Scientific Research Publishing

World Journal of Engineering and Technology, 2017, 5, 198-215
http://www.scirp.org/journal/wjet
ISSN Online: 2331-4249
ISSN Print: 2331-4222

Identifying Vehicular Crash High Risk Locations along Highways via Spatial Autocorrelation Indices and Kernel Density Estimation

Azad Abdulhafedh[1]

Department of Civil and Environmental Engineering, University of Missouri-Columbia, MO, USA
Email: asa8cd@mail.missouri.edu

How to cite this paper: Abdulhafedh, A. (2017) Identifying Vehicular Crash High Risk Locations along Highways via Spatial Autocorrelation Indices and Kernel Density Estimation. *World Journal of Engineering and Technology*, **5**, 198-215.
https://doi.org/10.4236/wjet.2017.52016

Received: February 27, 2017
Accepted: May 7, 2017
Published: May 10, 2017

Abstract

Identifying vehicular crash high risk locations along highways is important for understanding the causes of vehicle crashes and to determine effective countermeasures based on the analysis. This paper presents a GIS approach to examine the spatial patterns of vehicle crashes and determines if they are spatially clustered, dispersed, or random. Moran's *I* and Getis-Ord Gi* statistic are employed to examine spatial patterns, clusters mapping of vehicle crash data, and to generate high risk locations along highways. Kernel Density Estimation (KDE) is used to generate crash concentration maps that show the road density of crashes. The proposed approach is evaluated using the 2013 vehicle crash data in the state of Indiana. Results show that the approach is efficient and reliable in identifying vehicle crash hot spots and unsafe road locations.

Keywords

Spatial Autocorrelation, Kernel Density, Moran's *I*, Gi* statistic, Hot Spots Analysis

1. Introduction

Identifying vehicular crash high risk locations along highways is a useful tool that can help transportation agencies allocate limited resources more efficiently, and find effective countermeasures. A crash hot spot is a location showing concentration of incidents, and hot spot analysis is a method for analyzing the spatial tendency between points or events within this location [1]. If a feature's spatial tendency is high, and the values of its neighboring features is also high, it is a part of a hot spot, and if the tendency of a feature and its neighborhoods is low, it is a part of a cold spot. Spatial patterns of traffic crash data can be analyzed by

[1]PhD in Civil Engineering

spatial autocorrelation, which is a measure of the correlation of an observation with other observations through space. The spatial autocorrelation phenomenon can be summarized by the Tobler's first law of Geography that everything is usually related to all else but those which are near to each other are more related when compared to those that are further away [2]. Most statistical analyses are based on the assumption that the values of observations in each sample are independent of one another. Spatial autocorrelation violates this assumption, because samples taken from nearby locations are related to each other, and hence, they are statistically not independent of one another [1] [3]. To assess spatial autocorrelation, a distance measure must be specified in order to define what is meant by two observations being close together. These distances are usually presented in the form of a weight matrix, which defines the relationships between locations at which the observations occur [4]. If data were collected at n locations, then the weight matrix will be $n \times n$ with zeroes on the diagonal. The weight matrix is often row-standardized, (*i.e.* all weights in a row sum to one), and can be constructed given a variety of assumptions, such as [1]:

- A constant distance that represents the weight for any two different locations.
- A fixed weight for all observations within a specified distance.
- k nearest neighbors that represents a fixed weight, and all others non-neighbors are zero.
- Weight could be proportional to the inverse distance, or inverse distance squared.

There are a number of indices or statistics that attempt to measure spatial autocorrelation for continuous data, such as Moran's I, Geary's C, and Getis-Ord G_i statistic [5]. These indices can be used as Global or Local measures depending on the scope of the analysis. Global implies that all elements in the weight matrix are included in the calculation of spatial autocorrelation providing a single measurement of spatial autocorrelation for an entire data set. Local indices calculate spatial autocorrelation for all areal units of analysis. In other words, the global autocorrelation is the extent to which points that are close together in space have similar values, and the local autocorrelation is the extent to which points that are close to a given point or area have similar values. Anselin [6] outlined a general class of local indicators of spatial autocorrelation termed the Local Indicator of Spatial Autocorrelation (LISA) statistic, which implies that the *LISA* statistic decomposes global results into their local parts. For example, a significant global index at a given spatial point or section may hide large spatial patches of no autocorrelation, and *LISA* can detect this and show us the location of these insignificant patches in space. Conversely, an insignificant global result may hide patches of strong autocorrelation, and *LISA* can detect this again. Generally, both hot spots and cold spots can be identified as locations for which the *LISA* statistic is significant [6]. High and significant values of Moran's I and G_i statistic in a spot indicate a high spatial clustering (hot spot), whereas low and significant values indicate a low spatial clustering (cold spot). The type of clus-

tering and its statistical significance is evaluated based on a confidence level and on the output z-scores and the correspondent p-values. These will determine whether a data point or a location belongs to a hot spot (denoted by High-High, HH), cold spot (denoted by Low-Low, LL) or an outlier (a high data value surrounded by low data values or vice versa, denoted by High-Low, HL or Low-High, LH). Other methods for studying spatial patterns of crash data as point events have recently been developed. One of the most widely used is the Kernel Density Estimation (KDE). The goal of KDE is to develop a continuous surface of density estimates of discrete events such as road crashes by summing the number of events within a search bandwidth. Many recent studies have used the 2-D planar KDE for hot spot analysis. However, this method has been criticized in relation to the fact that road crashes usually happen on the road links and need to be considered in a road network space represented by 1-D dimension. Therefore, some studies have extended the planar KDE to network spaces, which estimates the crash density over a distance unit in a 1-D measurement instead of an area unit [7] [8] [9]. However, a major weakness of the KDE methods is that it cannot be tested for statistical significance [8] [10].

