwin, 2010). It is usually applied to areal data with numerical ratio or interval values (Rogerson, 2001; O'Sullivan and Unwin, 2010). The original Moran's I is a global statistic. For practical applications like traffic accident hot-spot detection, local spatial statistics are better suited in that they can help identify and examine where unusual clusters of events occur based on a formal assessment of statistical significance. Local Moran's I is fully developed by Anselin (1995) and its formation is shown in the following equation:

$$I_i = z_i \sum_j w_{ij} z_j \tag{3}$$

where z_i, and z_j are the deviations from the mean, and w_{ij} is the weight defined for each of the neighboring spatial unit j of i under a particular definition of neighborhood.

Moran's I is usually applied to areal data, but it is also used for data of other geometry types such as points (Ebdon, 1985; Truong and Somenahalli, 2011) and linear features (Moons et al., 2009). In a study of traffic accident hot spots by Truong and Somenahalli (2011), the severity of each accident is quantified first, and then the severity value is used as the attribute to compute the local Moran's I for each accident location. In contrast, Moons et al. (2009) aggregated accidents to 100-m segments and used the count of each segment as the attribute value for computing the Moran's I value.

This study differs from previous approaches in that we analyze the hot spots of traffic accidents by examining the spatial autocorrelation of accident density values resulting from NetKDE, instead of accident points or a tabulated count at a road segment. This design is more attractive for several reasons. Conceptually, an accident does not occur at a dimensionless, precise point location, but occupies and affects a certain length along a roadway,

especially when considering the relatively fine scales needed for local hot-spot analysis. The use of points largely simplifies the dimension and ignores the impact on the nearby roadway. It does not consider the uncertainty added by accident logging processes. Conversely, an accident may be more properly characterized as a spatial event that carries with it a spread of risk. The spread of risk can be defined as the neighboring area around an observed accident location where there is a different likelihood for an accident to occur depending on how close areas are to the accident (Anderson, 2009). KDE (and NetKDE) is naturally suited to quantifying the spread of risk of an accident. Hence, it is possible to compute the local Moran's I for the density dataset derived from the NetKDE once the density values can be interpreted as risk.

Some may argue that density values calculated from KDE (and NetKDE) should inherently be positively spatially autocorrelated, and nearby density values should be similar to each other since neighboring points within the distance of a bandwidth are used (in fact, weighted) in the KDE process. Thus there is no need to carry out additional spatial autocorrelative analysis on density values. As pointed out before, one of the major limitations for KDE and NetKDE is that no statistical inference can be assessed in the process and there is no indication of a density threshold above which a hot spot can be confidently declared. We may guess that locations with high density values could possibly be hot spots, but no mechanism has been available in KDE to assess their statistical significance. By conducting a statistical significance analysis on density values with local statistics such as local Moran's I, it is possible to evaluate formally the statistical significance of the extensiveness of locations with high density values, and to determine if hot spots of traffic accidents indeed exist consistently along certain portions of a roadway network.

2.3. Monte Carlo simulation and significance assessment

For a more quantitative hot-spot analysis, it is necessary to have a statistical significance test, which is missing in the NetKDE (Xie and Yan, 2008). Different approaches are discussed in the literature to perform the significance test. Two different null hypotheses were generally assumed for a spatial process: (1) normality and (2) randomization (Ebdon, 1985; Bailey and Gatrell, 1995; Mitchell, 2005). Under a randomization null hypothesis, a real spatial pattern is assumed to be only one of the realizations of a conditional spatial permutation process, by shuffling the given attribute values of the observed data among those spatial units. Basically, data values are fixed, while spatial arrangements change (Mitchell, 2005). Under a normalization null hypothesis, both the data values and their spatial arrangements are subject to random changes, but the values are assumed to be normally distributed. The expected values of Moran's I and its standard deviation can be estimated under these two hypotheses, and a z-score can be used to test the significance of the observed spatial pattern (Ebdon, 1985; Anselin, 1995; Mitchell, 2005;). In one recent study of accident hot spots by Truong and Somenahalli (2011), the *z*-score of Moran's I is used to test the significance of the spatial autocorrelation of traffic accident severity at selected points.

However, the use of a *z*-score for significance testing often has an implicit assumption that Moran's I values follow a normal distribution. The normal distribution assumption may not necessarily be true, as demonstrated experimentally by Moons et al. (2009). In fact, the distribution of Moran's I values may be better characterized as being intangible (Besag and Newell, 1991; Moons et al., 2009). Under such a circumstance, a Monte Carlo simulation seems a more appropriate approach for inferring a "pseudo-significance" value (O'Sullivan and Unwin, 2010) for Moran's I (Anselin, 1995; Moons et al., 2009). By adopting a Monte Carlo simulation, not only can we avoid making some unreliable assumptions, but also we can make the significance test more robust and versatile, since it is largely immune to any assumption of data distribution. In fact, a Monte Carlo simulation intends to reveal data distribution through a large number of experiment runs, such as 999 times. This obviously carries with it an added drawback – for example, a relatively intensive computational load – but this is not a significant limiting factor given current computation technology.

