Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 2151284, 12 pages
https://doi.org/10.1155/2019/2151284

Research Article

Identification of Accident Blackspots on Rural Roads Using Grid Clustering and Principal Component Clustering

Ling Shen[1,2,3] **Jian Lu**[1,2,3] **Man Long,**[1,2,3] **and Tingjun Chen**[4]

[1] *Jiangsu Key Laboratory of Urban ITS, Southeast University, Nanjing 21189, China*
[2] *Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Southeast University, Nanjing 211189, China*
[3] *School of Transportation, Southeast University, Nanjing 211189, China*
[4] *Taizhou Public Security Bureau, China*

Correspondence should be addressed to Jian Lu; lujian_1972@seu.edu.cn

Received 24 October 2018; Accepted 6 January 2019; Published 21 January 2019

Academic Editor: Alessandro Gasparetto

Identifying road accident blackspots is an effective strategy for reducing accidents. The application of this method in rural areas is different from highway and urban roads as the latter two have complete geographic information. This paper presents (1) a novel segmentation method using grid clustering and K-MEDOIDS to study the spatial patterns of road accidents in rural roads, (2) a clustering methodology using principal component analysis (PCA) and improved K-means to create recognition of road accident blackspots based on segmented results, and (3) using accidents causes in police report to analyze recognition results. The proposed methodology will be illustrated by accident data in Chinese rural area in 2017. A grid-based partition was carried on by using intersection as a basic spatial unit. Appended hazard scores were then added to the segments and using K-means clustering, a result of similar hotspots was completed. The accuracy of the results is verified by the analysis of the cause extracted by Fuzzy C-means algorithm (FCM).

1. Introduction

Traffic accidents are contingent events and are defined by a series of variables—the accident index, hidden danger index, and risk index—that explain them. When data is difficult to obtain in detail or changes greatly (such as in rapidly developing rural areas), latent variable models will be more suitable for safety evaluation. With the increase of car ownership and accidents in rural areas, developing countries like China are increasingly aware of the importance of rural road safety. By the end of the "Twelfth Five-Year Plan" (2015), the total mileage of rural roads in China reached more than 3.95 million kilometers. By the end of 2016, the number of household cars per 100 rural households was 17.4 (2016 Social Development Statistics Bulletin, 2017). At the same time, about two-thirds of all traffic accident deaths occurred on rural roads in 2016 (China Ministry of Transport, 2017). The Chinese government has put forward the slogan of "Four Good Rural Roads" and regards it as the main task of the Thirteenth Five-Year Plan.

One of the major difficulties in traffic safety evaluation is the heterogeneity of the data [1]. The threshold of selected variables is only used for accident black spots recognition, not considering the relationship between similar accidents, thus isolating the specific relationship between variables. In the establishment of the model for black spot recognition of accidents, the creation of multiple variables will have a certain degree of multicollinearity. Therefore, the model based on this contains vast amounts of redundant information [2]. Cluster analysis was used to identify black spots with the advantage of taking historical statistics and theoretical calculations into account [3]. It not only enables better clustering of similar segments, but also embodies the characteristics of different segments. It solves the problem of historical statistics.

Discretization of continuous attributes is an important preprocessing step in data mining. In the process of identifying the black spots of the accident, it is necessary to divide the intricate road network into continuous road segment for the

road black spots identification. In the identification process of accident black spots on highways, the road segments are divided according to fixed length, and data processing only selects the appropriate pile spacing. When identifying black spots of urban roads, GIS (Geo-Information system) [4] and Kernel density estimation [5] are well used because of the complete geographic information of urban roads and accident points. However, when identifying black spots in rural roads, especially for developing countries, the geographical location of rural roads is incomplete, and the description of accident locations is vague. This makes the segmentation process of rural roads different and more complicated than highways and urban roads. de Ona [6] uses Latent Class Cluster (LCC) as a preliminary tool for segmentation of accidents on rural highways in Granada. de Ona divides accident data into multiple hidden clusters according to the condition and severity, while geographic information has not been taken into account. Based on the basic idea of gridding-based cluster, this paper quantifies the analysis object into limited road segments. Being different from the CLIQUE algorithm [7] setting the grid of the established step size, this paper uses the intersection as the unit and clusters the rural road accident points according to the threshold of density. This is the preparatory work for the following principal component clustering. To the best of our knowledge, this is the first time that both approaches have been used together.