2. Literature Review

Vehicle crashes have been investigated from different spatial and temporal perspectives by different researchers using varied procedures. The Black and Thomas [11] paper was the first major work that clearly distinguishes traffic crash hotspots. Their work indicated that a positive network autocorrelation of road crashes can cause the spatial clustering of traffic hotspots. However, the analysis focused on network autocorrelations at a global level within an entire dataset by using Moran's I and the associated z-score tests. Flahaut [12] introduced the use of the local indices of spatial autocorrelation (*LISA*) to examine the crash patterns of road networks in a Belgian province. This work explained the advantages of hotspots and further developed logistic regression models to explain traffic crash hotspots with road characteristics and local environmental conditions. Yamada and Thill [7] explained and compared hotspot methods by Kernel Density Estimation (KDE) at the planar and network-constrained. Their work indicated that the planar KDE analysis can produce over-detecting cluster patterns. More recently, Yamada and Thill [13] introduced a method called local indicators of network-constrained clusters (LINCS) to identify hotspots by using the network-based approach. Theoretically, traffic crashes can occur at every possible location over the entire road network. However, it is impractical to examine the clustering pattern at every possible point using the network-based approach. Hence, they suggested using reference points along the network with an equal interval distances. Schweitzer [14] used a process called kernel smoothing for hot spot analysis. This process creates local estimates of the measure of the spatial intensity using the count of frequency of points within a given distance of each point, relative to symmetric distribution. Xie and Yan [8] used a planar and a network-based KDE approach to examine traffic crashes in the state of Ken-

tucky. In implementing the network KDE, they suggested using a linear segment of roads as the basic unit for aggregating crashes, calculating density, and for visualization. They found that segments of shorter length are more capable of showing the local variations of the segments, and concluded that the network-constrained KDE is more appropriate than the planar KDE for traffic crash analysis. Erdogan *et al.* [15] used a repeatability analysis to identify hotspots with the highest 5% and 1% area of the Poisson distribution over ten years, using a bandwidth of 500 m. They concluded that repeatability analysis determined more hot spot locations than the Kernel Density analysis. Yamada and Thill [16] applied the network-based framework to analyze highway crashes that occurred on a small highway network in Buffalo, New York. The method was implemented in conjunction with Monte Carlo simulation to obtain criteria against which statistical inferences from the observed patterns can be made. They found that incorporating GIS and spatial statistical approaches can effectively detect crash hotspots.

3. Moran's *I*

Moran's *I* [17] is one of the oldest indices of spatial autocorrelation and can be used to test for global and local spatial autocorrelation among continuous data. For any continuous variable, x_i a mean $\bar{x}$, can be calculated and the deviation of any observation from that mean can be calculated based on the cross products of the deviations from the mean. The statistic then compares the value of the variable at any one location with the values at all other locations [18] [19] [20]. For n observations on a variable x at locations i, j, Moran's I is calculated by Equation (1) as follows:

$$I = \frac{n}{S_0} \frac{\sum_i^n \sum_j^n w_{ij} \left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_i^n \left(x_i - \bar{x}\right))^2} \tag{1}$$

where,

$\bar{x}$: is the mean of the variable x;

w_{ij}: are the elements of the weight matrix;

S_0: is the sum of the elements of the weight matrix: $S_0 = \sum_i^n \sum_{ji}^n w_{ij}$

Values for this index typically, range from −1.0 to +1.0, where a value of -1.0 indicates negative spatial autocorrelation, and a value of +1.0 indicates positive spatial autocorrelation. When nearby points or segments have similar values, their cross product is high. Conversely, when nearby points or segments have dissimilar values, their cross-product is low. The expectation of Moran's I is:

$$E\left(I\right) = \left(\frac{-1}{n-1}\right) \tag{2}$$

with a Moran's I value larger than $E(I)$, indicates positive spatial autocorrelation, and a Moran's I less than $E(I)$, indicates negative spatial autocorrelation. In Moran's formulation, the weight variable, w_{ij}, is a contiguity matrix. If zone j is adjacent to zone $_i$, the product receives a weight of 1.0. Otherwise, the product

receives a weight of 0.0. The *z*-scores of Moran's *I* can be computed by Equation (3):

$$Z_i = \frac{I - E(I)}{\sqrt{V(I)}} \tag{3}$$

where *E(I)* is the expected value of *I*, and *V(I)* is the variance of *I*, as shown in Equation (4):

$$V(I) = E(I^2) - E^2(I) \tag{4}$$

The distribution of the *z*-scores is assumed to be approximately normal with a mean of 0.0 and a variance of 1.0 (Cliff and Ord 1981). A statistically significant positive *z*-score indicates that the distribution of the observations is spatially autocorrelated producing High-High (HH) clusters, whereas a negative *z*-score indicates that the observations tend to be more dissimilar producing Low-Low (LL) clusters. A *z*-score close to zero indicates that observations are randomly and independently distributed in space. By assuming a *z*-score is from a standard normal distribution, their associated *p*-value can be obtained, and can be used to determine the significance of the index at each location [4]. To determine if the *z*- score is statistically significant, it should be compared to the range of values for a particular confidence level. For example, at a significance level of 95%, a *z*-score would have to be less than –1.96 or greater than + 1.96 to be statistically significant. The null hypothesis H_0 is that there is no spatial autocorrelation among the observations. The null hypothesis can be rejected, if the *p*-value shows that the *z*-score is significant.

4. Getis-Ord G_i Statistic

The Getis-Ord G_i statistic is another index of spatial autocorrelation [21] that can distinguish between positive spatial autocorrelation with high values from positive spatial autocorrelation with low values. The General (Global) G_i statistic computes a single statistic for the entire study area, while the local G_i statistic is an indicator for local autocorrelation for each data point. There are two types of G_i statistics, although almost the two types produce identical results [22] [23]. The first one, G_i, does not include the autocorrelation of a zone with itself, whereas the G_i^* includes the interaction of a zone with itself (*i.e.* the G_i statistic does not include the value of X_i itself, but only the neighborhood values, but G_i^* includes X_i as well as the neighborhood values), and formally both can be computed by the formulae [5]:

$$G_i(d) = \frac{\sum_{j\neq i}^{n} w_{ij}(d)\, x_j}{\sum_{j\neq i}^{n} x_j} \tag{5}$$