In this study, the distribution of local Moran's I values is established with a large number of runs of Monte Carlo simulations. Two types of Monte Carlo simulations were performed based on two hypothesized spatial processes respectively. The first is via a conditional (spatial) permutation process, by shuffling the accident counts of all source segments among these segments many times. In these simulations, only the original real source segments, namely those segments with one or more traffic accidents assigned, are used as source segments, although the accident counts may change through the permutation process. In addition, all tests have the same segment accident count distribution; that is, the possible accident values for these segments are given. The second type of Monte Carlo simulation is complete randomness: each of those accidents is randomly assigned to any road segment. Any segment can serve as a source segment and the possible accident counts for segments will change. We need to point out that these values may not follow a normal distribution. In either type of simulation, after the completion of assigning accidents to source segments, the NetKDE density values are calculated and they are then used to compute local Moran's I. The Moran's I from all the Monte Carlo simulations under respective hypotheses are used to establish the pseudo-distribution of Moran's I values. Finally the pseudo-significance is derived by comparing the Moran's I of the real observed dataset against the pseudo-distribution.

It should be noted that not all segments and their computed Moran's I are used for establishing the distribution as discussed below. There are different types of spatial autocorrelation that can be revealed by examining the combinations of attribute value at one location and those at its neighbors: the "low-low" (LL), or "high-high" (HH) for positive autocorrelation, and "low–high" (LH) or "high-low" (HL) for negative autocorrelation (O'Sullivan and Unwin, 2010). The cases relevant for traffic accident detection are those HH cases; these are basically locations with high values surrounded by neighboring locations also with high values (thus hot spots), which could be statistically unusual (Moons et al., 2009). Similar to Moons et al. (2009), the HH cases are defined as each of those locations (i) satisfying two conditions: (1) $(y_i - \bar{y}) > 0$; (2) $\sum_j w_{ij}(y_j - \bar{y}) > 0$, where y_j is the NetKDE density value at the jth neighbor location of location i. Note again that the data values we examined are the density values computed with NetKDE for each segment, while a count number was used in Moons et al. (2009). The local Moran's I is computed for these HH segments using the NetKDE density estimated from the real traffic accident points or the NetKDE density in the Monte Carlo simulations.

After identifying the segments with statistical significance, these segments are aggregated to form accident hot spot clusters. The aggregation is performed by merging significant segments with their direct significant neighbors. This aggregation process should be very useful for practical applications when the real goal is to identify hot spots as linear regions instead of points or short segments.

3. Experiments and results

3.1. Experiment data: road segmentation and accident assignment

The data for this study include a real transportation network system in the Bowling Green, Kentucky, area, with traffic-accident

data for 2005 (Fig. 1). The traffic accident dataset is provided by the Kentucky State Police (KSP, 2005). The point location of each accident is recorded as a longitude and latitude pair via the on-board GPS unit in a reporting police vehicle. A total of 3226 traffic accidents are analyzed in the case study. The two datasets are both projected in the NAD83, UTM 16N coordinate system.

Different segment lengths can be used in NetKDE depending on application settings. To reveal hot spots at a relatively finer scale, 10 m is adopted as the basic segment length in the experiments. A total of 84,030 segments resulted, with a mean length of 9.70 m and a standard deviation of 1.38 m. The 3226 accidents were then assigned to their nearest segments. A total of 2637 segments had an accident(s) assigned to them, with a maximum of 12 accidents, a mean of 1.22, and a standard deviation of 0.65.

3.2. Experiment settings

A series of experiments were conducted to test the proper approaches for detecting accident hot spots, based on NetKDE and Moran's I (Table 1). The basic segment length was the same (i.e., 10 m). The first four experiments (I–IV) were NetKDE based. For these NetKDE-based experiments, the NetKDE computation procedure and parameters were the same. All used a Quartic kernel, and the kernel bandwidth was also the same (i.e., 100 m). The first two (I, II) used the same Monte Carlo simulation (complete randomness), while conditional permutation was applied to experiments III and IV. Experiment I differs from Experiment II in that, when computing local Moran's I, the search neighborhood for Experiment I was 100 m; that is, all the segments within this network distance of 100 m are deemed neighbors and they are given the same weights, while only the direct segment neighbors were counted for Experiment II. Experiments III and IV are the "duplicates" of Experiments I and II respectively, but with a different type of Monte Carlo simulation (conditional permutation instead of complete randomness). For comparison between NetKDE density based and count-based approaches for significance tests, two more experiments (V and VI) were conducted. In Experiment V, a conditional permutation Monte Carlo simulation was performed to assign accidents to segments, while a complete random Monte Carlo simulation was performed to assign accidents to segments in Experiment VI. In both Experiments V and VI, the number of accidents tabulated for source segments was used directly for the local Moran's I computation. The search neighborhood was 100 m for calculating local Moran's I values, the same as that for Experiments I and III, but neighbor segments were given different weights following a distance decay function in the same form of the NetKDE kernel, i.e. quartic function. The distance decay function is used to perform a fair comparison between NetKDE density based and count-based approaches, because a distance decay function is already adopted in NetKDE.

