Identifying Traffic Risk Hotspots Using Spatial-temporal Network Kernel Density Estimation: A Novel Optimal Parameter Selection Method with Dual Dataset Validation

Shengtai Yao
Department of Mechanical Engineering
Tsinghua University, Beijing, P.R.China, 100084
Email: yaost19@mails.tsinghua.edu.cn

Huiping Li
Department of Civil Engineering
Tsinghua University, Beijing, P.R.China, 100084
Email: lihp11@tsinghua.org.cn

Xiao Hu
Department of Civil Engineering
Tsinghua University, Beijing, P.R.China, 100084
Email: x-hu21@mails.tsinghua.edu.cn

Klaus Hermann
Research and Development Hardware for Automated Driving
Mercedes-Benz Group China Ltd. Beijing, P.R. China, 100102
Email: klaus.kh.hermann@mercedes-benz.com

Ke Zhang
Department of Civil Engineering
Tsinghua University, Beijing, P.R.China, 100084
Email: zhangkethu@mail.tsinghua.edu.cn

Yunxuan Li
Beijing Key Laboratory of Traffic Engineering
Beijing University of Technology, Beijing, P.R.China, 100124
Email: liyunxuan@bjut.edu.cn

Meng Li, Corresponding Author
Department of Civil Engineering
Tsinghua University, Beijing, P.R.China, 100084
Email: mengli@tsinghua.edu.cn

Word Count: 4851 words + 2 table(s) × 250 = 5351 words
Submission Date: September 28, 2025

ABSTRACT

Traffic risk assessment is crucial for enhancing urban traffic safety and efficiency. However, traditional research encounters the bottleneck that hotspot identification is less liable to verify accurately in the spatial-temporal network. In this study, we propose a novel approach for traffic risk hotspot detection in urban areas by leveraging two complementary datasets: probe vehicle event data and user-reported traffic event data. First, we employ a novel spatial-temporal network kernel density estimation (ST-NKDE) to identify risk hotspots in the road network. Through a systematic parameter selection process, we determine the optimal bandwidth values for distance and time. Furthermore, we employ Jensen Shannon divergence (JSD) to quantify the differences between the risk distributions from the two datasets. The visualization of risk maps with the comparison of parameters further demonstrate the consistency and reliability of our traffic risk hotspot detection model. Generally, by integrating two datasets, this research contributes to the advancement of urban traffic safety measures and has favorable implications for urban planning and traffic management strategies.

Keywords: Spatial-temporal network kernel density estimation (ST-NKDE); Optimal parameters; Risk hotspot visualization; Dual dataset validation;

INTRODUCTION

Urban transportation systems play a crucial role in modern society. However, with the rapid urbanization and increasing vehicle volume, traffic safety has become a significant concern for both policymakers and the general public. Traffic incidents, such as collisions, congestion, and road closures, not only result in economic losses but also bring severe risks to human lives. Therefore, identifying traffic risk hotspots and understanding the underlying patterns of risky driving behavior and traffic incidents are of great importance to enhance traffic safety and optimize transportation infrastructure (*1*).

In recent years, data-driven approaches have been powerful tools for traffic risk hotspot detection (*2*). These methods utilize large-scale and various datasets, including floating car data and traffic event data, to model and predict traffic risk in both spatial and temporal dimensions. Among these approaches, Kernel Density Estimation (KDE) has been widely used for its ability to estimate the probability density function of traffic risk in a continuous space. Traditional KDE methods consider the spatial distribution of traffic events on a 2D plane, but in the context of road networks, such an approach may oversimplify the risk estimation and fail to capture the complexities of road topologies.

To address these limitations, we use a novel approach, namely Spatial-Temporal Network Kernel Density Estimation (ST-NKDE), to identify traffic risk hotspots in urban transportation networks. ST-NKDE extends traditional KDE to a network space, where the risk is considered to spread along road links rather than evenly distributed on a 2D plane. Moreover, by incorporating both spatial and temporal dimensions, ST-NKDE allows for a more comprehensive representation of risk patterns, leading to accurate hotspot detection and risk assessment.

The primary contributions of this paper are as follows:

(a). Novel ST-NKDE Method with Optimal Parameter Selection: We present the ST-NKDE method, which integrates spatial and temporal information in a road network to estimate traffic risk distributions. Moreover, we propose a systematic approach to determine the optimal bandwidth parameters for the KDE process, ensuring the effectiveness of the hotspot detection results.

(b). Dual Dataset Validation: To demonstrate the robustness and generalizability of our method, we validate ST-NKDE on two heterogeneous datasets. The first dataset comprises large-scale, long-term probe vehicle data, providing insights into risky driving behaviors across the urban road network. The second dataset consists of user-reported traffic event data, capturing the occurrence of various traffic incidents. Through this dual dataset validation, we verify the applicability of ST-NKDE and the consistency of its performance across different data sources.

(c). Similarity Analysis of Dual Datasets: We conduct a similarity analysis using Jensen Shannon divergence (JSD) to examine the relationship between risky driving behavior and actual traffic incidents. By comparing the risk distributions obtained from the probe vehicle data and traffic event data, we uncover the correlation between risk hotspots and reported incidents, revealing the underlying patterns and factors contributing to traffic risk.

In the remainder of this paper, we present a comprehensive review of existing literature in traffic hotspot detection. Subsequently, we introduce the methodology of ST-NKDE employed in this study. Additionally, we provide the details regarding the process of optimal parameter selection. Afterwards, we present the results of the dual dataset validation and the similarity analysis. Finally, the article concludes with a discussion and summary of our findings, highlighting the contributions and potential future directions of our research.

LITERATURE REVIEW

Traffic risk hotspot identification

Traffic risk hotspot identification is a crucial to identify locations with relatively high risk levels for traffic incidents. Over the years, various methods have been proposed for hotspot identification, each with its strengths and limitations.

One category of hotspot identification methods is based on statistical analysis of historical traffic incident data (2, 3). For instance, logistical regression models have been commonly used to model the influencing factors of traffic incidents and identify hotspots based on the about 400 sets of accidents data of 10 major roads in Beijing city (4). Additionally, Geographic Information System (GIS) techniques have been employed to analyze spatial patterns and detect clusters of traffic accidents (5, 6) . While these statistical methods have been widely used, they often neglect the temporal dimension and the underlying network structure of road systems.

