Spatial Point Analysis of Road Crashes in Shanghai: A GIS-based Network Kernel Density Method

Becky P.Y. Loo[1], Shenjun Yao[1*], Jianping Wu[2]
[1]Department of Geography, University of Hong Kong, Hong Kong
[2]Key Lab of Geographical Information Science, Ministry of Education, East China Normal University, Shanghai, China
[*]Corresponding author, e-mail: shenjun_yao@126.com

***Abstract*—As road crashes are constrained to a one-dimensional space, this paper analyzes the spatial distribution of road crashes with a GIS-based network-constrained kernel density method. A dissolving procedure is introduced before road segmentation, which can significantly reduce the undesirable effects during the segmentation process. The result of the sensitivity analysis reflects that the bandwidth imposes great impacts on the spatial distribution of density estimates. Different bandwidths may be considered for different types of traffic crashes. In particular, vehicle-pedestrian crashes in downtown areas tend to be highly localized and a narrower bandwidth is more appropriate. Vehicle-vehicle crashes at the suburb and rural areas, however, tend to happen in a less concentrated manner along a continuous stretch of dangerous road segments; and a wider bandwidth is more powerful in identifying these hot zones. Based on our results, administrations can gain more information on hazardous road locations, conduct investigations and propose improvement measures.**

Keywords-network; crashes; trafffic; kernel density; GIS

I. INTRODUCTION

World Health Organization estimated that approximately 1.3 million people died each year on roads, and that between 20 and 50 million sustained non-fatal injuries [1]. Road traffic injuries remain one of the most critical public health problems, especially for developing countries such as India and mainland China. In 2009, there were more than 200,000 crashes in mainland China, causing over 65,000 deaths and 275,000 injuries [2]. As traffic crashes entail socio-economic costs and create a heavy burden on the society, road safety has been a major concern of the Chinese Government. In Shanghai -- the most populous city in China, road safety performance is unsatisfactory. According to the Ministry of Public Security of the People's Republic of China, the annual average number of crashes per 1,000 km of roads in Shanghai during the period from 2006 to 2008 was about 289.24, ranking first among all 31 provinces, autonomous regions and municipalities [3]. Against this background, this study analyzes the spatial distribution of traffic crashes in the urban and suburban areas of Shanghai. In particular, this research focuses on two major types of crashes, vehicle-pedestrian and vehicle-vehicle road crashes. We attempt not only to identify crash clusters (known as "black spots" or "hot spots") for each type of road crashes, but also to explore whether the spatial patterns of the two categories of crashes vary systematically within the study area.

A better understanding of the spatial pattern of crashes plays a key role in addressing road safety problems. Spatial analysis on traffic crashes has thus received a great deal of attention in the literature. As road crashes can be regarded as a network-constrained phenomenon, research can be categorized into two main categories. One group can be described as link-attribute-based analyses. Researchers aggregate road crashes by road links and perform spatial analysis based on the spatial structure and attributes of road segments such as the number of crashes [4-7]. The work reported in this paper belongs to the second category, known as event-based analyses. This type of studies focuses on the spatial distribution of individual crashes along the road network by using one-dimensional methods for event-based analysis [8-11]. This study examines the spatial distributions of vehicle-vehicle and vehicle-pedestrian crashes by using a Network Kernel Density Estimation (NKDE) approach. As it is not practical to examine the density for every possible location over the entire road network, Xie and Yan [10] suggested dividing the road network into basic linear units with equal interval (*lixel*) or basic spatial units (BSUs). Density values are only calculated for center points of the BSUs (named "reference points" hereafter). However, for a link-node network system, the equal interval condition is likely to be violated at junctions or near end nodes of roads, resulting in a great number of BSUs with lengths shorter than the interval. Moreover, they snapped road crashes to the nearest reference points and only measured distances among reference points. Though the algorithm improves computational efficiency, it affects the measured distance and resultant density values. This research tries to bridge these gaps by introducing a dissolving procedure before segmentation and calculating network distance with the precise locations of crashes.

The examination of spatial heterogeneity of different crash patterns is useful when crash analysts are interested in hazardous road locations for one particular type of road crashes in comparison with another category of crashes. The ratio of Kernel Density Estimation is used to compare density functions for a pair of point patterns observed over the same study area [12]. Bithell suggested a logistic transformation of the density ratio to symmetrize variances [13]. Following Kersall and Diggle [14], this study explores the spatial heterogeneity of two crash patterns by using the natural logarithm of the ratio of spatial densities for vehicle-pedestrian and vehicle-vehicle crashes.