3.3. Experiment results

The six experiments were applied to the real accidents (3226) and real road segments (84,030 segments of 10 m or less) in the case study area. The number of statistically significant (HH) segments and hot-spot clusters for the experiments are shown in Table 2 for different levels of significance. The resulting clusters

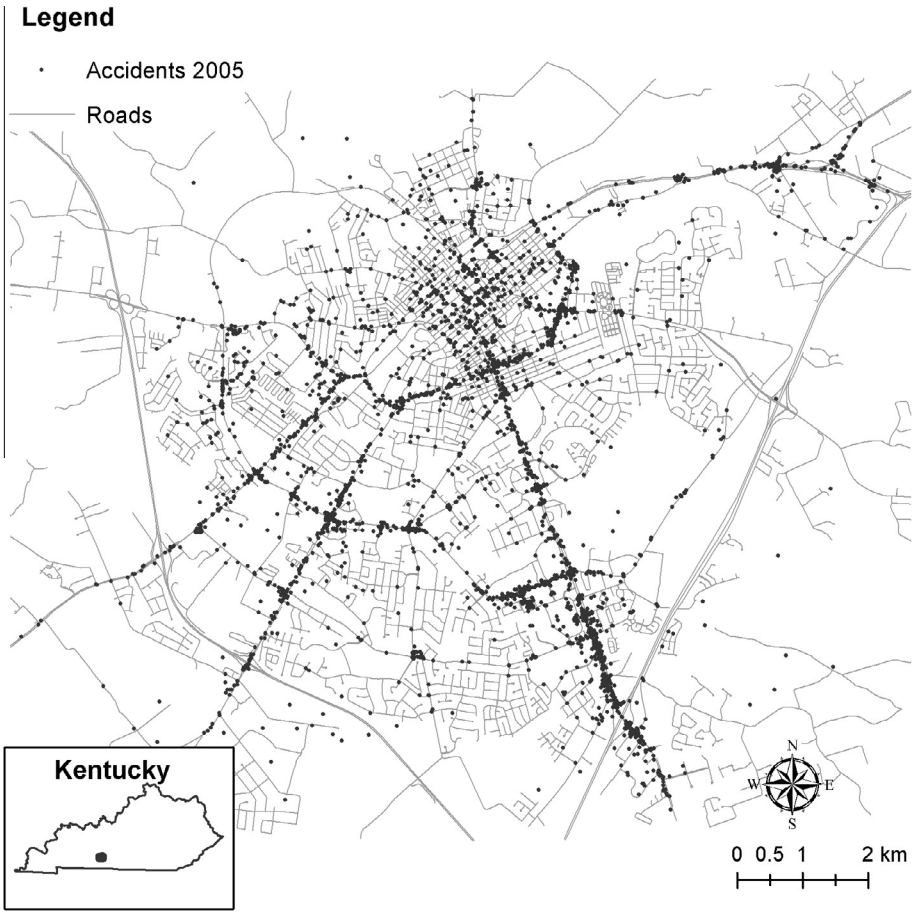

Fig. 1. Study area, the road network and the 2005 traffic accident data.

Table 1
The six experiments and their settings.

Experiment	Attribute value Calculating Moran's I	Monte Carlo simulation	Number of simulation	Moran's I neighborhood
I	NetKDE density	Complete random	100	100 m; equal weight
II	NetKDE density	Complete random	100	Direct neighbor, equal weight
III	NetKDE density	Conditional permutation	100	100 m; equal weight
IV	NetKDE density	Conditional permutation	100	Direct neighbor, equal weight
V	Accident count	Complete random	100	100 m; weight calculated as quartic function
VI	Accident count	Conditional permutation	100	100 m; weight calculated as quartic function

at a 0.05 significance level (roughly generated with 100 permutations in the Monte Carlo simulation) are also shown in Fig. 2.