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)\, x_j}{\sum_{j=1}^{n} x_j} \tag{6}$$

where, d is the neighborhood (threshold) distance, and w_{ij} is the weight matrix

that has only 1.0 or 0.0 values, 1.0 if j is within d distance of i, and 0.0 if its beyond that distance. These formulae indicate that the cross-product of the value of X at location i and at another location j is weighted by a distance weight, w_{ij} which is defined by either a 1.0 if the two locations are equal to or closer than a threshold distance, d, or a 0.0 otherwise. The G statistic can vary between 0.0 and 1.0. The statistical significance of the local autocorrelation between each point and its neighbors is assessed by the z-score test and the p-value. The expected G value for a threshold distance, d, is defined as:

$$E\left[G(d)\right] = \frac{W}{n(n-1)} \tag{7}$$

where, W is the sum of weights for all pairs of locations ($W = \sum_i^n \sum_j^n w_{ij}$), and n is the number of observations. Assuming normal distribution, the variance of $G(d)$ is defined as [24]:

$$Var\left[G(d)\right] = E\left(G^2\right) - E^2(G) \tag{8}$$

The standard error of $G(d)$ is the square root of the variance of G. Therefore, a z-test can be computed by:

$$S.E.\left[G(d)\right] = \sqrt{Var\left[G(d)\right]} \tag{9}$$

$$Z\left[G(d)\right] = \frac{G(d) - E\left[G(d)\right]}{S.E.\left[G(d)\right]} \tag{10}$$

Where, a positive z-value indicates spatial clustering of high values, while a negative z-value indicates spatial clustering of low values. Sometimes, the G statistic may not follow a normal standard error, and the distribution of the statistic may not be normally distributed, such as the case of a skewed variable with some points having very high values while the majority of other points having low values. In this case, a permutation type simulation should be used [6] [25], with a randomization distribution to test the null hypothesis of no local autocorrelation (H_0). This will maintain the distribution of the variable z but will estimate the value of G under random assignment of this variable, and the user can take the usual 95% or 99% confidence intervals based on the level used.

5. Planar Kernel Density Estimation

Kernel Density Estimation is a non-parametric method to estimate the probability density function of a variable that produces a smooth density surface of point events over a 2-D geographic space (*i.e.* planar space). Kernel density estimations are closely related to histograms, but can be constructed with properties such as smoothness or continuity by using a suitable kernel. The disadvantages of histograms provide the motivation for kernel estimation. When we construct a histogram, we need to consider the width of the bins in which the whole data interval is divided by, and the end points of the bins. As a result, the problems with histograms are that they are not smooth, and therefore we can alleviate these problems by using kernel density estimation that centers a kernel function

at each data point [26]. KDE tends to produce a smooth density surface of point events over space by computing event intensity as density estimation. The general form of a KDE in a 2-D space is given by [8]:

$$\lambda(s) = \sum_1^n \frac{1}{\pi r^2} k \frac{d_{is}}{r} \tag{11}$$

Where $\lambda(s)$ is the density at location s, r is the search radius (bandwidth) of the KDE, k is the weight of a point i at distance d_{is} to location s. The kernel function k is usually considered as a function of the ratio between d_{is} and r. As a result, the longer the distance between a point and location s, the less that point is weighted for calculating the overall density. All points within the bandwidth r of location s are summed for calculating the density at s. A number of distributions can be used to measure the spatial weights k, such as Gaussian, Quartic, Conic, Minimum variance function, negative exponential, and epanichnekov [27] [28]. Some of the mostly used forms of kernel functions are [29]:

The Gaussian function:

$$k \frac{d_{is}}{r} = \frac{1}{\sqrt{2\pi}} EXP\left(-\frac{d_{is}^2}{2r^2}\right) \tag{12}$$

The Quartic function:

$$k \frac{d_{is}}{r} = K\left(1 - \frac{d_{is}^2}{r^2}\right) \tag{13}$$

Where K is a scaling factor to ensure the total volume under Quartic curve is 1.0, and usually used as ¾.

The minimum variance function:

$$k \frac{d_{is}}{r} = \frac{3}{8}\left(3 - 5\frac{d_{is}^2}{r^2}\right) \tag{14}$$

To find the KDE value, two key parameters must be chosen: the kernel function k; and the search radius (bandwidth) r. Many studies have found that the type of the distribution of the kernel function k has a very little effect on the results compared to the choice of search bandwidth r [1] [26] [29] [30] [31]. The value of search bandwidth r is very important because it usually determines the smoothness of the estimated density and can affect the outcome. If it is too small, it will not produce a continuous smooth surface, and too large bandwidth will suppress spatial variation of events. Therefore, the bandwidth of the kernel density estimation often proves to be more influential to outcomes than the kernel shape distribution. Hence, an optimal value of r must be chosen that minimizes the sum of the squared errors of the kernel estimation [32]. There is no unique definition of the optimal search radius, and different optimality criteria have been used. For example, ESRI ArcGIS 10.2 uses the following formula as a default optimal search radius [33]:

$$r = 0.9 \times \min\left(SD, \sqrt{\frac{1}{\ln(2)}} \times D_m\right) \times n^{-0.2} \tag{15}$$

where,

SD: the standard distance

D_m: the median distance

n: the number of points if no population field is used, or if a population field is supplied, n is the sum of the population field values, and min: means that whichever of the two options that results in a smaller value will be used. Silverman [26] suggested a rule of thumb for calculating the optimal search radius as follows:

$$r = SD\sqrt[5]{\frac{4}{3n}} \tag{16}$$

where, the SD is the standard deviation of the samples provided that the kernel function is Gaussian type and that samples follow normal distribution. The cell size depends on the user choice and the dataset. Okabe *et al.* [9] suggested using a cell size of (r/10) as a rule of thumb.

6. Network Kernel Density Estimation

In the real world, there are many kinds of network-constrained events, such as traffic crashes, street crimes, leakages in gas pipe lines along roadways, and river contamination. In planar KDE, the space is characterized as a 2-D homogeneous Euclidian space and density is usually estimated at a large number of locations that are regularly spaced over a grid. However, in analyzing the hot spots of network-constraint events, the assumption of homogeneity of 2-D space does not hold and the relevant KDE methods may produce biased results [9]. Therefore, the planar KDE has been extended to the network KDE, which differs from the planar KDE in several aspects: (i) the network space is used as the point event context; (ii) both search bandwidth r and kernel function k are based on network distance (calculated as the shortest path distance in a network) instead of straight-line Euclidean distance; and (iii) density is measured per linear unit instead of area unit. The network KDE is a 1-D measurement while the planar KDE is a 2-D measurement [9]. The network KDE is an extension of the planar 2-D KDE and it uses the following equation for the density estimation of network-constrained point events in a network space [8]:

$$\lambda(s) = \sum_i^n \frac{1}{r} k\left(\frac{d_{is}}{r}\right) \tag{17}$$

Instead of calculating the kernel density over an area unit, the equation estimates the density over a linear unit, and any of the different forms of kernel functions k may be used.