Supported by the Research Grant Council (HKU 747809H)

978-1-61284-848-8/11/$26.00 ©2011 IEEE

II. METHODOLOGY

A. Data Collection and Extraction

The principal database for this study is the police crash data of 2006 from Shanghai 110 Calling Center. The database contains crash information such as time, location and road user type. In particular, the database records X and Y coordinate with Shanghai Local Projected Coordinate System, whereby crashes can be conveniently plotted onto GIS maps. We collected information about 278,307 vehicle-vehicle crashes and 18,602 vehicle-pedestrian crashes within the urban (within the Middle Ring) and suburban (between the Outer Ring and the Middle Ring) areas of Shanghai. Another crucial database is the link-node road network system of Shanghai for the year 2006. The database records road information including road name and road type. We extracted roads within the study area, which was about 4,400 kilometers long with 17,153 links in total.

B. Analytical Tools

In this section, the basic algorithm for NKDE and the log density ratio is firstly introduced. It is followed by a discussion on some key implementation issues.

1) Basic algorithm

The procedures presented below show the general steps for the NKDE and the log ratio:

a) Cut the entire road network into BSUs with equal interval l.

b) Obtain center points of BSUs as reference points.

c) For each reference point i, calculate the density estimate $f_{vp}(i)$ for N_{vp} vehicle-pedestrian crashes and $f_{vv}(i)$ for N_{vv} vehicle-vehicle crashes by:

$$f_{vp}(i) = \frac{1}{N_{vp} b} \sum_{j=1}^{N_{vp}} Kern(\frac{d_{ij}}{b}), \quad (1)$$

$$f_{vv}(i) = \frac{1}{N_{vv} b} \sum_{j=1}^{N_{vv}} Kern(\frac{d_{ij}}{b}), \quad (2)$$

where b is the bandwidth (searching distance) with length no less than half of l; d_{ij} is the network distance between reference point i and crash j; Kern(.) is a kernel function that will be discussed in the following sub-section.

d) For each reference point i, calculate the natural logarithm of the ratio of the two spatial densities r(i) by:

$$r(i) = \log \frac{f_{vp}(i)}{f_{vv}(i)}. \quad (3)$$

Note that $f_{vp}(i)$ and $f_{vv}(i)$ need to be calculated with the same bandwidth and kernel function.

2) Key implementation issues

As mentioned earlier, the network segmentation (Step 1 of the algorithm) can produce a large number of small BSUs (< *l*), particularly in a high-density link-node road network system. In this study, there were 17,008 (32.5% of the total) short BSUs, even though the road network is divided with 100m interval. A practical approach used here is to dissolve road links based on road name and road type before the segmentation process. After the dissolving performance, the total number of links in the road network database is reduced from 17,153 to 6,830. If the roads are still segmented with 100m, the percentage of BSUs with length less than 100m can be decreased to 14.3%.

Another issue is the choice of kernel function. A number of kernels can be used to measure the "distance decay effect", such as Uniform, Triangle, Quartic (biweight), Triweight and Gaussian [12]. It is generally accepted that the form of kernel function is not critical [14]. Here we choose Quartic function, a commonly used kernel form is determined by:

$$Kern(\frac{d_{ij}}{b}) = \begin{cases} \frac{15}{16}(1 - \frac{d_{ij}^2}{b^2})^2 & if\ 0 < \frac{d_{ij}}{b} \le 1 \\ 0 & otherwise \end{cases}. \quad (4)$$

Although the precise form of *Kern*(.) weakly influences density estimates, the bandwidth can have a profound impact. While large bandwidths result in smoothing estimate, small bandwidths retain more local characteristics. In this regard, this study will conduct a sensitivity analysis of spatial pattern to the choice of bandwidth.