Among the methods of identifying black spots, data mining technologies are approved for the reliability and prospects. Many previous studies have focused on compressing and identifying key factors that have an impact on the severity of road accidents. Neural network [8], rough set [9], fuzzy logit, and decision tree learning [10, 11] have been applied for recognition. The establishment of the recognition model requires multiple variables, while the existing relationships between the variables are easily ignored, so the establishment of the dimensions and weights of the variables can be important [12]. In order to avoid the multicollinearity problem between multivariables, this paper uses PCA to quantify the information of each road segment and extracts the principal components. On this basis, the improved K-means clustering is used to identify the black spots of the accident. In order to verify the reliability of identification results, the causes of the accidents are chosen for analysis. Fuzzy c-means algorithm (FCM) is widely used to identify the causes of accident [13]. This paper further improves its accuracy and noise immunity.

The foci of this paper are as follows: (1) to present a methodology for the segment division of rural highway, (2) based on PCA's hazard scoring on the segment, to use K-means clustering to identify the black spots of the accident, and (3) to connect the cause of the accident to analyze and test the identified black spots of the accident.

2. Methodology

This chapter firstly introduces the accident segment division method based on gridding-based clustering. On this basis, the principal component cluster is introduced, which includes using the principal component to score the segment and using K-means to cluster the scoring results. Finally, fuzzy cluster is introduced to test the aforesaid results.

2.1. Gridding-Based Clustering. When analyzing conventional clustering problems, the Euclidean distance formula is generally used to measure the distance between two points. However, for the road traffic accidents, it is necessary to consider the spatial distribution difference between them and other general events; that is, traffic accidents are strictly restricted to road traffic networks. Being different from the one-dimensional linearity of the expressway, when analyzing the accident points in rural roads, the vehicles are strictly bound in the road network. If the Euclidean distance is used to describe the distance between the accidents, many actual errors will be generated in the road network which is easy to amplify the danger.

The gridding-based clustering algorithm refers to quantify an object space into a finite road segment to form a grid-like structure. This approach will increase processing speed and constrain the disorganized points in the space to the grid for analysis [14, 15], which brings the possibility of simplification to the rural road black spots featured by linear complexity and inaccurate road network. Classic grid clustering ideas, such as the CLIQUE algorithm, segment each dimension into nonoverlapping communities, so that the entire embedded space of the data object is segmented into units, and presupposed density thresholds can identify dense units. Gridding-based clustering requires two presupposed parameters: one is the step size of the grid and the other is the threshold of the density. This paper, when analyzing rural roads, replaces the units segmented by the established steps with intersections. The critical distance between dense intersections has no explicit provision. Referring to the *"General Principles for the Design of Chinese Civil Buildings"*, Anderson [5], Benedek [16], and Ulak [17] which define the window width of the accident intersection, scholars generally believe that it is reasonable to set up 100-200 meters in a city. For expressways, the distance is considered to be longer than 500 meters. The rural road network studied in this paper is relatively sparse compared with the urban road network. Hence, the critical distance between adjacent segments is set to 200 meters in this paper. The specific process is as follows:

(1) Scanning all grids. When the first dense grid is found, the grid begins to expand. Divided segment S_j includes the incident record of dense units u_i:

$$x_{s_j} = \sum_{u_i \in S_j} count\left(u_i\right) \tag{1}$$

$count(u_i)$ is the number of incidents contained in u_i. The extension principle is that if a gridding is adjacent to a gridding in a known dense region and is itself dense, then the gridding is added to the dense region until no such gridding is found. The average of the area coverage and the