Another approach to hotspot identification involves the use of machine learning and data mining techniques. Machine learning models (7–9), such as Support Vector Machines (SVM) (10), Random Forest (7, 11, 12), and Artificial Neural Networks (ANN) (13), have been applied to predict traffic incidents and identify hotspots. These methods can capture complex relationships between risk factors and traffic incidents, but they may require a large amount of labeled data for training and may be computationally expensive.

However, most of the existing methods overlook the spatial and temporal interactions among traffic incidents and the network structure of road networks, limiting their effectiveness in capturing the dynamic nature of traffic risk.

Kernel density estimation

Kernel Density Estimation (KDE) is a widely used non-parametric method for estimating probability density functions from data. In the context of traffic risk hotspot identification, KDE can be applied to estimate the probability density function of traffic incidents in a spatial context, assuming the road network as a 2D plane (14). KDE has been successfully used in various applications, including traffic accident analysis (15). However, traditional KDE methods overlook the network structure of road networks, which can significantly impact the spatial distribution of traffic incidents.

To address the limitations of traditional KDE methods, researchers have introduced

Network Kernel Density Estimation (NKDE) (*11*, *16*–*18*) and Spatial-Temporal Kernel Density Estimation (ST-KDE) techniques (*19*, *20*). NKDE considers the road network as a graph and incorporates spatial relationships and connectivity among road segments. Chen et al. applied an NKDE method for traffic risk hotspot detection, where traffic incidents are modeled as spreading along the road links rather than evenly distributed on a 2D plane (*16*). This approach improves hotspot detection accuracy by accounting for network topology. ST-KDE is an improved version of KDE that incorporates spatial-temporal information, which allows for more accurate and efficient representations of data patterns and trends over time and space (*19*).By combining the advantages of the two types of method, ST-NKDE further extends the NKDE approach by incorporating the temporal dimension, enabling the modeling of traffic incidents over both space and time. It allows for the identification of risk hotspots that vary with time and location (*21*). Therefore, ST-NKDE provides a more comprehensive understanding of traffic risk patterns and dynamics.

The effectiveness of KDE-based models, including ST-NKDE, depends on the selection of optimum parameters, particularly the bandwidth. The bandwidth determines the diffusion range of each data point and influences the overall estimation of risk density (*22*). Finding the optimal bandwidth is challenging, as it involves balancing the trade-off between under-smoothing and over-smoothing the data.Various methods have been proposed for selecting the optimal bandwidth. Cross-validation techniques, such as K-fold cross-validation, can be used to evaluate the performance of different bandwidth values and select the one that minimizes prediction errors. Grid search is another common approach, where a range of bandwidth values is specified, and the best value is determined through an exhaustive search. Moreover, heuristic methods, such as Scott's Rule and Silverman's Rule, provide guidelines for bandwidth selection based on the characteristics of the data.

To sum up, while ST-NKDE is a promising approach for traffic risk hotspot identification, there is a research gap in its practical application and the selection of its best parameters. More specifically, the real-world application and effectiveness of ST-NKDE on diverse datasets from different cities and regions need further exploration to ensure its generalizability and robustness. Furthermore, the aforementioned methods for determining optimal parameters are typically used in supervised learning or purely based on experience. For the hotspot identification, which is an unsupervised learning, there still lacks efficient methods to obtain the optimal parameters. Thus, this paper will propose an approximate approach to determine the optimal parameter values.

METHODOLOGY

In this paper, a spatial-temporal network kernel density estimation (ST-NKDE) method is developed to identify risk hotspots in road networks. To effectively reveal the spatio-temporal relationship within the road network, we improved this method from the traditional kernel density estimation method and provided a method for calculating the optimal parameters.

Basics in kernel density estimation

Kernel density estimation is a non-parametric method for estimating probability density functions which are unknown. Assume that $\boldsymbol{x_1,x_2,...,x_n}$ are n sample points of independent and identically distributed random variables. The probability density of this distribution $f(\boldsymbol{x})$ is unknown, and kernel density estimation can be used to estimate it, defined as Eq.(1):

$$\widehat{f_h}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x_i}), \tag{1}$$

where, $\widehat{f_h}$ is the estimated density distribution function, K is the kernel density function, and h is the bandwidth. The conditions that the kernel density function K needs to meet are: non- negative, integral in the domain is 1, and it conforms to the probability density property. The bandwidth h is a non-negative number used to control the smoothness of the estimation in the domain.

In the task of finding traffic risk hotspots, the locations of different traffic events on the map (each event has a different risk value) are given, and the requirement is to estimate the probability density distribution of the risks on the whole map. This is a problem on a two-dimensional plane and cannot simply use the one-dimensional and density diffusion shown in Eq.(1). The general form of a kernel density estimator on a 2-D plane is given by Eq.(2):

$$\lambda(s) = \sum_{i=1}^{n} \frac{1}{\pi r^2} k\left(\frac{d_{is}}{r}\right), \tag{2}$$

where, $\lambda(s)$ is the density at location s on the plane, r is the search radius (bandwidth) of the KDE (only points within r are used to estimate $\lambda(s)$, k is the weight of a point i at distance d_{is} to location s. The weight k is modeled as a kernel function of the ratio between d_{is} and bandwidth r. The further away event i occurs from location s, the less impact event i has on location s. Common kernel functions include Gaussian function, Quartic function, minimum variance function, etc.

Spatial-temporal network kernel density estimation

Basic principles of network kernel density estimation

The main target of the KDE method is to find the probability density throughout the 2-D plane. However, in the road traffic system, the road links compose a network rather than a plane. The traffic risk should be modeled as spreading along the road links instead of evenly spreading on a 2-D plane. Therefore, network kernel density estimation (NKDE) is proposed as an extension of the standard KDE method to estimate the probability density distribution in a network space. The network kernel density estimator is given by Eq.(3):

$$\lambda(s) = \sum_{i=1}^{n} \frac{1}{r} k\left(\frac{d_{is}}{r}\right), \tag{3}$$

where, the variables have the same meaning as in Eq.(2), and use Gaussian kernel function, given by Eq.(4):

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{4}$$

Compared to standard 2-D KDE, NKDE assumes the risk spreads along the road links. Thus, the density is calculated over a linear unit rather than an area unit. Besides, the distance d_{is} is calculated as the shortest distance in the network space between event i and location s.