Focusing on Equation (1), (2) and (4), one may notice that the kernel function can produce zero estimates for either $f_{vp}(i)$ or $f_{vv}(i)$. Zero estimates for $f_{vv}(i)$ lead to instabilities (division by zero) in the resultant r(*i*). Following Bithell [13], this study introduces a modification to r(*i*) by adding a small constant to both $f_{vp}(i)$ and $f_{vv}(i)$, which can be defined as:

$$r(i)_\sigma = \log \frac{f_{vp}(i) + \sigma}{f_{vv}(i) + \sigma}. \quad (5)$$

III. RESULTS

With 100 meters as *the BSU* interval, the entire road network was divided into 47,181 segments. 47,181 reference points were derived. For each reference point, the density values for vehicle-pedestrian and vehicle-vehicle crashes were calculated. Four bandwidths (100m, 250m, 500m and 1000m) were used with the same kernel function (Quartic function). Fig. 1 shows the results of density patterns for vehicle-pedestrian and vehicle-vehicle crashes with the four different bandwidths. It is observed that bandwidth exerts great impacts on the network density pattern. As shown in Fig. 1, the density pattern becomes smoother with increasing bandwidth. It appears that the narrow bandwidth (100m and 250m) preserves more local features and is therefore more appropriate for identifying crash

clusters at precise locations ("hot spots"). As the bandwidth is increased from 250m to 1000m, the local hot spots gradually combined with their neighbors, resulting in crash clusters at larger spatial scales ("hot zones" or even "hot areas"). Using density patterns with the bandwidth of 1000m as an example, crashes are clustering nearly everywhere in the north and west of urban area within the Inner Ring of the city. As one of the purposes of this research is to identify the spatial pattern of crashes at relatively small spatial scales (that is, to identify crash hot spots), narrow bandwidth (100m or 250m) may be more suitable for exploring spatial distributions of crashes. In 2001, the Ministry of Public Security conducted a project for the identification of hazardous road locations in mainland China, which suggested 500m as the spatial scale of crash hot spots [15]. Following this project, 250m is chosen in this study as the searching bandwidth.