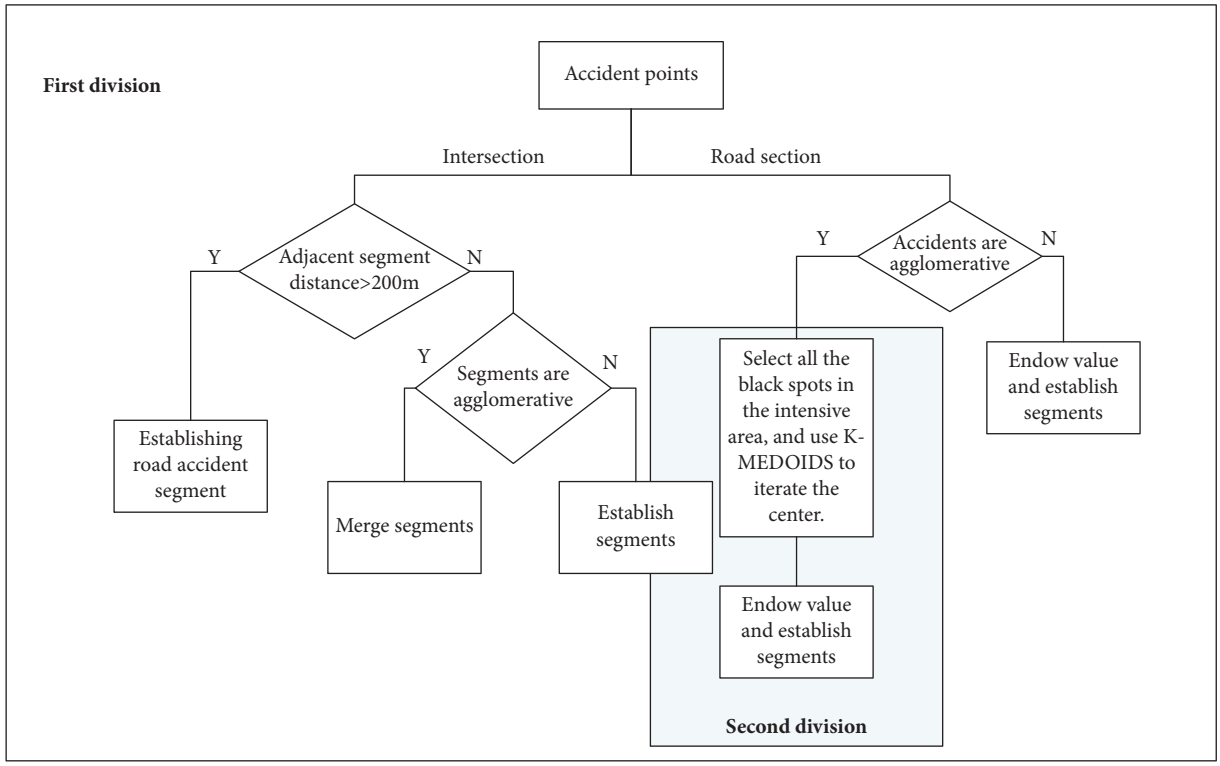

Figure 1: Road segment division flow chart.

difference between each unit and the average in the segment are calculated, and the sum of them is the objective function:

$$CL(i) = \log_2(u_I(i)) + \sum_{1<j<i} \log_2\left(\left|x_{s_j} - u_I(i)\right|\right) + \log_2(u_P(i)) + \sum_{i+1<j<n} \log_2\left(\left|x_{s_j} - u_p(i)\right|\right) \quad (2)$$

The aggregate R is selected, and the aggregate P is not selected.

(2) Continuing to scan the gridding and repeating the aforesaid process until all grids are scanned.

In order to solve the problems like the accident black spots of the same accident cause attribute or close distance divided into two different road units, the accident unit division of this paper will divide the second road segment division of the accident data accumulated on the road section. In order to avoid the interference of isolated points on rural roads [18], the K-MEDOIDS algorithm has been chosen for the second division.

The path distance D_i between the accident point i and the nearest intersection is the attribute variable of the sample. The specific analysis process is as follows:

① Determining the clustering center and completing the initial classification

K points are randomly selected from the accident points that need to be analyzed in the dense area as the original cluster center. Considering the general case, the relevant radius of the road network accident is 100 m. According to the dense local distance, a suitable K value is calculated, K=⌊L/100⌋+1, where L is the length of the accident-intensive area, and the unit is m.