Establishment of road network data

To estimate the probability density distribution of events on the road network, we need to determine the specific location where each event occurs. Therefore, it is necessary to segment the route. Specifically, the road links are split into short segments (which are termed lixels) (17), which are no longer than a preset length (e.g., 10 meters). We assume risk density is uniformly distributed on each lixel. Thus, each lixel can be treated as a node. On the other hand, the connection relationships can be treated as the links that connect the lixels. In this way, we can build a lixel network from the original road network. Risk density estimation will be calculated based on the lixel network.

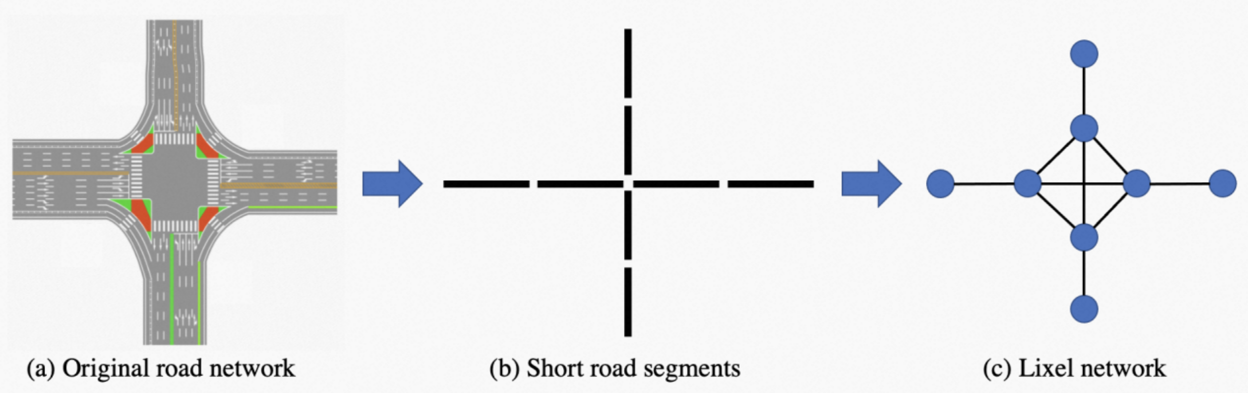

FIGURE 1: Establishment of lixel network

Spatial-temporal network kernel density estimation (ST-NKDE)

However, as is well known, in many locations, the risk is only high during peak hours in the morning and evening. If we can obtain the risk of each location at a specific time, we can provide some warnings to car drivers when driving through a high-risk location. In order to achieve this, we can choose to spread risk along the network and time axis, which is so called spatial-temporal network kernel density estimation (ST-NKDE). Meanwhile, the multi-variable KDE method is called generalized KDE (23). Spatial-temporal spreading completely differs from sampling at different time periods and then spreading separately in space. ST-NKDE can reveal the impact of events on the timeline, such as event i occurring at time 12:00, which can contribute to the risk of location s at time 12:10.

For the ST-NKDE, the density function is given by Eq.(5):

$$\lambda\left(s\right) = \Sigma_{i=1}^{n} \frac{1}{r_d r_t} k(\frac{d_{is}}{r_d}) k(\frac{t_{is}}{r_t}), \tag{5}$$

where, r_d and r_t represent bandwidths of different variables, and s is a 2-D vector with the position in the network and the time. Besides, different kernel functions can be chosen for different variables. As we have divided links into lixels, we divide the time axis into time slices to make the problem computationally tractable. And then, we can get the risk distribution for different time slices. A detailed analysis reveals that the impact of the risk of traffic events occurring at location i on location s not only decays in space but also in time.

To calculate the risk density, we take the following steps in Algorithm 1:

Algorithm 1: ST-NKDE calculation steps

```
Data: The road link data and the event data
Build up the lixel network;
Slice the time axis;
foreach event do
    Determine the spatial-temporal position;
    Add it's risk level to the position;
end
foreach position in spatial-temporal domain do
    if risk level of position $\neq 0$ then
        Find all it's neighbor position within the given bandwidths;
        foreach neighbor do
            Calculate risk density by $\frac{1}{r_d r_t} k(\frac{d_{is}}{r_d}) k(\frac{t_{is}}{r_t}) \times risk\ level\ of\ position$ ;
            Add to it's estimated risk density;
        end
    else
end
Result: Risk density on each spatial-temporal position
```

Derivation of optimum parameters

According to the operation steps in Algorithm 1, we can already determine which point on the spatial-temporal map is a risk hotspot. However, due to the fact that the diffusion bandwidths r_d, r_s are selected based on experience, we do not have sufficient reasons to suggest that these locations do pose high traffic risks. Therefore, we need to provide a method of selecting bandwidths r_d, r_t to make the results more convincing.

First, introduce a mathematical tool called Kullback-Leibler (KL) loss. KL loss is used to measure the amount of information lost when using a distribution to approximate the original distribution. The loss function in the case of discrete random variable is given by Eq.(6):

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x_i)(logp(x_i) - logq(x_i)), \tag{6}$$

where, p is the original distribution, and q is the distribution we predicted. The greater the loss, the less close our predicted distribution is to the original distribution. Our model defines $\lambda_0(s)$ as the potential original distribution, which is unknown to us, but we suppose it exists. Let $\lambda(s)$ represent the distribution we obtained from the above section. Since the routes have been cut into lixels and the time axis has been sliced, it can be assumed that each spatial-temporal position corresponds to a risk value, therefore it can be considered that the risk density function $\lambda(s)$ is discrete. Then, the loss in the prediction is given by Eq.(7):

$$D_{KL} = \Sigma_F \lambda_0(s)\left(log\lambda_0(s) - log\lambda(s)\right) = \Sigma_F \lambda_0(s)\left(log\frac{\lambda_0(s)}{\lambda(s)}\right). \tag{7}$$