7. Data

The analysis is conducted on a dataset that presents a road network in the state of Indiana as shown in **Figure 1**. The data includes a crash point layer and a road layer with crash records of the year 2013 that includes all types of crashes (*i.e.* fatal, injury, and property damage). The dataset includes 2983 crash point

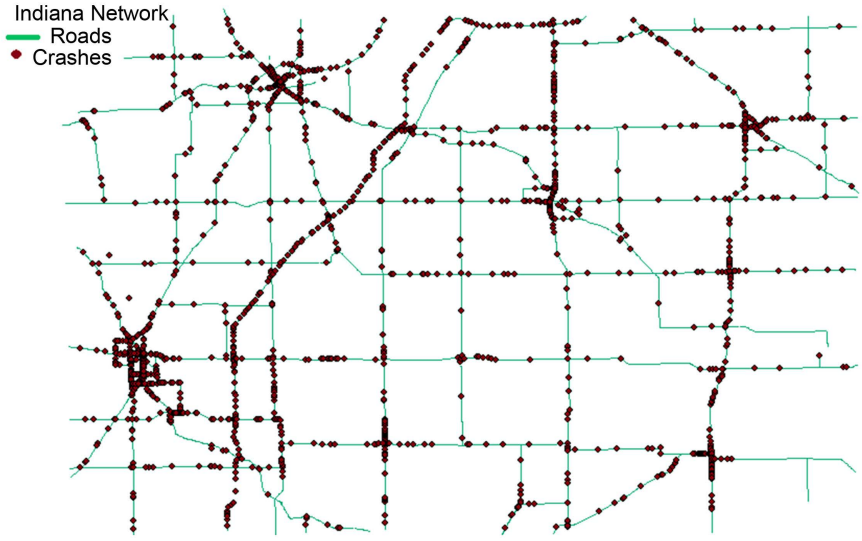

Figure 1. Roads and Crashes of Indiana network used in the analysis.

events on the network.

8. Results and Discussion

The Global Moran's *I* evaluates whether the overall network crashes are clustered, dispersed, or random, and assesses the overall spatial pattern of the crash data. The GIS spatial statistics tool is used to compute the Global Moran's *I*, and five values are generated from running this tool: The Moran's *I* Index, the Expected Index, the Variance, the *z*-score, and the *p*-value as shown in Table 1.

The results of the analysis are interpreted within the context of the null hypothesis. For the Global Moran's *I* statistic, the null hypothesis states that the attributes (*i.e.* crashes) being analyzed are randomly distributed among the features in the study area (*i.e.* no global spatial autocorrelation exists for the entire network). However, since the p-value being generated is less than 0.01 (using a confidence level of 99%), then this indicates that the Global Moran's *I* spatial autocorrelation is significant, and hence, we can reject the null hypothesis, and state that it is quite possible that the spatial distribution of the overall network crashes is the result of clustering pattern, and there is less than 1% probability that this pattern could be the result of random process.

Similarly, the Global (General) Getis-Ord G_i^* statistic evaluates whether the overall network crashes are clustered, dispersed, or random, and assesses the overall pattern and trend of the crash data, and five values are generated from running the ArcMap spatial statistic tool: The General *Gi* statistic, the Expected Index, the Variance, the *z*-score, and the *p*-value as shown in Table 2.

Since the *p*-value being generated is less than 0.01 (using a confidence level of 99%), then this indicates that the General G_i statistic spatial autocorrelation is significant, and hence, we can reject the null hypothesis, and state that it is quite possible that the spatial distribution of the overall network crashes is the result of clustering patterns, and there is less than 1% probability that this pattern

Table 1. Global Moran's *I* Summery for Indiana road network.

Global Moran's *I*	Expected Index	Variance	*z*-score	*p*-value	Decision
0.135847	−0.000335	0.000006	53.817107	0.000000	significant

Table 2. General *Gi* statistic Summery for Indiana road network

General Gi	Expected Index	Variance	*z*-score	*p*-value	Decision
0.128449	0.106109	0.000001	19.233837	0.000000	significant

could be the result of random process. This result is analogous to the Global Moran's *I* in determining the overall clustering pattern of the crashes.

Next, the statistically significant hot spots, cold spots, and spatial outliers are identified using the Anselin Local Moran's *I*, and the local G_i^* statistic. The *z*-scores and *p*-values can be used to evaluate the statistical significance of the computed index values. This method can distinguish between a statistically significant cluster of high values (HH), cluster of low values (LL), and outliers in which a high value is surrounded by low values (HL), and outliers in which a low value is surrounded by high values (LH). Table 3 shows the HH, LL, HL, LH identified by both Moran's *I* and G_i^* statistic. Figure 2 shows the HH, LL, HL, LH identified by Moran's *I*. Figure 3 shows the HH, LL, HL, LH of Moran's *I* with rendering that clearly illustrates the range of the *z*-scores of the identified clusters between the range LL < - 2.0, and the HH > 2.0. It can be seen that the G_i^* statistic has identified a larger number of significant hot spots (157 HHs) and significant cold spots (307 LLs) than the Moran's *I* (102 HHs, 287 LLs). However, Moran's I has identified a larger number of significant outliers (79 HLs, 82 LHs) than the G_i^* (48 HLs, 0.0 LHs).

In addition, the extent and locations of hot spots, cold spots and outliers differ from one method to the other. For example, cluster # 1 is identified by the G_i^* statistic as purely HH hot spot, while it has been identified as a mixed HH and LH hot spot by Moran's *I*. Clusters # 3, 4, and 5 are identified by the G_i^* as purely cold spots, while they have been identified as mixed HH, LH, and non-significant hot spots by Moran's *I*. Clusters # 2 and 6 are identified as non-significant spots by both methods.

Figure 4 shows the HH, LL, HL, LH identified by the G_i^* statistic. Figure 5 shows the HH, LL, HL, LH of G_i^* with rendering that clearly illustrates the range of the *z*-scores of the identified clusters between the range of LL < - 2.0, and the HH > 2.0.