The initial classification $G^0 = \{x_1, x_2 \ldots x_k\}$

② Calculating each accident point and assigning each accident data to its nearest cluster

The accident point parameters are defined as $Point_\alpha$, $\alpha\epsilon\{a, b, c \ldots\}$ and the path distance $|D_{\text{Point}_\alpha} - D_i|$, $i \in \{1, 2 \ldots k\}$ is compared, so the minimum value i is selected. The accident $Point_\alpha$ will be assigned i with corresponding cluster. By analogy, the accident points are assigned to the corresponding clusters.

③ For each point in each cluster, the similarity between each point and other points should be calculated separately (that is the distance from the cluster center). The point where the sum of the distances is the smallest is updated to a new center point of each cluster.

④ Entering iteration (2) (3) until the quality of the cluster meets the specified threshold (in this case, the corresponding cluster selected for each object no longer changes)

⑤ Outputting final classification $G^n = \{x_1, x_2 \ldots x_k\}$, n is the number of iterations

The final G^n is the result of the second division of the road unit. The road unit division flow chart is shown in Figure 1.

2.2. Principal Component Clustering. The method used in the previous section divides the accidents into appropriate segments. In this section, principal component analysis is used to score the hazards of the segment, and cluster analysis is performed on the basis of the score. Figure 2 shows the different spatial levels of road accidents.

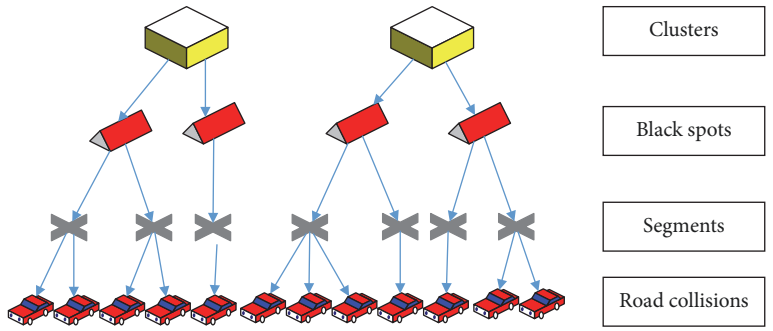

Figure 2: Spatial levels of road accidents identification.

2.2.1. Principal Component Analysis Quantitative Rating. The basic idea of principal component analysis is to process a more parsimonious description of the data provided, using fewer, more preferable association of variables to explain the variance of multivariates [19]. The principal component clustering model uses the mediation, trend, or internal relationship of accident data to indirectly analyze road traffic accidents. PCA analyzes rural roads in developing countries with two advantages: (1) allowing data loss and change within a certain range, avoiding excessive reliance on road accident data, which is suitable for the rapid construction of rural roads in developing countries; (2) using the eigenvalues to determine the main influence factors of the accident black spots and avoid the multivariate collinearity problem that occurs in traditional mathematical models.

The original dataset $X = [X_1, X_2 \dots X_p]$ can be integrated into a matrix X of n∗p. The data matrix contains n observation, each consisting of p variables. Let $R = [R_1, R_2 \dots R_p]$ be the correlation matrix of X. If the principal component analysis is effective, there should be K <n principal components, and the general formula of the ith principal component can be represented as $Z_i = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$.

The principal component is obtained to maximize the variability between individuals and is also constrained by $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1$ and COR $[Z_1, Z_2 \dots Z_i] = 0$. The eigenvalue of the matrix X of the sample variance-covariance is the variance of the principal component. The corresponding feature vector provides a coefficient that satisfies the constraint. The P × P sample variance-covariance matrix, the sum of the eigenvalues λ_p of this matrix is equal to the sum of the diagonal elements. Since the sum of the diagonal elements represents the total sample variance, and the sum of the feature values is equal to the trace of the matrix, the ratio of the total variance of the j principal component interpretation is $VAR_j = \lambda_j/(\lambda_1 + \lambda_2 + \cdots + \lambda_p)$, j = 1, 2, . . . , p.

Where VAR_j is the total variance of the j principal component interpretation and λ_j is the eigenvalue corresponding to the j principal component. In order to avoid excessive influence on the unit of measurement, principal component analysis needs to be conducted on a standardized matrix of variance-covariance. $Z_{ij} = (X_{ij} - \overline{X_j})/\sigma_j$; $i = 1, 2, \dots, n$; $j = 1, 2, \dots, p$.