However, we cannot adopt Eq.(7) directly because there might be a few points in field F with zero density, resulting in D_{KL} to positive infinity. Here, some limitations need to be imposed on this situation, such as the limit approaching positive infinity to approaching 1. We can use function $1 - \frac{1}{x}$ which approaches $log(x)$ to a certain extent. And then, the KL loss is given by Eq.(8):

$$D_K L = \Sigma_F\ \lambda_0\ (s)(1 - \frac{\lambda\ (s)}{\lambda_0\ (s)}) = \Sigma_F\ (\lambda_0\ (s) - \lambda(s)). \tag{8}$$

To minimize D_{KL}, given that $\lambda_0(s)$ is a certain distribution, it is equal to maximize the sum of $\lambda(s)$in field F, given by Eq.(9):

$$E(\lambda) = \Sigma_F \lambda(s). \tag{9}$$

To intuitively understand why bandwidths can be selected by maximizing the sum, the KDE method is to smoothly and accurately estimate unknown distributions. Initially, the bandwidths are zero, and diffusion does not occur. All values are concentrated at a few points, resulting in a small sum across the entire field. At this point, the estimated distribution is not smooth. To make it smooth, we increase the bandwidths and spread the value around. However, if bandwidths increase from zero, the sum first increases and then decreases. Furthermore, the sum in the field will decrease to zero if the bandwidths increase to infinity. Thus, suitable bandwidths are found to some extent while the sum is maximum. We can get the optimal bandwidths by solving the optimization problem in Eq.(10):

$$\begin{aligned} \max \quad & \Sigma_F \lambda(s) \\ s.t. \quad & r_d \in Q_d, \\ & r_t \in Q_t, \end{aligned} \tag{10}$$

where, $\lambda(s)$ is the risk density function; r_d is the distance bandwidth; r_t is the time bandwidth; Q_d and Q_t are reasonable intervals for distance and time bandwidth, respectively.

DATA DESCRIPTION

In this study, the traffic safety data were obtained from two distinct datasets. The first dataset was gathered from onboard detectors installed in probe vehicles. These detectors capture anonymized information about safety-related events such as emergency brakes, slippery roads, and potholes, as depicted in Table 1. The second dataset was collected from user-reported traffic events through Amap. This dataset contains both map data and traffic event data, enabling us to analyze and pinpoint the actual hotspot locations accurately, as depicted in Table 2. Using multiple data sources offers the advantage of enhancing the calibration and verification of the hotspot detection model through data complementarity. By combining these datasets, we can achieve a more comprehensive and robust understanding of traffic safety patterns and effectively identify areas prone to safety incidents.

TABLE 1: Data fields of the probe vehicle car data

Data label	Data field	Description
id of the data	Vin	Desensitized unique ID for each probe vehicle
	Uuid	Session ID for each data record
event data	Event type	Type of event (e.g., brake, acceleration)
	Event time	Sample time of the event
	Event value	Detailed value for each event (e.g., brake torque, acceleration value)
GPS data	Time	Sample time for the GPS point
	Longitude	Longitude of the point
	Latitude	Latitude of the point

The probe vehicle data was collected from April 1, 2022, to March 31, 2023, encompassing a span of one year. The user-reported traffic event data covers a more extensive period, from January 2021 to March 2023, totaling 27 months. Both datasets are concentrated in the Wangjing area in Beijing. The floating car data comprises event and GPS data from various vehicles operating in this region. The dataset includes event data with 6,714 vehicles and 26,595,248 event entries, GPS data with 7,149 vehicles and 3,455,001 entries, each sampled independently at different intervals. In order to aggregate data from events, we utilize the fuzzy matching technique. This technique involves matching data points based on certain criteria, such as matching by vehicle identification number (VIN) and a time window of ±5 seconds.

The user-reported data from the Amap navigation app encompasses traffic events reported within the same area. To augment the floating car data, we collected all event data within the same range as the floating car data. After removing any data that was not relevant to the main road network, we obtained a total of 17,811 traffic event records. All the data fields and their corresponding types are shown in Table 1 and 2.

TABLE 2: Data fields of user-reported traffic event data

Data label	Data field	Description
map data	Link id	Unique id for each link in the map
	Link shape and coordinate	WKT format data of link's shape and GPS coordinate, containing the position of the two intersections that the link connects
traffic event data	Event id	Unique id for each accident event
	Event type	Type of the traffic event (e.g., congestion, road closed, traffic accident, serious accident)
	Event start time	Start time of the event
	Event end time	End time of the event
	Longitude	Longitude of the event
	Latitude	Latitude of the event
	Event description	Description of the traffic event in plain language
	Link id	ID of the link where the event happens

RESULTS

Network kernel density estimation

In the user-reported traffic event data, for different traffic events, risk levels are artificially labeled based on the severity of the event (e.g., the risk level of congestion events is 1, the risk level of general traffic accidents is 3, and the risk level of serious traffic accidents is 5). In the probe vehicle data, risk levels are classified based on the magnitude of braking torque or acceleration. We used a 10 meters long lixel to divide the route and slice the time axis into 5-minute segments. After calculating the risk density on the map according to Algorithm 1, we normalized the data and set the spatial-temporal location density with the highest risk density throughout the day to 1.

Figure 2(a) show the risk value distributed on road links based on the proposed ST-NKDE method. In the contrast in Figure 2, it can be seen that using kde on a plane will mark many areas that are not along the route as risky, while using network kde can more finely mark the specific locations of risks on the road network.