The planar Kernel Density Estimation is determined by the ArcMap spatial analyst tools. Kernel Density calculates the density of a point around each output raster cell, and a smoothly curved surface is fitted over each point. Density surfaces show where point features are concentrated. The surface value is highest at the location of the point being analyzed and decreases with increasing distance from the point, reaching zero at the search radius distance from the point. Figure 6 shows the hot spots identified by the planar KDE, and their average densi-

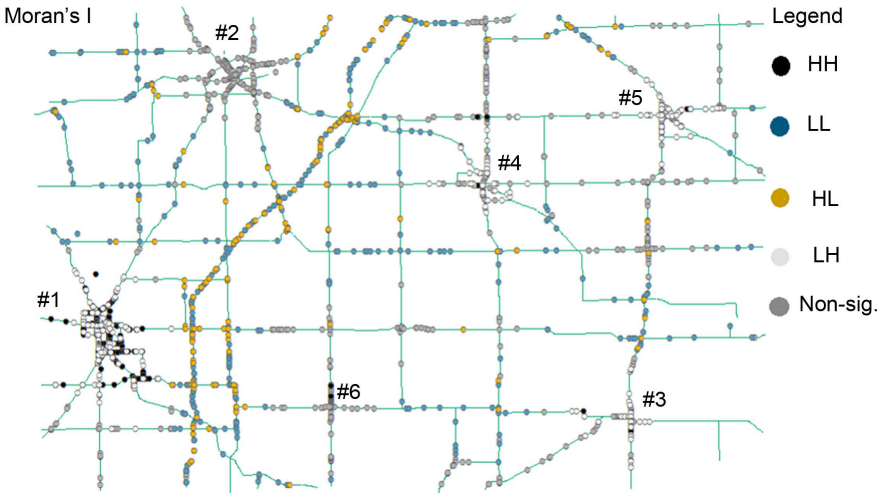

Figure 2. Hot spots and outliers by Anselin Moran's I.

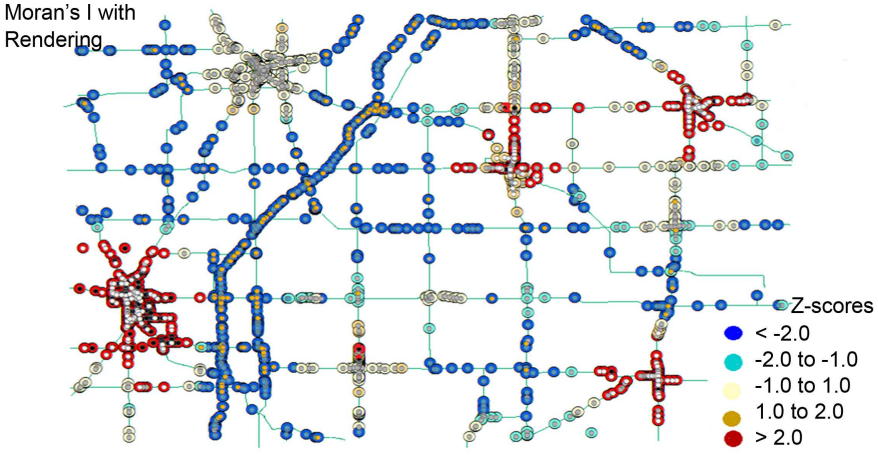

Figure 3. Hot Spots by Moran's I with z-scores rendering.

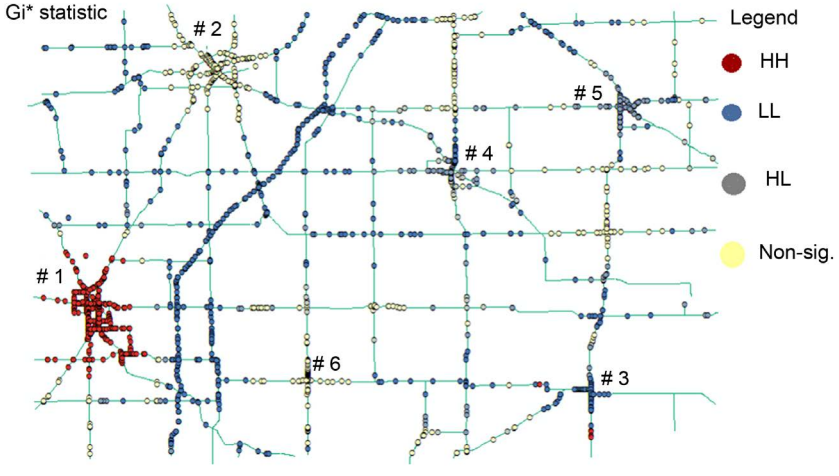

Figure 4. Hot Spots and Outliers by Gi* statistic.

ties per km^2. It can be seen that the planar KDE has identified seven clusters with different crash densities. For example, cluster # 1 contains 8 density levels rang-

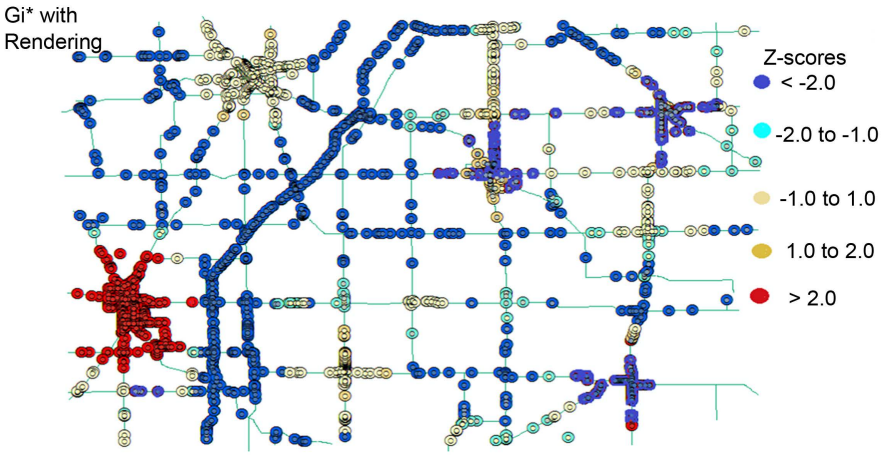

Figure 5. Hot Spots by Gi* statistic with z-scores rendering.