When an observation is different from most of the data or is sufficiently unlikely under the assumed probability model for the data, the observation is considered to be an outlier. The common variables of the analysis of the traffic accident are limited, which makes the extracted number of principal components less. It makes outlier have a big influence on the identification (Zheng, 2013). The sum of squares of the standardized principal component scores is given by

$$\sum_{i=1}^{P} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_p^2}{\lambda_p} \tag{3}$$

corresponding to a chi-square distribution with q degrees of freedom. Considering the minor components, $\sum_{i=p-r+1}^{p}(y_i^2/\lambda_i)$ should be used to detect observation. For a given significance level α, an observation x is an outlier if

$$\sum_{i=p-r+1}^{p} \left(\frac{y_i^2}{\lambda_i}\right) > x_r^2(\alpha) \tag{4}$$

where $x_r^2(\alpha)$ is the value for chi-square distribution with r degrees of freedom testing at a given signification level α.

2.2.2. Improved K-Means Algorithm. K-Means and K-MEDOIDS mentioned above belong to partitional clustering, dividing data into K clusters, and completing segmentation according to $\arg_s \min \sum_{k=1}^{K} \sum_{x \in S_k} \|x - \mu_k\|^2$, where μ_k is the accident center of cluster C_k. The aim of K-means clustering is to minimize the distance among clusters, i.e., to minimize the objective function $J = \sum_{n=1}^{N} \sum_{k=1}^{K} \delta_{nk} \|x_n - \mu_k\|^2$, where δ_{nk} is the indicator function and if n belongs to cluster k, the value of δ_{nk} equals to 1, otherwise δ_{nk} equals to 0.

In the process of accident data analysis, various data leads to outliers, thus influencing the selection accuracy of cluster centers [20]. Add indicator function to objective function, which is defined as follows:

$$\varphi_k^l = \frac{\sum_{n=1}^{N} \delta_{nk}^l}{\sum_{n=1}^{N} \delta_{nk}} \tag{5}$$

where φ_k^l approaches to 1, indicating that cluster C_k only contains category l, otherwise nearly not. Additionally, set the φ_k^l in objective function as follows:

$$\mathrm{J}\left(\mu_k^l, \delta_{nk}^l\right) = \sum_{n=1}^{N}\left[a\sum_{k=1}^{K}\sum_{l=1}^{L}\delta_{nk}^l\left\|x_n - \mu_k^l\right\|^2\varphi_k^l + (1-a)\sum_{k=1}\delta_{nk}\left\|x_n - \mu_k\right\|^2\right] \tag{6}$$

where initial value of δ_{nk}^l can be calculated by Laplacian smoothing; μ_k^l can be calculated by partial derivation of both sides of objective function; a is the ratio of supervised clustering to unsupervised clustering. Algorithm implementation process is the same as K-MEDOIDS in Section 2.1, and the difference lies in the calculation of objective function in iteration, which is calculated by μ_k^l and δ_{nk}^l. When the value of $\mathrm{J}(\mu_k^l, \delta_{nk}^l)$ does not change, iteration is completed.

2.3. Fuzzy Cluster Identification. In order to verify the reliability of black spots, the following causes of the accident in the data should be analyzed. Fuzzy cluster analysis refers to the mathematical method of describing and classifying things according to certain requirements by fuzzy mathematical language, which is widely used in the identification of accident causes. The difference between fuzzy clustering and other analytical methods is that when it involves fuzzy boundaries between things, it classifies things according to specific requirements. In the analysis of the cause of the black spot of the accident, the influence of the cause is difficult to accurately quantify with the data model. Since there is no absolute difference and clear boundary between objective things, in the actual situation, the existence of ambiguity will be more in line with the analysis of the cause of black spots.