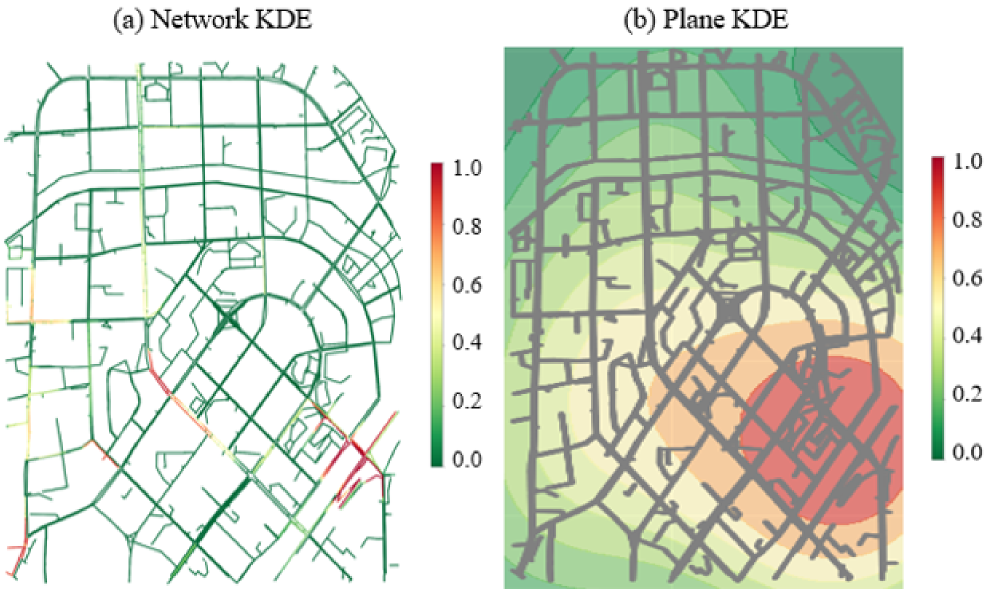

FIGURE 2: The risk map in 8:00-8:05 using the user-reported traffic event dataset, (a) is the result from network KDE; (b) is the result from plain KDE

Optimal parameter selection

To verify the effectiveness of the proposed kernel density estimation method, we determine the optimal bandwidths for spatial and temporal dimension for the traffic event dataset firstly. Figure 3(a) shows that when the distance bandwidth is 350m, the objective function is the largest. In Figure 3(b), when the time bandwidth is 10 minutes, the objective function is the largest, which can be regarded as the optimal parameters.

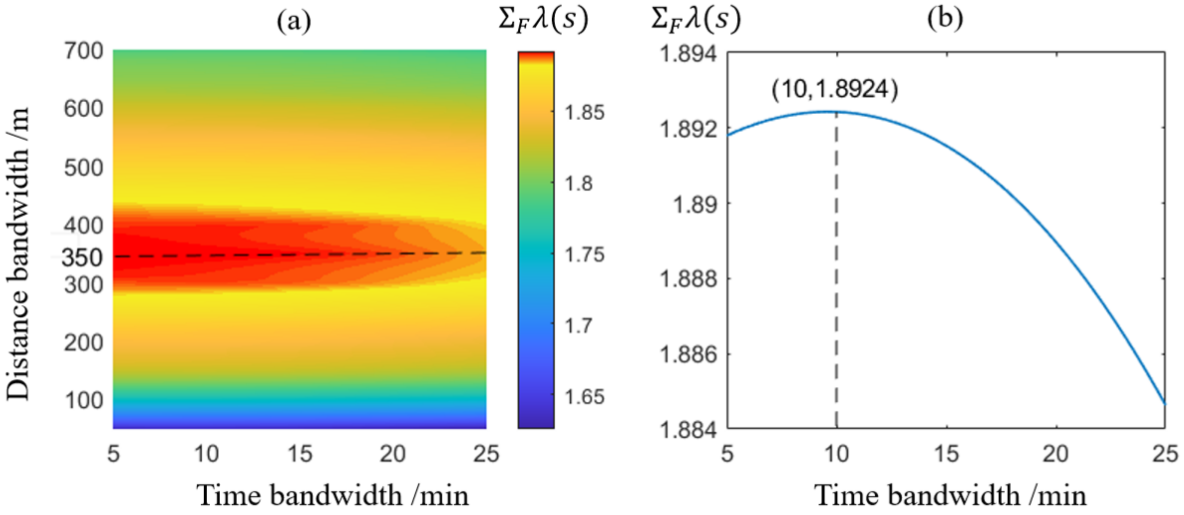

FIGURE 3: The optimal parameters from the user-reported traffic event dataset (a) is a plan view of the variation of the objective function with bandwidths ; (b) is a section view when distance bandwidth=350m

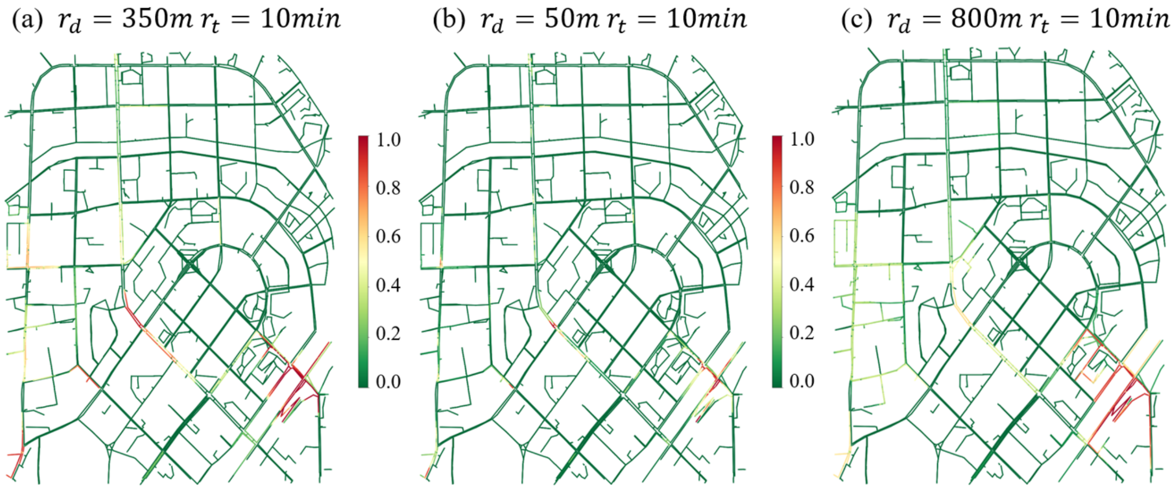
(a) $r_d = 350m\ r_t = 10min$
(b) $r_d = 50m\ r_t = 10min$
(c) $r_d = 800m\ r_t = 10min$
1.0
0.8
0.6
0.4
0.2
0.0

FIGURE 4: Risk map of 8:00-8:05 based on traffic event data (a) is the result using the optimal parameters $r_d = 350m, r_t = 10min$; (b) is the result using the parameters $r_d = 50m, r_t = 10min$; (c) is the result using the optimal parameters $r_d = 800m, r_t = 10min$

From the comparison of Figure 4 , it can be seen that when the optimal parameters are used in Figure 4(a), the changes in risk on the road network are smoother compared to using smaller distance bandwidth in Figure 4(b), and there are rarely "point like" high-risk areas. At the same time, many high-risk road sections marked with the optimal parameters are not reflected when the optimal parameters are not used in Figure 4(c). From the comparison, we can see the advantages of using the optimal parameters.