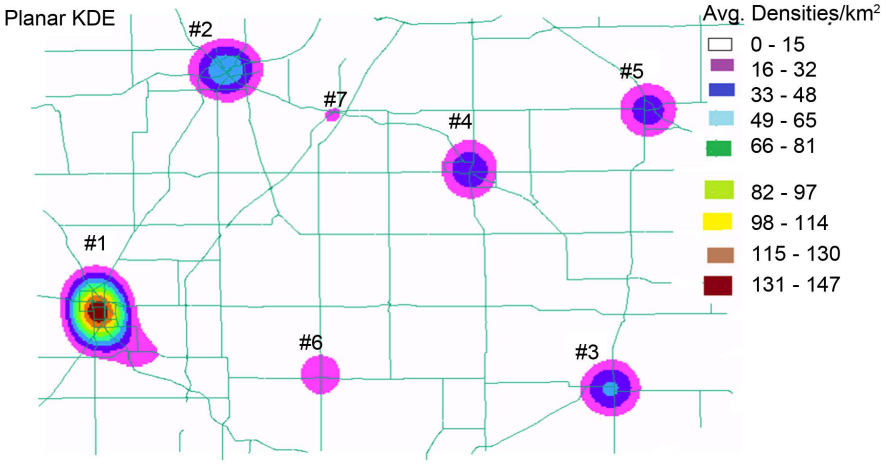

Figure 6. Hot Spots by Planar Kernel Density Estimation.

Table 3. Hot spots, Cold Spots, and outliers of Indiana network by Moran's I and G_i^*.

Method	Hot Spot HH	Cold Spot LL	Outlier HL	Outlier LH
Anselin Moran's I	102	287	79	82
G_i^* statistic	157	307	48	0.0

ing from the highest density value of 147 crashes/km^2 to the lowest density value of 16 crashes/km^2. Cluster # 2 and # 3 contain 3 density levels ranging from the highest density of 65 crashes/km^2 to the lowest density of 16 crashes/km^2. Cluster # 4 and # 5 contain 2 density levels that decreases from the highest value of 48 crashes/km^2 to the lowest value of 16 crashes/km^2. Cluster # 6 and # 7 contain only one density level of at least 16 crashes/km^2. The remaining white raster area contains the minimum density (between 0 – 15 crashes/km^2).

The network-constrained Kernel Density Estimation is determined using the SANET V4.1 software [34]. Traditionally, network events are analyzed with spatial methods assuming Euclidean distance on a 2-D plane, however, this assumption does not hold in practice when analyzing network events, because Euclidean

distances and their corresponding network shortest-path distances are significantly different. Alternatively, network spatial analysis assumes the shortest-path distance on networks that enables more practical investigation of network events than planar spatial analysis. Hence, planar KDE is likely to lead to false conclusions when applied to network events [35]. A clear example is provided in Figure 7, which shows that for a road segment AB, the planar KDE considers 14 crashes within the circular area surrounding the segment, while the network KDE considers only 11 crashes on that segment.

Figure 8 shows the hot spots identified by the network KDE, and their average densities per linear km. Figure 9 shows the hot spots by the network KDE with rendering of their *z*-scores. It can be noticed from Figure 8 that the network KDE has identified different clustering patterns. For example, cluster # 1 is identified by the network KDE as purely high density hot spot similar to the G_i^* pattern. The average density at this junction is (between 4.8 to 5.7 crashes/km). Clusters # 3, 4, and 5 are identified by the network KDE as purely high density hot spots (their density between 4.8 to 5.7 crashes/km), while they have been identified as mixed HH, LH, and non-significant hot spots by Moran's *I*, and purely LL cold spots by G_i^*. Clusters # 2 and 6 are identified by the network KDE as mixed high density hot spots and low density spots (density for the low spots is between 1.8 to 2.7 crashes/km, and density for the high spots is between 4.8 to 5.8 crashes/km), while they are identified as non-significant spots by both Moran's *I* and G_i^* statistic. The crash density for the cold spots is between 2.8 to 4.7 crashes/km as shown in Figure 8.

Since each method has identified different clustering patterns, therefore we recommend using a combination of these methods in hot spot analysis. Comparable results can show more diverse and flexible interpretations among the clustering patterns. Using only one method can result in misleading conclusions. For example, as we saw above, cluster # 1 is identified as a pure HH hot spot in G_i^*

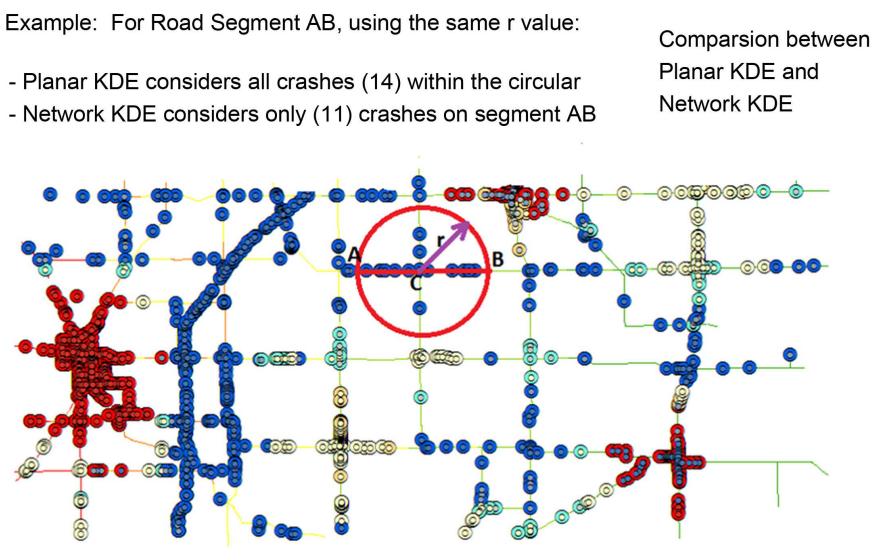

Figure 7. An example of different outcomes between the planar KDE and the network KDE.