As shown in Table 2, as expected, the number of significant segments continually decreases with the increasing significance level from 0.50 to 0.05 for every experiment. The number of hot-spot clusters (numbers in parentheses) shows the same trend. Between those two hypotheses, experiments with the completely random Monte Carlo simulation (Experiments I, II, V) produced a higher number of significant segments than their corresponding counterparts with the conditional permutation Monte Carlo simulation (Experiments III, IV, VI), no matter whether NetKDE density or count is used. Similar observations can be made when comparing the number of hot-spot clusters between the two kinds of Monte Carlo simulations.

Table 2 also shows the clear impact of Moran's I neighborhood definition on the significance assessment and cluster detection. Between Experiments I and II, both of which adopt a completely random Monte Carlo simulation, Experiment I with a 100-m neighborhood results in far fewer hot-spot clusters than Experiment II with direct neighbors at each significance level. The search neighborhood shows the same impact with a conditional permutation Monte Carlo simulation. Experiment III with a 100-m neighborhood results in far fewer hot-spot clusters than Experiment IV with direct neighbors at each significance level. This makes sense, since the clusters identified with larger neighborhoods are often indicative of the existence of more extensive clusters along the network.

When comparing the NetKDE-based experiments (Experiments I–IV) and count-based ones (Experiments V and VI), the number of significant segments produced by the count-based approach is generally much lower, with one exceptions at the 0.05 significance level for the completely random simulation (Experiment V). In terms of hot-spot clusters, the count-based approach results in a drastically higher number of clusters at every significance level compared to the NetKDE-based approach. For example, there are still 773 and 122 clusters at the 0.05 significance level for Experiment V and VI respectively, while the NetKDE-based approach (Experiments I–IV) has a smaller number of clusters (53, 93, 27, 55 respectively). The complete random simulation (assignment of accidents to segments) in Experiment V seems not very useful for count-based approaches since there are too many clusters at the 0.05 significance level. The higher number of clusters in count-based experiments is due to the fact that these significant segments are not as spatially contiguous as NetKDE-based ones, as shown in a zoom-in portion of the significant segments at the 0.05 significance level (Fig. 3).

To compare further the significant segments output from the six experiments, pairwise spatial overlay was conducted at the 0.05 significance level and the number of identical segments between each pair is shown in Table 3. Among NetKDE-based approaches, a very interesting association can be observed: all the significant segments in Experiment III are also part of those in Experiment I, while all the significant segments in Experiment IV are also part of those in Experiment II. Because Experiments I and III both adopt a 100-m neighborhood, while Experiments II and IV use direct neighbors, this suggests that the neighborhood definition has an important role in determining which segments to include in the significant segments. General clusters identified with larger neighborhoods are indicative of the existence of more extensive clusters. On the other hand, a complete random simulation, in NetKDE-based approaches, seems to result in more significant segments than the conditional permutation, since Experiments III and IV both use a conditional permutation simulation, while Experiments I and II both use a complete random simulation. A similar result is observed for count-based approaches. Fewer significant segments are produced by conditional permutation (Experiment VI) and all these segments are part of the significant segments determined by the complete random simulation (Experiment V). It should be noted that all the significant segments in Experiment VI are part of those in Experiment I and II, and most of them are also part of Experiments III and IV. This suggests that significant segments in a count-based approach through conditional permutation can also be identified via a NetKDE-based approach.

While Tables 2 and 3 present the quantitative differences of the significance assessment and cluster detection between the two types of simulations, the impacts of Moran's I neighborhood definition, as well as the differences between NetKDE-based and count-based experiments, are more clearly visualized in Figs. 2 and 3.

In summary, it appears that a smaller number of spatially contiguous clusters is identified by a conditional permutation Monte Carlo simulation, combined with 100-m search neighborhood, based on the NetKDE density, while other combinations produce a relatively higher number of segments and/or clusters. In particular, the count-based approaches result in a drastically higher num-

Table 2
The number of significant (HH) segments for the six experiments at different significance levels. The number in the parenthesis is the number of hot spot clusters.

Sig. Level	Experiment					
	I	II	III	IV	V	VI
0.50	7975(98)	9057(182)	7439(91)	8566(173)	1733(1206)	1017(608)
0.40	6941(85)	6974(160)	5864(84)	6667(165)	1662(1138)	833(483)
0.30	5888(86)	5426(147)	4359(75)	4935(138)	1549(1035)	666(376)
0.20	4751(77)	4439(122)	2938(61)	3335(108)	1445(948)	505(291)
0.15	4126(76)	3948(113)	2208(50)	2539(94)	1404(912)	428(251)
0.10	3458(65)	3201(106)	1542(42)	1725(86)	1342(861)	330(194)
0.05	2528(53)	2352(93)	835(27)	896(55)	1230(773)	197(122)