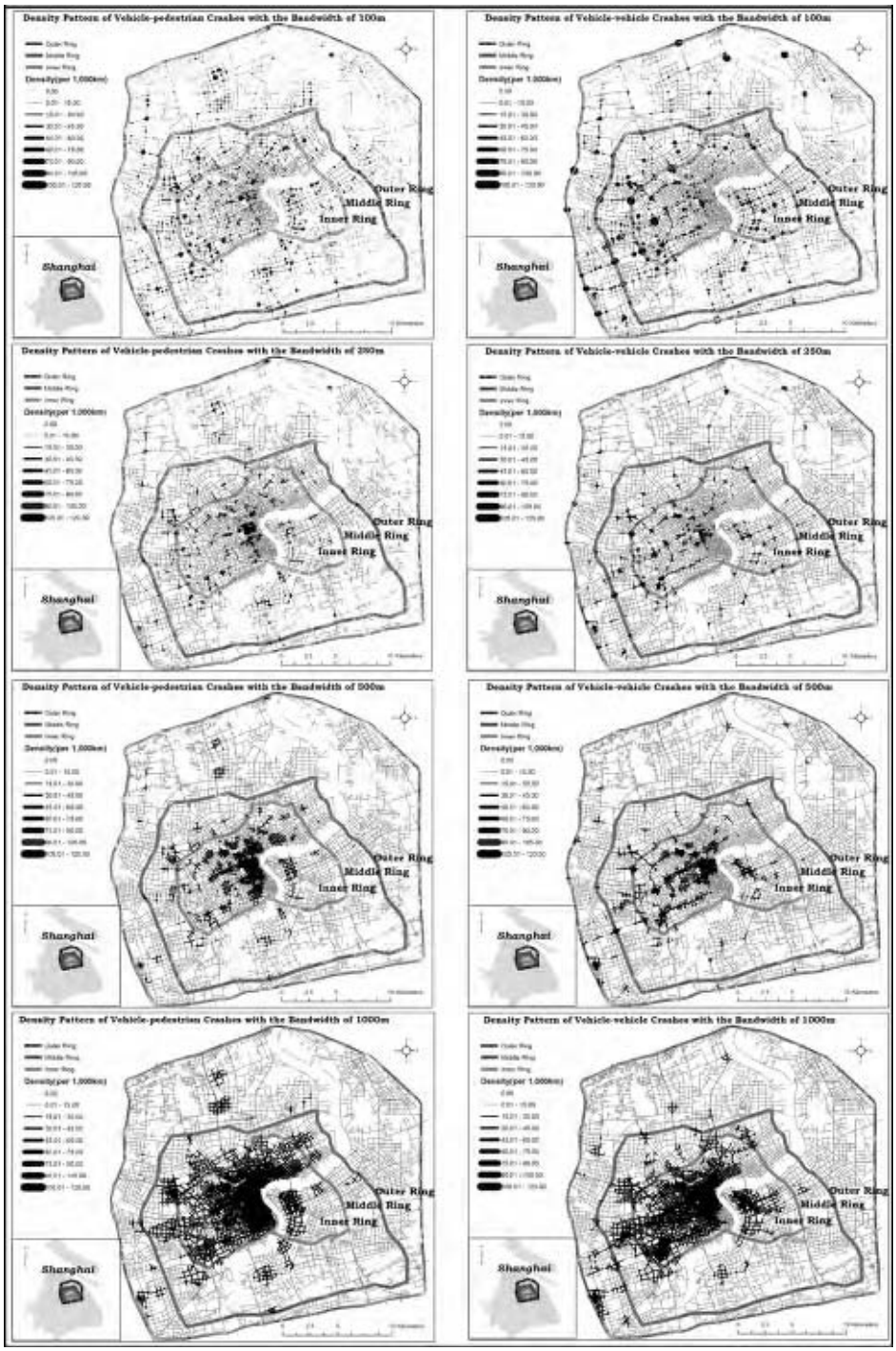
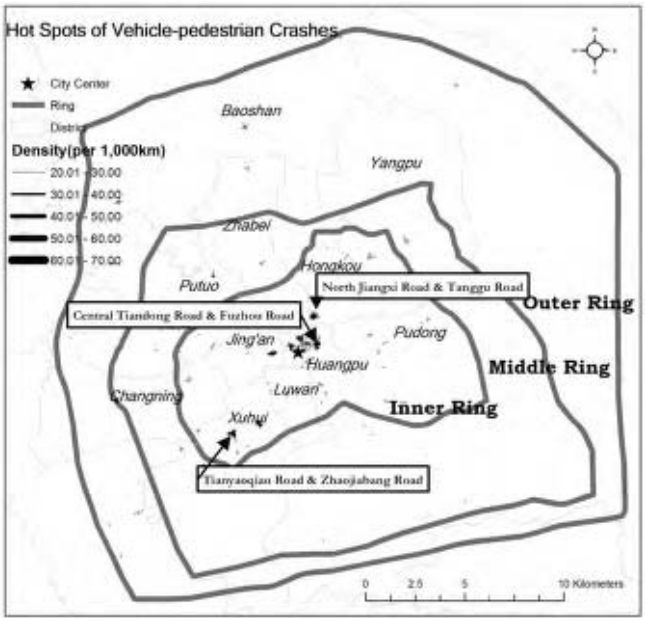
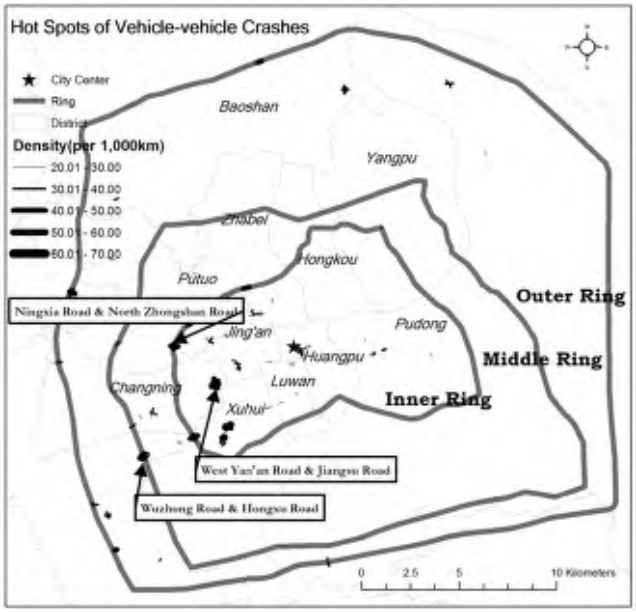


Figure 1. Density patterns for vehicle-pedestrian and vehicle-vehicle crashes with bandwidths of 100m, 250m, 500m and 1km.