Dual dataset validation

Simple visualization cannot quantify the difference between the results obtained from the two datasets. Therefore, we introduce Jensen Shannon divergence (JSD) (24), which originates from Kullback Leibler divergence (KLD), to describe the differences between two distributions. There are two existing distributions P and Q. In our example, we assume that the P distribution is the risk distribution obtained from MB data, and the Q distribution is the risk distribution obtained from gaode data. Using KL divergence to measure the difference between two distributions is equal to one Cross entropy minus one information entropy, given by Eq.(11) and Eq.(12):

$$KL(P||Q) = \Sigma p(x) \log \frac{p(x)}{q(x)}, \tag{11}$$

$$KL(Q||P) = \Sigma q(x) \log \frac{q(x)}{p(x)}. \tag{12}$$

However, due to the asymmetry of KL divergence, $KL(P||Q) \neq KL(Q||P)$, we introduce symmetric JS divergence to describe the differences between two distributions defined as Eq.(13):

$$M = \frac{1}{2}(P+Q), \quad JSD(P||Q) = JSD(Q||P) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M). \tag{13}$$

Furthermore, we bring the KL divergence formula into the expansion as Eq.(14):

$$JSD(P||Q) = \frac{1}{2}\Sigma p(x)\log\frac{2p(x)}{p(x)+q(x)} + \frac{1}{2}\Sigma q(x)\log\frac{2q(x)}{p(x)+q(x)}, \tag{14}$$

where, $p(x)$ and $q(x)$ are normalized from the risk density $\lambda(s)$ obtained from two datasets. The value of JS divergence is between 0 and 1, and a smaller value indicates that the two distributions are approximately similar.

We further use the similar method to find the optimal parameters for the probe vehicle dataset. In Figure 5, it can be seen that the objective function achieves the peak when the distance bandwidth is 450m and the time bandwidth is 10min, which are similar to the optimal parameters for the traffic event dataset.

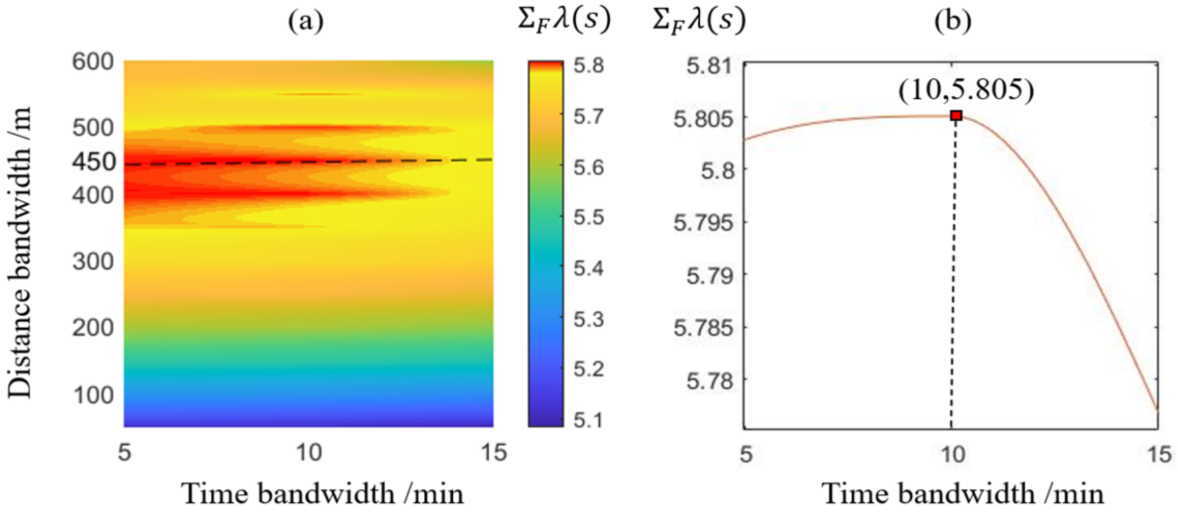

FIGURE 5: The optimal parameters from probe vehicle dataset (a) is a plan view of the variation of the objective function with bandwidths ; (b) is a section view when distance bandwidth=450m

We can choose the same time bandwidth for diffusion, and the selected bandwidth is relatively optimal on both datasets so that we can compare the differences in risk distribution between the two datasets within the same time period. In our two datasets, it is obvious that 10 minutes is the optimal time bandwidth for both of them. For these two datasets, we compare the differences in risk distribution between them during peak hours in the morning and evening. The risk maps of the two datasets in morning and evening peak hours are shown in Figure 6.