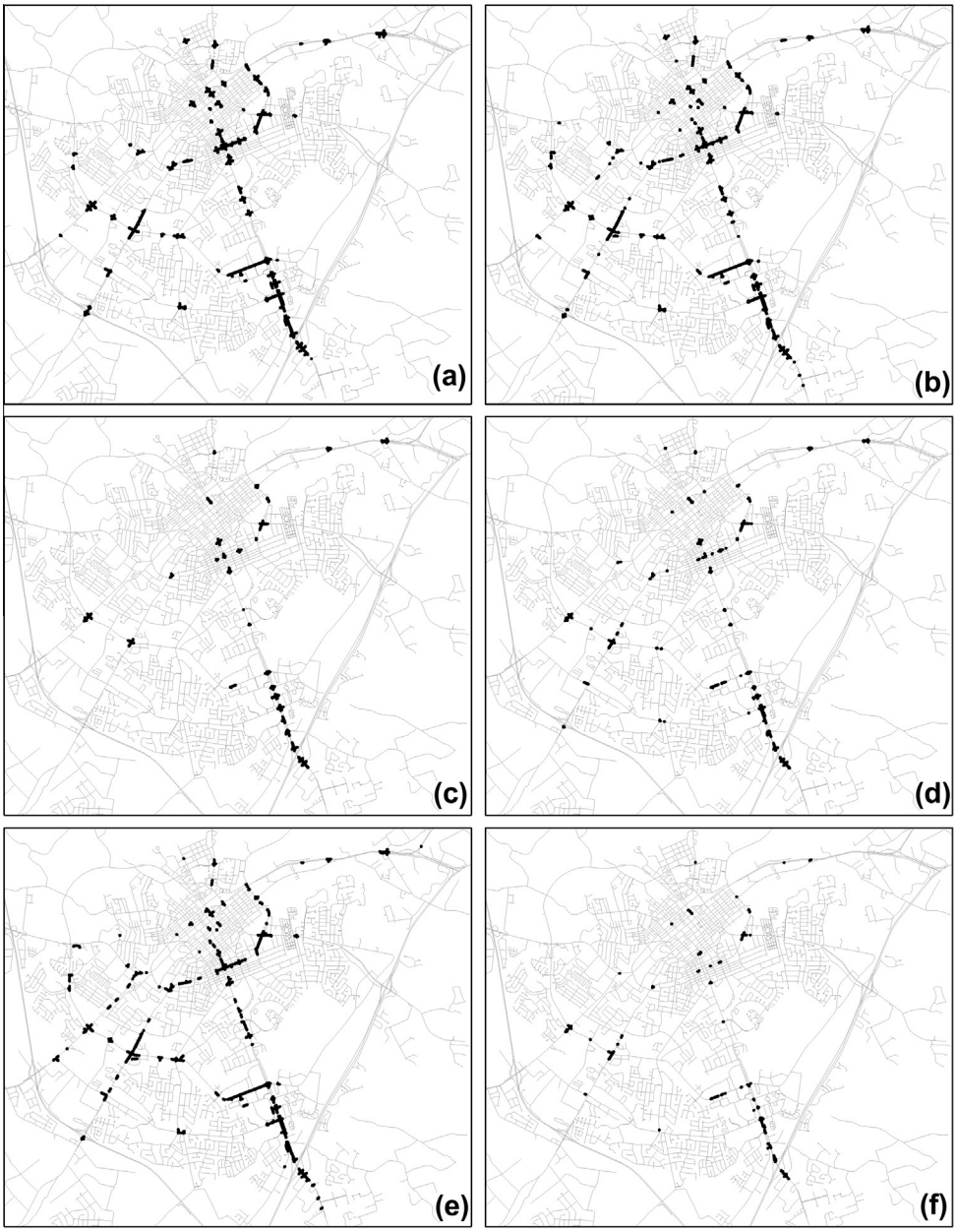

Fig. 2. The significant segments (dark segments) for the six experiments at the 0.05 significance level, (a) Experiment I, (b) Experiment II, (c) Experiment III, (d) Experiment IV, (e) Experiment V, (f) Experiment VI.

ber of clusters that are not as spatially contiguous as NetKDE-based ones.

4. Discussions and conclusions

Traffic accidents are one of the major threats to human well-being and safety, and the identification of traffic accident hot spots is essential for the allocation of limited resources for transportation safety improvements. A kernel-density estimation approach has long been used for detecting traffic accident hot spots, and NetKDE has proven to be more appropriate in accident analysis over a network space. Yet, both planar KDE and NetKDE are still used largely as a visualization tool, due to the absence of quantitative statistical assessment. On the other hand, after the NetKDE density estimation, road segments can be regarded as a basic spatial unit or *lixel* (Xie and Yan, 2008) and they are naturally exposed to a wider set of spatial analysis methods. This paper integrated the NetKDE with local Moran'I, a popular local spatial statistic, for hot-spot detection, and the results clearly show the effectiveness of such an integration approach. It seems that the NetKDE-based approach may be of more practical value for decision making because identified hot-spot clusters are spatially more contiguous and the number of clusters is much smaller. In addition, the adoption of NetKDE helps mitigate potential instability in observed accident counts.