Fuzzy C-means algorithm (FCM) is the most reliable algorithm in existing fuzzy clustering algorithms. FCM obtains the membership degree of each sample point to all class centers by optimizing the objective function, thereby assigns the sample points to the nearest class, and finally completes the classification of the sample data. However, considering the shortcomings of traditional FCM in terms of convergence speed, accuracy, and anti-interference, this paper selects the fuzzy C-means clustering method based on weight entropy improvement to cluster the cause of black spots. The effects of various types on the clustering results are different, and the bias of the objects belonging to each category will be different. Usually, the weights are used to extend the traditional cost function, which indicates that different attributes have different effects on the clustering results. Frigui and Nasraoui [21] proposes a new cost function F_1 for fuzzy clustering. Among them, when the cost function is the smallest, the cluster is optimal:

$$F_1(T, W, C) = \sum_{l=1}^{k}\sum_{j=1}^{n}\sum_{i=1}^{m}\tau_{lj}\omega_{li}^{\beta}\left(x_{ji} - c_{li}\right)^2 \tag{7}$$

Restricted conditions

$$\sum_{l=1}^{k}\tau_{lj} = 1, \quad 1 \le j \le n, \ \tau_{lj} \in \{0,1\}$$
$$\sum_{i=1}^{m}\omega_{li} = 1, \quad 0 \le \omega_{li} \le 1, \ 1 \le l \le k \tag{8}$$

where ω_{li}^{β} is the weight of the i attribute of the l cluster. β is a parameter greater than 1. T is a matrix of k ∗ n for τ_{lj}, W is a matrix of k ∗ m for ω_{li}, and C is a matrix of k ∗ m for c_{li}. In the formula, T, W, C are all unknown matrices to be sought, but the other two sets of values can be fixed to obtain the third set of values. The formula is as follows:

$$\tau_{lj} \in \{0,1\} \tag{9}$$

$$c_{li} = \frac{\sum_{j=1}^{n}\tau_{lj}x_{ji}}{\sum_{j=1}^{n}\tau_{lj}} \tag{10}$$

$$\omega_{li} = \frac{1}{\sum_{t=1}^{m}\left[\left(\sum_{j=1}^{n}\tau_{lj}\left(c_{li} - x_{ji}\right)^2\right)/\left(\sum_{j=1}^{n}\tau_{lj}\left(c_{lt} - x_{jt}\right)^2\right)\right]^2} \tag{11}$$

The specific process can set the initial matrices W and C firstly and then obtain T according to the principle of minimum distance. Based on the obtained T, the matrices W and C are updated. The process is then iterated until the presupposed condition is met or a predetermined number of times is reached.

Considering the above algorithms only meet the need to make the distance between the same clusters as small as possible in the clustering, and the distance between different clusters is as large as possible and is not constrained. This paper adds weight entropy to the original to correct the cost function. Entropy of weights:

$$\gamma\sum_{i=1}^{m}\omega_{li}\log\omega_{li} \tag{12}$$

where the parameter $\gamma > 0$, then correct the cost function by the sum of the squares of the distance between the global center and the class center:

$$F_2(T, W, C) = \sum_{l=1}^{k}\left[\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}\tau_{lj}\omega_{li}^{\beta}\left(x_{ji} - c_{li}\right)^2}{\sum_{i=1}^{m}\left(c_{li} - \overline{x_i}\right)^2} + \gamma\sum_{i=1}^{m}\omega_{li}\log\omega_{li}\right] \tag{13}$$

In the following paper, according to the three-dimensional data set of the original data, the Xie-Beni indicator is selected to evaluate the clustering effect. On this basis, Gaussian noise contrast is added to the original 3D dataset to evaluate the improved FCM accuracy and noise immunity.

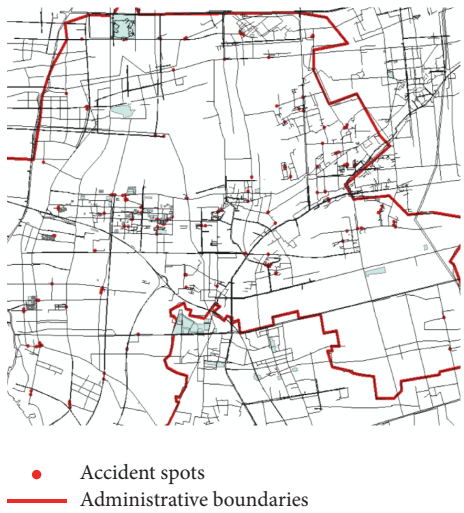

Figure 3: Statistical chart of rural road accidents in 2017.