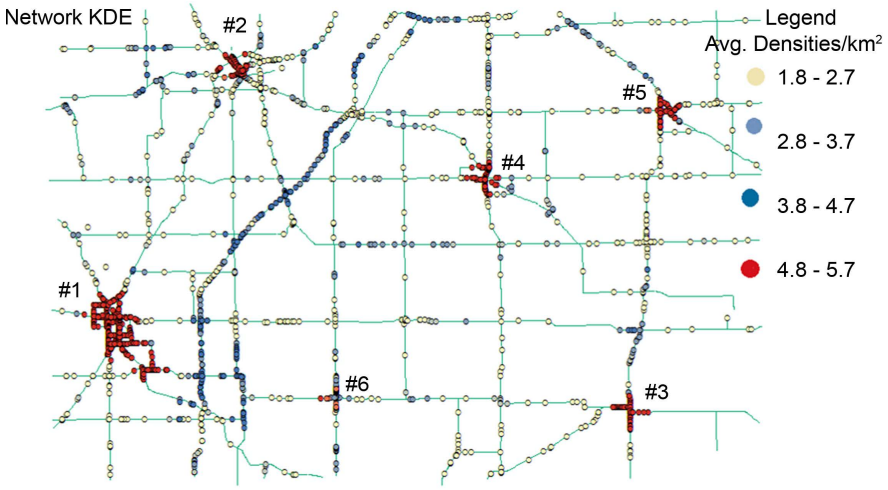

Figure 8. Hot Spots by Network Kernel Density Estimation.

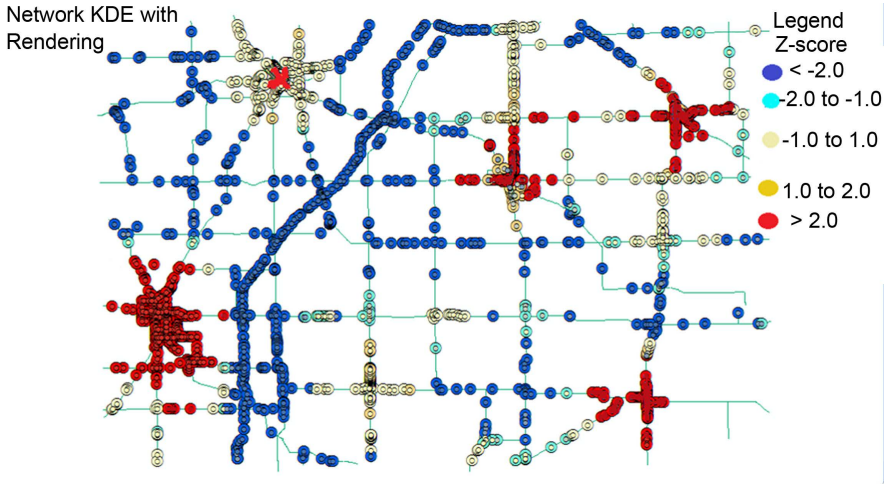

Figure 9. Hot Spots by network KDE with z-score rendering.

and high density spot in network KDE (density from 4.8 to 5.7 crashes/km), while it is identified as a mixed HH and LH in Moran's *I*. Likewise, cluster # 2 is identified as a pure non-significant spot in both Moran's *I* and G_i^*, while it is a mixed HH density spot (with density from 4.8 to 5.7 crashes/km) and low density spot (from 1.8 to 2.7 crashes/km) in network KDE. Hence, using a combination of these methods would probably produce more reliable results than using one method alone. Table 4 shows a comparison between some characteristics of Moran's *I*, G_i^* statistic, and network KDE in identifying hot spots.

9. Conclusion

Hot pot analysis focuses on highlighting areas which have higher than average incidence of events, and it is a valuable technique for visualizing the concentration of events on networks. This paper presented two methods: Moran's *I* and Getis-Ord G_i^* statistic based on network spatial autocorrelation and another third method, kernel density estimation (KDE) to examine the spatial patterns of

Table 4. Comparison between Moran's I, Gi* statistic, and network KDE.

Moran's I	Gi* statistic	Network KDE
measures spatial correlation, identifies hot spots with high-high values, and cold spots with low-low values	measures spatial correlation, identifies hot spots with high-high values, and cold spots with low-low values	measures probability density function, identifies only hot spots in term of density per linear distance unit
Identifies outliers (dispersed incidents) with high-low values and low-high values	does not identify outliers	does not identify outliers
Looking at the value in the context of its neighbors' values within the inverse distance between locations	Looking at the value in the context of its neighbors' values that fall within a specified distance of each other	conduct density calculation based on the user-specified search radius and raster cell size
does not include the interaction of a zone with itself but only with its neighborhoods in measuring spatial correlation	includes the interaction of a zone with itself in addition to its neighborhoods in measuring spatial correlation	does not include the interaction of a zone with itself but only with its neighborhoods in measuring kernel density
reports an index-value, and a z-score	reports a combined index-value and a z-score	reports a linear density value, and a z-score
reports a p-value	reports a p-value	does not report a p-value
presents the statistical significance of clustering	presents the statistical significance of clustering	does not present the statistical significance of clustering
ranges from -1.0 to + 1.0	ranges from 0.0 to + 1.0	any positive value

vehicle crashes and determines if they are spatially clustered, dispersed, or random using the 2013 vehicle crash data in the state of Indiana. The Global values of both Moran's I and G_i^* showed that it is quite possible that the spatial distribution of the overall network crashes is the result of clustering patterns, and there is less than 1% probability that this pattern could be the result of random process. The local Moran's I and G_i^* identified different clustering patterns on the road network. The G_i^* statistic has identified a larger number of significant hot spots (157 HHs) and significant cold spots (307 LLs) than the Moran's I (102 HHs, 287 LLs). However, Moran's I has identified a larger number of significant outliers (79 HLs, 82 LHs) than the G_i^* (48 HLs, 0.0 LHs). The kernel density estimation is evaluated as planar KDE and network KDE. In planar KDE, the space is characterized as a 2-D homogeneous Euclidian space that does not hold when analyzing network events. Therefore, the planar KDE has been extended to the network KDE, which is a 1-D measurement while the planar KDE is a 2-D measurement. In applying the planar KDE to the network, it identified seven clusters with different crash densities. Cluster # 1 contained 8 density levels, cluster # 2 and # 3 contained 3 density levels, cluster # 4 and # 5 contained 2 density levels, and cluster # 6 and # 7 contained only one density level. The network KDE identified different patterns of clusters than what Moran's I and G_i^* statistic have identified. For example, cluster # 1 is identified as a pure HH hot spot in both G_i^* and network KDE (density from 4.8 to 5.7 crashes/km), while it is identified as a mixed HH and LH in Moran's I. Likewise, cluster # 2 is iden-

tified as a pure non-significant spot in both Moran's I and G_i^*, while it is a mixed HH density spot (with density from 4.8 to 5.7 crashes/km) and low density spot (from 1.8 to 2.7 crashes/km) in network KDE. Since each method has identified different clustering patterns, therefore this paper recommends using a combination of these methods in hot spot analysis. Comparable results from these methods can produce more reliable interpretations among the clustering patterns. Using only one method can probably produce misleading results.