It should be noted that this research design (integration of NetKDE and Moran's I) has implicitly changed the representation of traffic accidents from simple points to events with a spread of risk. The statistical inference is adjusted accordingly. In particular, the local Moran's I is computed based on aggregated NetKDE density values and, as a result, its distribution is rather elusive, especially for those HH segments. Therefore, it is logical that the significance

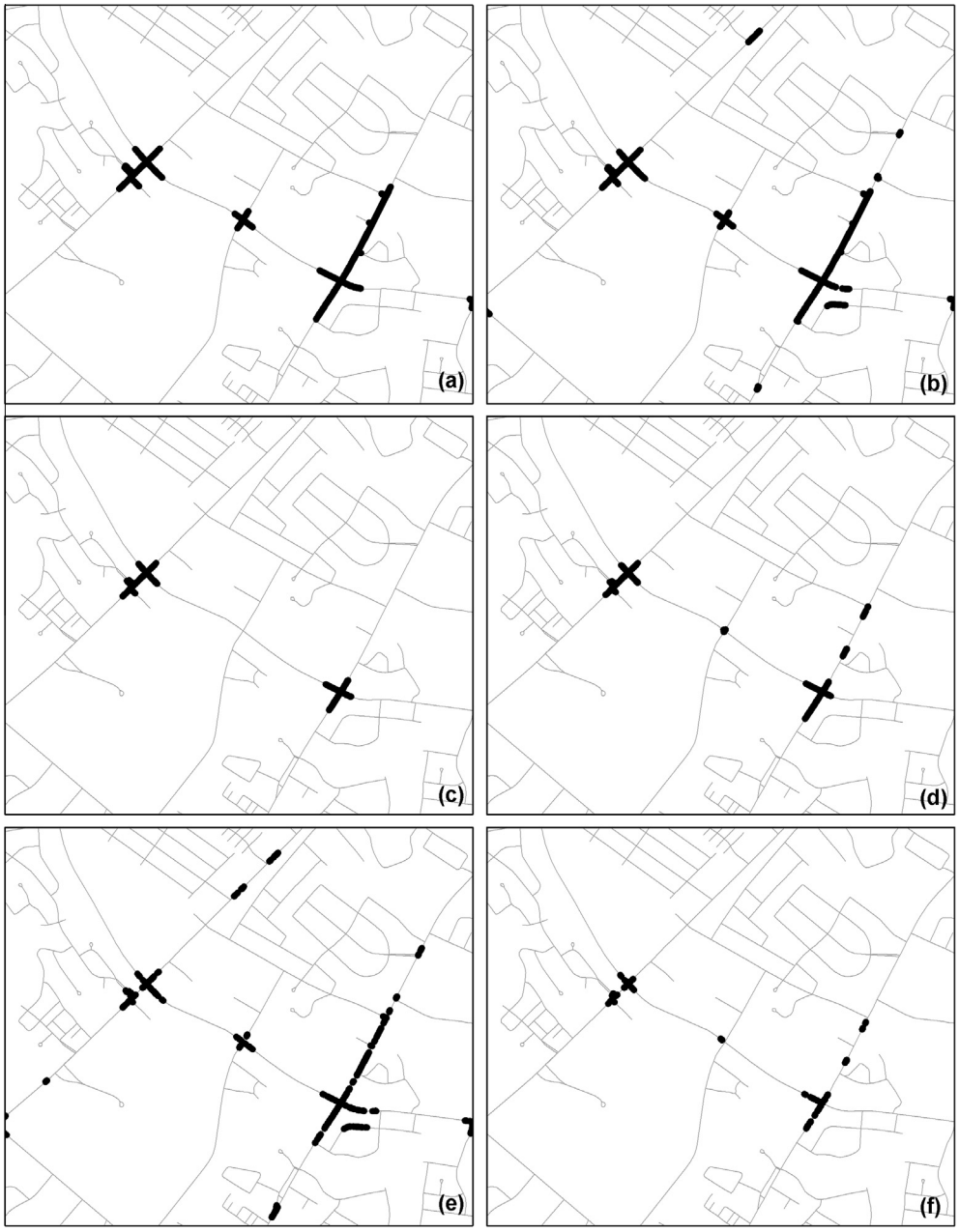

Fig. 3. The illustration of more spatial details of the significant segments (dark segments) for the six experiments at the 0.05 significance level, for a small portion of the study area, (a) Experiment I, (b) Experiment II, (c) Experiment III, (d) Experiment IV, (e) Experiment V, (f) Experiment VI.