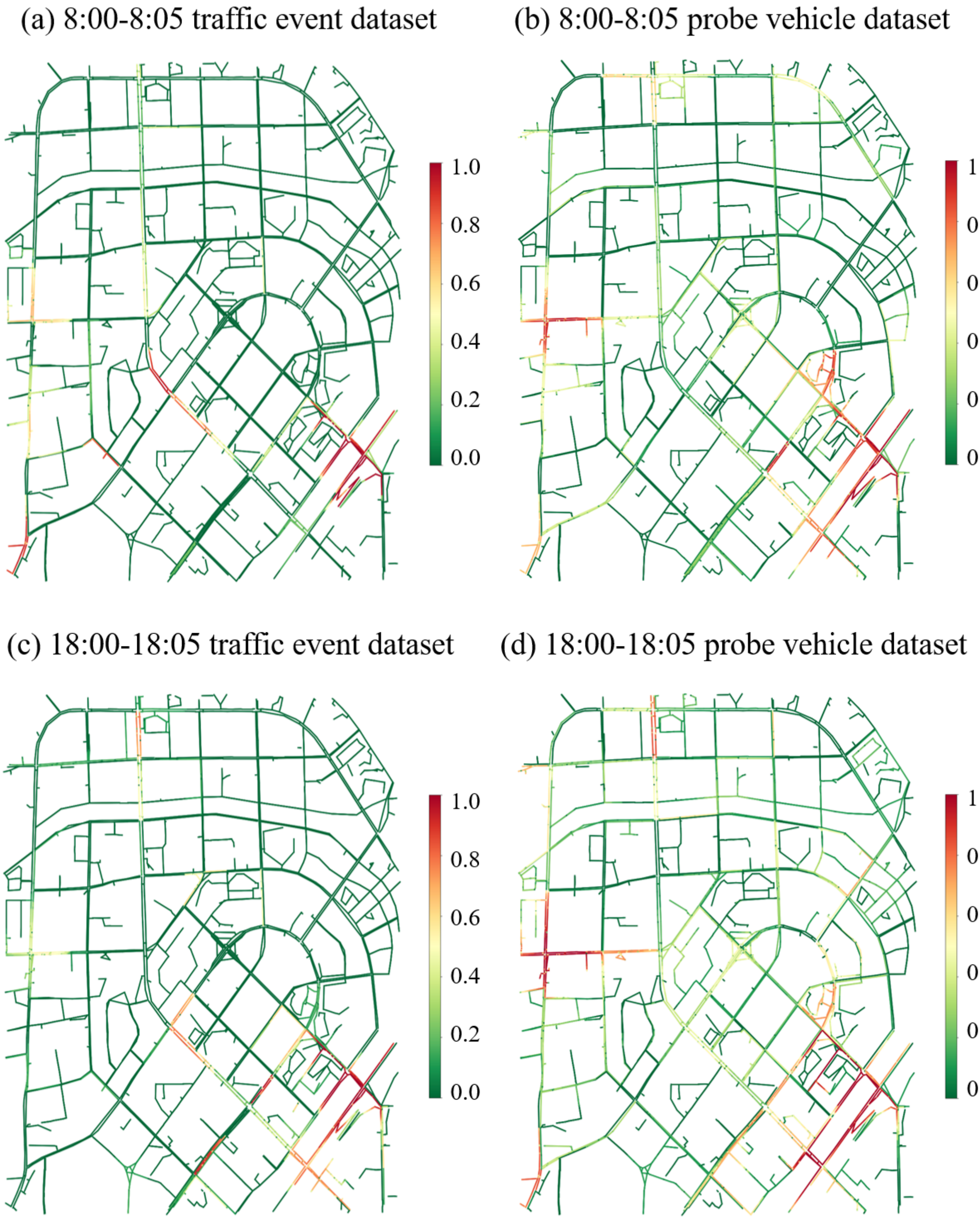
(a) 8:00-8:05 traffic event dataset
(b) 8:00-8:05 probe vehicle dataset
(c) 18:00-18:05 traffic event dataset
(d) 18:00-18:05 probe vehicle dataset
1.0
0.8
0.6
0.4
0.2
0.0

FIGURE 6: (a) is the risk map in the morning peak hours from the user-reported traffic event dataset using optimal parameters; (b) is the risk map in the morning peak hours from the probe vehicle dataset using optimal parameters; (c) is the risk map in the evening peak hours from the user-reported traffic event dataset using optimal parameters; (d) is the risk map in the evening peak hours from the probe vehicle dataset using optimal parameters

Subsequently, the JSD can be calculated for the normalized risk distribution based on the two datasets, respectively for the morning peak and evening peak:

$$JSD(P_m||Q_m) = 0.0939, \tag{15}$$

$$JSD(P_e||Q_e) = 0.1009, \tag{16}$$

where, P_m and P_e are the risk distributions obtained from MB dataset during the morning and evening peak period separately, Q_m and Q_e are the risk distributions obtained from the Amap dataset during the morning and evening peak period.

The results demonstrate that during the morning peak period, the divergence of JS is smaller, which indicates that the risk distribution obtained from the two datasets is more similar. By using the JSD method, we can obtain the similarity of risk maps from two datasets within each time slice. The smaller the JSD value, the more similar the two results are. In practical applications, a threshold can be set based on experience to determine whether the results obtained within each time slice are trustworthy. At the same time, it can also be used to explore the relationship between driver driving behavior and traffic accidents.

DISCUSSION AND CONCLUSION

In this study, we proposed a novel approach for traffic risk hotspot detection in urban areas by utilizing two complementary datasets: floating car data and user-reported traffic event data. The results demonstrate the effectiveness and reliability of our method in identifying high-risk areas on the road network. In this section, we discuss the implications of our findings and provide a comprehensive conclusion.

One contribution of our research lies in the effective combination of user-reported traffic incidents with probe vehicle data to produce a reliable traffic risk hotspot identification model. The probe vehicle data offers valuable insights into real-time vehicle trajectories and dynamics, enabling us to capture the underlying patterns of risky behaviors. On the other hand, the user-reported traffic event data provides complementary information and essential validation for our model, ensuring its accuracy and reliability. By combining these two datasets, we enhance the comprehensiveness and validity of our risk assessment.

Optimizing the accuracy of our model required careful consideration of the parameter selection procedure. We developed a more accurate and trustworthy method of risk assessment by using kernel density estimation to determine the ideal time and distance bandwidths. The selected parameters (distance bandwidth and time bandwidth) proved to be effective in capturing risk hotspots and providing a smooth risk distribution over the road network.

The comparison of risk maps during morning and evening peak hours further validated the consistency and effectiveness of our approach. The small Jensen Shannon divergence (JSD) values further indicated that the risk distributions obtained from the two datasets were relatively similar, emphasizing the robustness of our model.

Notably, the higher similarity during the morning peak period may be attributed to more predictable and regular traffic patterns during that time.

Our research may have important implications for urban traffic safety measures and city planning. By identifying traffic risk hotspots, city authorities can prioritize safety interventions and optimize traffic management strategies. The proactive identification of high-risk areas can lead to the implementation of targeted safety measures, such as traffic calming measures, improved signals, or enhanced enforcement in those specific locations.