Fig. 2 delineates hot spots of vehicle-pedestrian and vehicle-vehicle crashes by using 99 percentiles of density estimates as the cut-off value. Three "hottest" spots were labeled with road names. For both types of crashes, hot spots were more likely to be detected at road junctions. Focusing on vehicle-pedestrian crashes in Fig. 2a, one may observe that there were a number of hot spots at the city center. The reason might be that the most important commercial center of Shanghai, including the People's Square and Nanjing Street, are located in this area. Crashes involving pedestrians happened more frequently in these places mainly due to the higher exposure of pedestrians. While vehicle-pedestrian crashes had higher incidence on the secondary trunk and branch roads in commercial areas, a great number of vehicle-vehicle crash hot spots were found on expressways such as A20 Expressway (part of Outer Ring) and arterial roads such as Wuzhong Road (part of Middle Ring) and North Zhongshan Road (part of Inner Ring), probably because of the faster speed and heavy traffic volume.



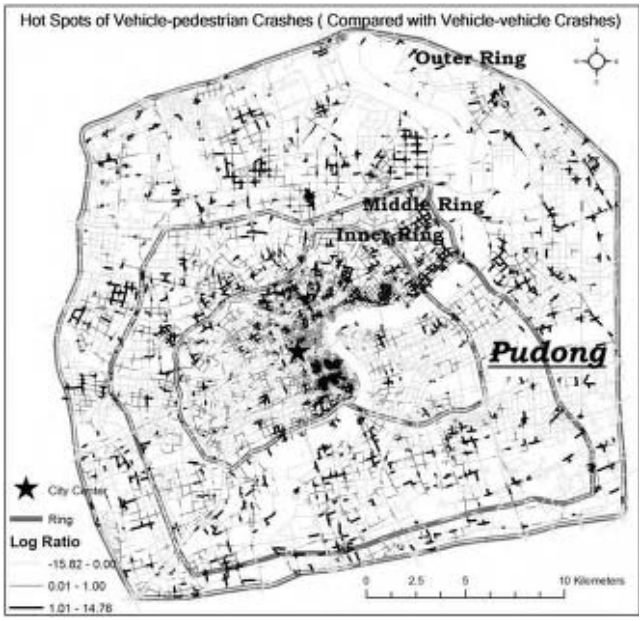
(a)



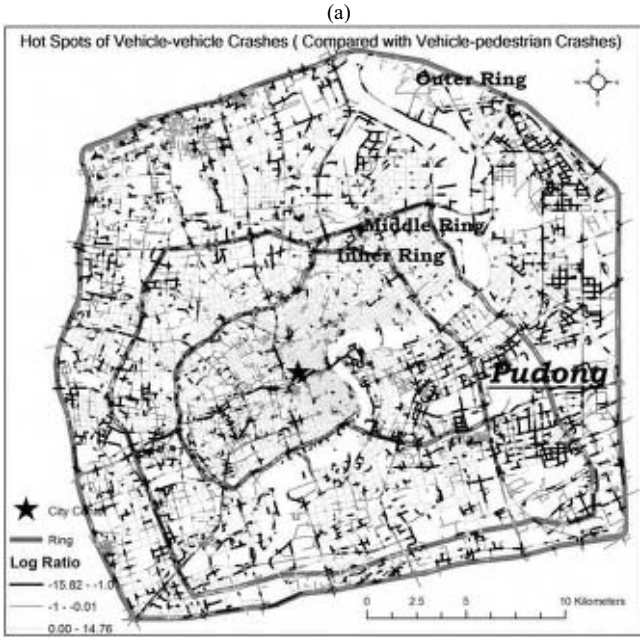
(b)

Figure 2. Hot spots of vehicle-pedestrian and vehicle-vehicle crashes.

In order to gain a better understanding of the spatial heterogeneity of vehicle-pedestrian and vehicle-vehicle distributions, the log density ratio was computed with σ equal to 0.000001. If one targets vehicle-pedestrian crashes and defines 1 as the threshold value, BSUs with log ratio more than 1 can be identified as hazardous road locations for pedestrians. Fig. 3a highlights hot spots (colored in black) of vehicle-pedestrian crashes in comparison with vehicle-vehicle crashes. Of these hazards, almost 70% were located in urban area (within the Middle Ring). The central urban area (within the Inner Ring) where the central business district (CBD) is situated had significantly higher clustering tendency of vehicle-pedestrian crash hot spots, due probably to higher exposure of pedestrians. Fig. 3b describes spatial pattern of log density ratios for vehicle-vehicle crashes, in which BSUs with log ratio less than -1 were regarded as hot spots. Unlike vehicle-pedestrian crashes, hot spots of vehicle-vehicle crashes were clustering in suburban area. On one hand, people in Shanghai still preferred to live in urban places, leaving the suburban area sparsely populated. Lower exposure of pedestrians might decrease the incidence of crashes involving pedestrians. On the other hand, expressways and many arterial roads are located in suburban area. The primary function of these roads is to deliver traffic from urban center to rural places or even to other cities. Heavy traffic and high speed might lead to more crashes occurring on these suburban locations.