Table 3
The number of identical segments between the output significant segment datasets for the six experiments at the 0.05 significance level. The number of significant segments (and clusters) for each experiment is listed in the headings.

Experiment	I 2528(53)	II 2352(93)	III 835(27)	IV 896(55)	V 1230(773)	VI 197(122)
I	2528	2149	835	896	870	197
II		2352	832	896	846	197
III			835	703	331	158
IV				896	356	177
V					1230	197
VI						197

test has to rely on a pseudo-distribution built through Monte Carlo simulation. Two hypotheses were tested in this study: complete randomness and conditional permutation. Our results show that the conditional permutation leads to fewer statistically significant HH segments and hot-spot clusters; hence it may be preferable when resources are limited and fewer clusters are to be identified. It is also important to note that the hot-spot clusters from the complete random simulation contain the hot-spot clusters from condi-

tional permutation (at least at the 0.05 significance level in this test), when relevant settings (segment length, kernel function, kernel bandwidth, Moran's I neighborhood) are the same.

One alterative randomness simulation design would be to fix one source segment at a time and randomly distribute the remaining accidents. However, in a unique spatial problem setting like accidents along a road network, with a large number of road segments (84,030 in this study), but very few (2637) source accident segments, the simulation results may not be much different between this design and complete randomness. In addition, in randomness simulations, by fixing one particular segment at a time, there also comes an increasing computational load, although computation is not a major concern these days.

Moran's I was selected in this study because it is one of the most commonly used spatial autocorrelation measures. Li et al. (2007a) argued that Moran's I is only a good estimator of the spatial autoregressive model's spatial dependence parameter when the parameter is close to 0 and they developed a closed-form measure of spatial autocorrelation called APLE. To determine the most appropriate measure in accident hot-spot detection, more research is needed to compare the performances of those local statistics measures. Also note that, in our experiments, we only ran NetKDE with 10-m segment lengths and a 100-m bandwidth, following the results from a previous study (Xie and Yan, 2008). In future research, sensitivity analysis may be needed using other parameter values. Furthermore, this study focused on a methodological demonstration of the integration of NetKED and Moran's I for accident hot-spot detection. The road network and traffic are largely simplified. For example, traffic volume is rarely constant as implied in the case study, and road segments with a higher traffic volume should naturally suffer more accidents. Therefore, traffic volume and other factors should be taken into consideration in a real application. Nevertheless, the study does show that the integration of NetKDE and local spatial statistics can be applied to traffic accident hot-spot detection, and provides enhanced statistical rigor. It is believed that such detection could be valuable for traffic accident prevention and safety improvement in the future.

Acknowledgement

We would like to thank three anonymous reviewers and the editor for their constructive comments that helped us improve the paper.