References

[1] Bailey, T.C. and Gatrell, A.C. (1995) Interactive Spatial Data Analysis. Addison Wesley Longman Ltd., Harlow, England.

[2] Tobler, W.R. (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, **46**, 234-240. https://doi.org/10.2307/143141

[3] Black, W. (1992) Network Autocorrelation in Transport Network and Flow Systems. *Geographical Analysis*, **24**, 207-222. https://doi.org/10.1111/j.1538-4632.1992.tb00262.x

[4] Cliff, A.D. and Ord, J.K. (1981) Spatial Processes: Models and Applications. Pion, London, UK.

[5] Fischer, M. and Wang, J. (2011) Spatial Data Analysis: Models, Methods, and Techniques. Springer, New York. https://doi.org/10.1007/978-3-642-21720-3

[6] Anselin, L. (1995) Local Indicators of Spatial Association—LISA. *Geographical Analysis*, **27**, 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

[7] Yamada, I. and Thill, J.C. (2004) Comparison of Planar and Network K-Functions in Traffic Accident Analysis. *Journal of Transport Geography*, **12**, 149-158.

[8] Xie, Z. and Yan, J. (2008) Kernel Density Estimation of Traffic Accidents in a Network Space. *Computers, Environment and Urban Systems*, **32**, 396-406.

[9] 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**, 7-32. https://doi.org/10.1080/13658810802475491

[10] Anderson, T.K. (2009) Kernel Density Estimation and K-Means Clustering to Profile Road Accident Hot Spots. *Accident Analysis & Prevention*, **41**, 359-364.

[11] Black, W. and Thomas, I. (1998) Accidents on Belgium's Motorway: A Network Autocorrelation Analysis. *Journal of Transport Geography*, **6**, 23-31.

[12] Flahaut, B. (2004) Impact of Infrastructure and Local Environment on Road Unsafety: Logistic Modeling with Spatial Autocorrelation. *Accident Analysis & Prevention*, **36**, 1055-1066.

[13] Yamada, I. and Thill, J.C. (2007) Local Indicators of Network-Constrained Clusters in Spatial Point Patterns. *Geographical Analysis*, **39**, 268-292. https://doi.org/10.1111/j.1538-4632.2007.00704.x

[14] Schweitzer, L. (2006) Environmental Justice and Hazmat Transport: A Spatial Analysis in Southern California. *Transportation Research Part D: Transport and Environment*, **11**, 408-421.

[15] Erdogan, S., Yilmaz, I., Baybura, T. and Gullu, M. (2008) Geographical Information Systems Aided Traffic Accident Analysis System Case Study: City of Afyonkarahisar. *Accident Analysis & Prevention*, **40**, 174-181.

[16] Yamada, I. and Thill, J.C. (2010) Local Indicators of Network-Constrained Clusters in Spatial Patterns Represented by a Link Attribute. *Annals of the Association of American Geographers*, **100**, 269-285. https://doi.org/10.1080/00045600903550337

[17] Moran, P. (1948) The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society*, **10**, 243-251.

[18] Anselin, L. (1992) Space Stat: A Program for the Statistical Analysis of Spatial Data. National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA.

[19] Goodchild, M.F. (1987) Spatial Autocorrelation. *Concepts and Techniques in Modern Geography*, **48**, 56-63.

[20] Griffith, D.A. (1987) Spatial Autocorrelation: A Primer. Resource Publications in Geography, The Association of American Geographers, Washington DC.

[21] Getis, A. and Ord, J.K. (1992) The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, **24**, 189-206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

[22] Getis, A. and Ord, J.K. (1996) Local Spatial Statistics: An Overview. Geo Information International, Cambridge, England.

[23] Berglund, S. and Karlstrom, A. (1999) Identifying Local Spatial Association in Flow Data Geographical Systems. *Transportation Geography*, **1**, 219-236.

[24] Lee, J. and Wong, D.W.S. (2005) Statistical Analysis with ArcView GIS and ArcGIS. Wiley & Sons, Inc., New York.

[25] Mobley, L.R., Kuo, T.M., Driscoll, D., Clayton, L. and Anselin, L. (2008) Heterogeneity in Mammography Use across the Nation: Separating Evidence of Disparities from the Disproportionate Effects of Geography. *International Journal of Health Geographics*, **7**, 32. https://doi.org/10.1186/1476-072x-7-32

[26] Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis. Chapman Hall, London.

[27] Levine, N. (2004) CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations. Ned Levine & Associates—The National Institute of Justice, Houston, TX; Washington DC.

[28] Gibin, M., Longley, P. and Atkinson, P. (2007) Kernel Density Estimation and Percent Volume Contours in General Practice Catchment Area Analysis in Urban Areas. *Proceedings of the GIScience Research UK Conference* (*GISRUK*) 2007, Maynooth, Ireland.

[29] Schabenberger, O. and Gotway, C.A. (2005) Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC, Boca Raton, FL.

[30] O'Sullivan, D. and Unwin, D.J. (2002) Geographic Information Analysis. John Wiley, Hoboken, NJ.

[31] O' Sullivan, D. and Wong, D.W.S. (2007) A Surface-Based Approach to Measuring Spatial Segregation. *Geographic Analysis*, **39**, 147-168. https://doi.org/10.1111/j.1538-4632.2007.00699.x

[32] Scott, D.W. (1992) Multivariate Density Estimation, Theory, Practice, and Visualization. John Wiley & Sons, New York. https://doi.org/10.1002/9780470316849

[33] ArcGIS Resources. http://resources.arcgis.com/en/help/main/10.2/index.html#//009z00000011000000

[34] Okabe, A., Okunuki, K. and Shiode, S. (2006) The SANET Toolbox: New Methods for Network Spatial Analysis. *Transactions in GIS*, **10**, 535-550. https://doi.org/10.1111/j.1467-9671.2006.01011.x

[35] Okabe, A. and Sugihara, K. (2012) Spatial Analysis along Networks: Statistical and Computational Methods. John Wiley & Sons, New York.
https://doi.org/10.1002/9781119967101