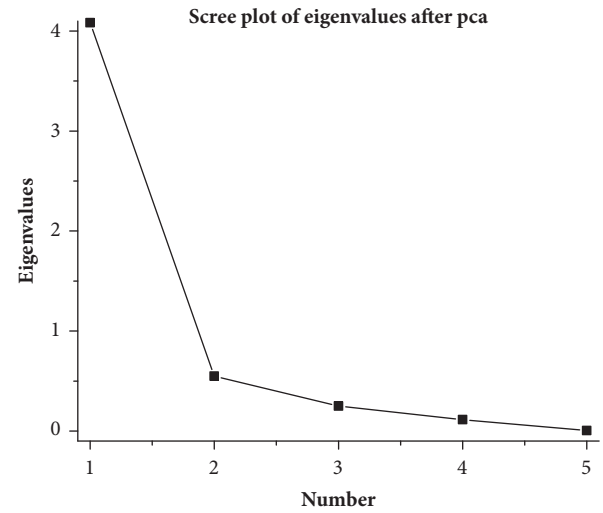

Figure 4: Scree Plot.

3. Data

The accident data was obtained from the Traffic Police Accident Section of the county-level city of Jiangsu Province (eastern coast of China). The accident data is of the rural road traffic police in the country recorded in 2017. It is based on the standard police reports used in China, with recorded variables: the number of deaths, the number of casualties, the economic loss, the location of the accident, the cause of the accident, the date of the accident, the number of the accident handling file, the handling of the accident police, name accident participants, and ID card number of accident participants. The statistical variables included in the paper are the number of deaths, the number of casualties, the economic loss, the location of the accident, and the cause of the accident. Due to China's political and confidentiality factors, the local government only provided data on accidents in 2017. In the next phase of the study, if support from the government is obtained, the accident data after 2017 will then be cross-checked to verify the accuracy of identifying black spots.

Traffic police record data and road network data (source: http://www.openstreetmap.org/) are imported into Arcgis software. According to the data of accidents, a total of 214 accident data records were obtained, including 163 accidents at intersections and 51 accidents at road sections. Accidents at intersections account for 76.2% of road accidents in rural areas. Hence, it will have a good prospect to conduct the intersection crash analysis using the accident identification in Section 4.2 in further studies. Figure 3 shows the distribution of accident locations.

4. Results and Discussion

4.1. Segmentation and Preprocessing. Based on the grid clustering analysis method (first division), units are directly established for accidents occurring at intersections, and adjacent units are merged. Based on the grid-based and K-MEDOIDS, a total of 56 segments were obtained from 214 sets of accident data. According to the traffic police accident record form, the variables (number of accidents, number of deaths, number of casualties, direct property losses, and accidents without any injuries involved) were extracted to analyze the road accidents. Considering the difference between the magnitude of the property loss and the number of deaths and injuries, the standardization process is required. The maximum and minimum normalization methods are used to linearly convert the initial data. The maximum and minimum normalization method is through the mapping formula: $v' = ((v - min_A)/(max_A - min_A))(new\_max_A - new\_min_A) + new\_min_A$, map the variable v to a new range $[new\_max_A, new\_min_A]$ and become the new value v', eliminating the magnitude difference. Data is compiled in Table 1.

4.2. Identification of Black Spots. Considering that principal component analysis only applies to related variables, the correlation of these five indicators is firstly analyzed. It has been verified that the correlation coefficient between the parameters is greater than 0.5 (the minimum DEATH and PROPERTY coefficients are 0.5443); that is, the indicators are significantly correlated with each other, which is consistent with the requirements for principal component analysis. Through principal component analysis, the Scree Plot of the principal component eigenvalues is obtained. Figure 4 shows the Scree Plot of eigenvalues.

As shown in the figure, the feature value of "Comp1" is much higher than other feature values. The change in the coordinates from "Comp1" to "Comp2" is extremely obvious. After "principal component 2", the eigenvalue changes tend to be gentle. According to the Cattell steep-order test rule (also known as the gravel map test, which analyzes the steep transition between the steep slope and the slow slope of the factor eigenvalues), the cumulative contribution rate of the variance needs to be greater than 0.8. The variance contribution rate of "principal component 1" was found to

Table 1: The results of the accident unit division and its standardized parameters.