However, some limitations should be acknowledged. Firstly, the datasets used in this study were localized in the Wangjing area of Beijing, which may limit the generalizability of our findings to other urban areas. Extending the study to diverse urban contexts can provide more comprehensive insights into traffic risk patterns. Additionally, the accuracy of user-reported traffic event data may be influenced by reporting bias and incomplete coverage, which can impact the reliability of the validation process. Future research could focus on refining the data collection process and exploring alternative validation methods.

In conclusion, our study presents an effective and reliable method for traffic risk hotspot detection in urban areas. By leveraging probe vehicle data and user-reported traffic events, we achieve a comprehensive and validated traffic risk assessment. The optimized parameter selection process enhances the accuracy of our model, while the JSD analysis confirms the consistency of risk distributions obtained from both datasets. This research contributes to the advancement of urban traffic safety measures and has significant implications for urban planning and traffic management strategies. Moving forward, further research and validation in diverse urban settings can expand the scope and applicability of our approach, leading to safer and more efficient urban transportation systems.

REFERENCES

1. Costescu, D., S. Raicu, M. Rosca, S. Burciu, and F. Rusca, Using intersection conflict index in urban traffic risk evaluation. *Procedia technology*, Vol. 22, 2016, pp. 319–326.
2. Zhao, X., Y. Ding, Y. Yao, Y. Zhang, C. Bi, and Y. Su, A multinomial logit model: Safety risk analysis of interchange area based on aggregate driving behavior data. *Journal of safety research*, Vol. 80, 2022, pp. 27–38.
3. Yang, D., K. Xie, K. Ozbay, H. Yang, and N. Budnick, Modeling of time-dependent safety performance using anonymized and aggregated smartphone-based dangerous driving event data. *Accident Analysis & Prevention*, Vol. 132, 2019, p. 105286.
4. Lu, T., Z. Dunyao, Y. Lixin, and Z. Pan, The traffic accident hotspot prediction: Based on the logistic regression method. In *2015 International Conference on Transportation Information and Safety (ICTIS)*, IEEE, 2015, pp. 107–110.
5. Shafabakhsh, G. A., A. Famili, and M. S. Bahadori, GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. *Journal of traffic and transportation engineering (English edition)*, Vol. 4, No. 3, 2017, pp. 290–299.
6. Al-Omari, A., N. Shatnawi, T. Khedaywi, and T. Miqdady, Prediction of traffic accidents hot spots using fuzzy logic and GIS. *Applied Geomatics*, Vol. 12, 2020, pp. 149–161.
7. Santos, D., J. Saias, P. Quaresma, and V. B. Nogueira, Machine learning approaches to traffic accident analysis and hotspot prediction. *Computers*, Vol. 10, No. 12, 2021, p. 157.
8. Yan, X., S. Richards, and X. Su, Using hierarchical tree-based regression model to predict train–vehicle crashes at passive highway-rail grade crossings. *Accident Analysis & Prevention*, Vol. 42, No. 1, 2010, pp. 64–74.
9. Qi, H., Y. Yao, X. Zhao, J. Guo, Y. Zhang, and C. Bi, Applying an interpretable machine learning framework to the traffic safety order analysis of expressway exits based on aggregate driving behavior data. *Physica A: Statistical Mechanics and its Applications*, Vol. 597, 2022, p. 127277.
10. Xia, D., Y. Zheng, Y. Bai, X. Yan, Y. Hu, Y. Li, and H. Li, A parallel grid-search-based SVM optimization algorithm on Spark for passenger hotspot prediction. *Multimedia Tools and Applications*, Vol. 81, No. 19, 2022, pp. 27523–27549.
11. Yao, S., J. Wang, L. Fang, and J. Wu, Identification of vehicle-pedestrian collision hotspots at the micro-level using network kernel density estimation and random forests: A case study in Shanghai, China. *Sustainability*, Vol. 10, No. 12, 2018, p. 4762.
12. Li, Y., M. Li, J. Yuan, J. Lu, and M. Abdel-Aty, Analysis and prediction of intersection traffic violations using automated enforcement system data. *Accident Analysis & Prevention*, Vol. 162, 2021, p. 106422.
13. Zhao, H., H. Cheng, T. Mao, and C. He, Research on traffic accident prediction model based on convolutional neural networks in VANET. In *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, 2019, pp. 79–84.

14. Anderson, T. K., Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, Vol. 41, No. 3, 2009, pp. 359–364.

15. Katicha, S. W. and G. W. Flintsch, A kernel density empirical Bayes (KDEB) approach to estimate accident risk. *Accident Analysis & Prevention*, Vol. 186, 2023, p. 107039.

16. Chen, X., L. Huang, D. Dai, M. Zhu, and K. Jin, Hotspots of road traffic crashes in a redeveloping area of Shanghai. *International journal of injury control and safety promotion*, Vol. 25, No. 3, 2018, pp. 293–302.

17. Xie, Z. and J. Yan, Kernel density estimation of traffic accidents in a network space. *Computers, environment and urban systems*, Vol. 32, No. 5, 2008, pp. 396–406.

18. Tang, L., Z. Kan, X. Zhang, F. Sun, X. Yang, and Q. Li, A network Kernel Density Estimation for linear features in space–time analysis of big trace data. *International Journal of Geographical Information Science*, Vol. 30, No. 9, 2016, pp. 1717–1737.

19. Hu, Y., F. Wang, C. Guin, and H. Zhu, A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied geography*, Vol. 99, 2018, pp. 89–97.

20. Li, Y., M. Abdel-Aty, J. Yuan, Z. Cheng, and J. Lu, Analyzing traffic violation behavior at urban intersections: A spatio-temporal kernel density estimation approach using automated enforcement system data. *Accident Analysis & Prevention*, Vol. 141, 2020, p. 105509.

21. Romano, B. and Z. Jiang, Visualizing traffic accident hotspots based on spatial-temporal network kernel density estimation. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2017, pp. 1–4.

22. Heidenreich, N.-B., A. Schindler, and S. Sperlich, Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Advances in Statistical Analysis*, Vol. 97, 2013, pp. 403–433.

23. Terrell, G. R. and D. W. Scott, Variable kernel density estimation. *The Annals of Statistics*, 1992, pp. 1236–1265.

24. Lin, J., Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, Vol. 37, No. 1, 1991, pp. 145–151.