References

Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hot spots. Accid. Anal. Prev. 41, 359–364.
Anselin, L., 1995. Local indicators of spatial association-LISA. Geogr. Anal. 27 (2), 93–116.
Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Longman Scientific, Harlow, UK.
Besag, J., Newell, J., 1991. The detection of clusters in rare diseases. J. Royal Stat. Soc. Ser. A 154, 327–333.
Black, W.R., Thomas, I., 1998. Accidents on Belgium's motorways: a network autocorrelation analysis. J. Transp. Geogr. 6 (1), 23–31.
Delmelle, E.C., Thill, J.-C., 2008. Urban bicyclists – a spatial analysis of adult and youth traffic hazard intensity. Transp. Res. Rec. 2074, 31–39.
Ebdon, D., 1985. Statistics in Geography. A Practical Approach. sedond ed.. Wiley-Blackwell, Malden, MA.
Erdogan, S., Yilmaz, I., Baybura, T., Gullu, M., 2008. Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. Accid. Anal. Prev. 40 (1), 174–181.
Evans, L., 2004. Traffic Safety. Science Serving Society, Bloomfield Hills, MI.
Gibin, M., Longley, P., Atkinson, P., 2007. Kernel density estimation and percent volume contours in general practice catchment area analysis in urban areas. In: Section 5A in Proceedings of the GIScience research UK conference (GISRUK), April 11–13, Maynooth, Ireland. <ncg.nuim.ie/gisruk/materials/proceedings/>.
Knox, P., McCarthy, L., 2005. Urbanization: an Introduction to Urban Geography, second ed. Pearson/Prentice Hall, Upper Saddle River, NJ.
Krisp, J.M., Durot, S., 2007. Segmentation of lines based on point densities – an optimisation of wildlife warning sign placement in southern Finland. Accid. Anal. Prev. 39 (1), 38–46.
KSP (Kentucky State Police). 2005. Traffic Collision Facts, 2005 Report. Frankfort, KY: Commonwealth of Kentucky.
Kuo, P., Zeng, X., Lord, D., 2012. Guidelines for choosing hot-spot analysis tools based on data characteristics, network restrictions, and time distributions. In: Proceedings of the 91 Annual Meeting of the Transportation Research Board, January 22–26, Washington, DC. <http://amonline.trb.org/1snld8/1> (accessed 25.05.2012).
Levine, N., 2004. CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations. The National Institute of Justice, Washington, DC.
Li, H., Calder, C.A., Cressie, N., 2007a. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geogr. Anal. 39 (4), 357–375.
Li, L., Zhu, L., Sui, D.S., 2007b. A GIS-based Bayesian approach for analyzing spatial–temporal patterns of intra-city motor vehicle crashes. J. Transp. Geogr. 15 (4), 274–285.
Lu, Y., Chen, X., 2007. False alarm of planar K-function when analyzing urban crime distributed along streets. Soc. Sci. Res. 36 (2), 611–632.
Mitchell, A., 2005. The ESRI Guide to GIS Analysis, vol. 2. ESRI Press, Redlands, CA.
Moons, E., Brijs, T., Wets, G., 2009. Improving Moran's index to identify hot spots in traffic safety. Geocomputation Urban Plan. 176, 117–132.
Moran, P.A.P., 1948. The interpretation of statistical maps. J. Royal Stat. Soc. Ser. B (Methodological) 10 (2), 243–251.
NHTSA (National Highway Traffic Safety Administration). 2006. Traffic Safety Facts: 2005 Data. <http://www-nrd.nhtsa.dot.gov/Pubs/810623.pdf> (accessed 12.06.2012).
Okabe, A., Sugihara, K., 2012. Spatial Statistics along Networks: Statistical and Computational Methods. John Wiley, Chichester, UK.
Okabe, A., Yamada, I., 2001. The K-function method on a network and its computational implementation. Geogr. Anal. 33, 271–290.
Okabe, A., Yomono, H., Kitamura, M., 1995. Statistical analysis of the distribution of points on a network. Geogr. Anal. 27 (2), 152–175.
Okabe, A., Satoh, T., Sugihara, K., 2009. A kernel density estimation method for networks, its computational method and a GIS-based tool. Int. J. Geogr. Inform. Sci. 23 (1), 7–32.
Okunuki, K., Okabe, A., 1999. A Computational Method for Optimizing the Location of a Store on a Continuum of a Network when Users' Choice Behavior Follows the Huff Model. Discussion Paper #19, Center for Spatial Information Science at the University of Tokyo. <www.csis.u_tokyo.ac.jp/english/dp/dp.html>.
Ord, J.K., Getis, A., 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geogr. Anal. 27, 286–306.
Oris, W.N., 2011. Spatial Analysis of Fatal Automobile Crashes in Kentucky. M.S. Thesis, Department of Geography and Geology, Western Kentucky University, Bowling Green, Kentucky.
O'Sullivan, D., Unwin, D.J., 2010. Geographic Information Analysis. John Wiley, Hoboken, NJ.
Pulugurtha, S.S., Krishnakumar, V.K., Nambisan, S.S., 2007. New methods to identify and rank high pedestrian crash zones: an illustration. Accid. Anal. Prev. 39 (4), 800–811.
Rodrigue, J.-P., Comtois, C., Slack, B., 2009. The Geography of Transport Systems, second ed. Routledge, New York, NY.
Rogerson, P.A., 2001. Statistical Methods for Geography. Sage Publications, London, UK.
Schabenberger, O., Gotway, C.A., 2005. Statistical Methods for Spatial Data Analysis. Chapman Hall/CRC, Boca Raton, FL.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman Hall, London, UK.
Truong, L.T., Somenahalli, S.V.C., 2011. Using GIS to Identify Pedestrian-Vehicle Crash Hot Spots and Unsafe Bus Stops. J. Public Transport. 14 (1), 99–114.
WHO (World Health Organization). 2004. World Report on Road Traffic Injury Prevention. <http://whqlibdoc.who.int/publications/2004/9241562609.pdf> (accessed 12.06.2012).
Xie, Z., Yan, J., 2008. Kernel density estimation of traffic accidents in a network space. Comput. Environ. Urban Syst. 35 (5), 396–406.
Yamada, I., Thill, J.-C., 2004. Comparison of planar and network K-functions in traffic accident analysis. J. Transp. Geogr. 12, 149–158.
Yamada, I., Thill, J.-C., 2007. Local indicators of network-constrained clusters in spatial point patterns. Geogr. Anal. 39 (3), 268–292.