(a)



(b)

Figure 3. Spatial pattern of log density ratio.

IV. CONCLUSIONS

As road crashes are constrained to a one-dimensional space, this paper analyzes the spatial distribution of road crashes in Shanghai with a network-constrained kernel density method. One key feature is that the entire road network is dissolved before segmentation. Such performance can significantly reduce the negative effects of the equal interval condition and the complex structure of link-node road system on the density pattern. It is also found that the bandwidth imposes great impacts on spatial distribution of density estimates. Using a narrow bandwidth that reveals more local effects, hot spots within urban and suburban area are identified. Based on the results, administrations can gain more information on hazardous road locations, conduct investigations and propose improvement measures such as installing pedestrian refuges or speed cameras.

It should be pointed out that the density value is only calculated for the reference point and is used to represent the whole corresponding BSU, in a similar way of planar kernel density method that only computes the estimate for the center point of a raster cell. This study chooses a relatively small value, 100m, as the segmentation interval. As a step further, the influences of the length of interval on network-constrained density estimates may be carefully examined. In addition, the road network is characterized by various spatial and non-spatial factors such as traffic volume, number of lanes and land use type. More research efforts are needed to identify the variables that could explain the spatial distribution of road crashes.

ACKNOWLEDGMENT

The authors would like to thank Mr. Lei Fang for his valuable comments and other help.

REFERENCES

[1] "Global status report on road safety," Switzerland: World Health Organization, 2009.

[2] "Statistics on road traffic crashes in 2009," available at http://www.mps.gov.cn/n16/n1252/n1837/n2557/2276407.html, Ministry of Public Security of the People's Republic of China, 2010.

[3] China's Road Crashes Statistical Book, Ministry of Public Security of the People's Republic of China, 2006, 2007, 2008.

[4] W. R. Black and I. Thomas, "Accidents on Belgium's motorway: a network autocorrelation analysis," Journal of Transport Geography, vol. 6 (1), 1998, pp: 23-31.

[5] B. P. Y. Loo, "The identification of hazardous road locations: a comparison of the blacksite and hot zone methodologies in Hong Kong," International Journal of Sustainable Transportation, vol. 3 (3) , 2009, pp: 187-202.

[6] E. Moons, T. Brijs, and G. Wets, "Identifying hazardous road locations: hot spots versus hot zones," Transactions on Computational Science, vol. 5730, 2009, pp: 288-300.

[7] I. Yamada and J. C. Thill, "Local indicators of network-constrained clusters in spatial patterns represented by a link attribute," Annals of the Association of American Geographers, vol. 100(2), 2010, pp: 269-85.

[8] A. Okabe, K. Okunuki, and S. Shiode, "SANET: A toolbox for spatial analysis on a network," Geographical Analysis, vol. 38(1), 2006, pp: 57-66.

[9] I. Yamada and J. C. Thill, "Local indicators of network-constrained clusters in spatial point patterns," Geographical Analysis, vol. 39(3), 2007, pp: 268-92.

[10] Z. Xie and J. Yan, "Kernel density estimation of traffic accidents in a network space," Computers Environment and Urban Systems, vol. 32(5), 2008, pp: 396-406.

[11] A. Okabe, T. Satoh, and K. Sugihara, "A kernel density estimation method for networks, its computational method and a GIS-based tool," International Journal of Geographical Information Science, vol. 23(1), 2009, pp: 7-32.

[12] L. Waller and C. Gotway, Applied Spatial Statistics for Public Health Data. New York: Wiley, 2004.

[13] J. F. Bithell, "An application of density-estimation to geographical epidemiology," Statistics in Medicine, vol. 9(6), 1990, pp: 691-701.

[14] J. E. Kersall and P. J. Diggle, "Kernel estimation of relative risk," Bernoulli,vol. 1, 1995, pp: 3-16.

[15] F. Lu, W. Jiang, and S. Ma, "Locations identification of hazardous road," Journal of Chang'an University(Natural Science Edition), vol. 23(1), 2003, pp: 97-90.