Unit	No. of accidents	No. of deaths	No. of injuries	Direct property losses	No. of deaths and injuries
1	0.08696	0.33333	0.09524	0.05312	0
2	0.04348	0	0	0.01848	0.16667
3	0.13043	0	0.19048	0.09469	0
4	0.08696	0	0.09524	0.03002	0.16667
5	0	0	0	0.00346	0
6	0.56522	0.33333	0.61905	0.30254	0.50000
⋮	⋮	⋮	⋮	⋮	⋮
53	0.08696	0	0.09524	0.09353	0
54	0	0	0	0.00808	0
55	0	0	0	0.00462	0
56	0	0	0	0.02194	0

Table 2: Variable value of "Comp 1".

Variable	Accidents	Death	Injure	Property	No damage
Comp1	0.4821	0.4379	0.4721	0.3633	0.4710

be 0.8166, and the condition was satisfied. Variable value of "Comp 1" can be seen in Table 2.

This paper selects "Comp1" as the principal component variable. Variable value was used as coefficient value to modify the equation. The specific explanation of Comp1 is as follows:

$$\begin{aligned} Comp1 = {} & 0.4821 \times accidents + 0.4379 \times death \\ & + 0.4712 \times injure + 0.3633 \times property \\ & + 0.4710 \times nodamage \end{aligned} \quad (14)$$

From the composition of "principal component 1", this value has a significant positive correlation with all absolute index values. Relatively speaking, the parameter property has a lower impact (that is because the property loss coefficient is smaller), while the parameter coefficients of the other four related personnel casualties and accidents are relatively high. On the other hand, there are some problems such as collinearity and dimension weight among the accident-related accident parameters. The principal component analysis method proposes a solution to this problem by weighing them. In this paper, the variance contribution rate of the selected components is used as a weight to summarize F to calculate the risk level of the accident. Results are shown in Table 3.

However, during the process of grading road black points by cluster analysis, it is found that the value of k will have a great influence on the specific selection of critical indicators. The defect of K-means clustering is that the K value needs to be artificially defined before the cluster starts. In order to select the most suitable k value, this paper is based on sum of squared errors (SSE) and Silhouette Coefficient to improve the process of K-means in the process of black spot identification.

SSE is an indicator used to measure clustering effects. The smaller the value of SSE, the closer the distance to the centroid and the better the clustering effect. The formula of SSE is defined as follows: SSE $= \sum_{i=1}^{k} \sum_{x \in C_i} dist(c_i, x)^2$, where C_i is the aggregate of the ith data and C_i is the centroid of the cluster C_i k indicating that the data set can be divided into a set of K clusters, and dist is the Euclidean distance between the two objects. It is not difficult to find out from the algorithm that SSE is actually a strict coordinate descent process. Therefore, the value of SSE is gradually decreasing as the value of k increases. The current mainstream view is that the elbow in the statistical chart is the best. However, relying solely on the "elbow" evaluation may have a large error. In this paper, SK is used to replace SSE with min(SSE $*$ k). Since the optimal value of SK is a local minimum, it is easier to obtain than the optimal value of SSE.

The contour coefficients take the cohesion and separation of the clustering results into account when evaluating the effects of clustering. The contour factor can range from -1 to 1. Within the range, the contour coefficients are positively correlated with the clustering effect. For the element x_i, the contour factor $s_i = (b_i - a_i)/\max(b_i, a_i)$, where a_i represents the average of the distance between x_i and all other elements in the same cluster and is used to quantify the degree of cohesion within the cluster. b_i denotes the minimum value of the average distance between all clusters, x_i and all points in cluster b, for quantifying the degree of separation between clusters. Based on the above formula, it is not difficult to find that if s_i is less than 0, it means that the average distance between x_i and its elements is smaller than the nearest other cluster, and no good clustering effect is obtained. When a_i tends to 0 or b_i is large enough, the value of s_i will approach 1, indicating that a good clustering effect is achieved.

In summary, SSE and contour coefficients evaluate the appropriateness of k-values for K-means clustering from different aspects.

The comprehensive risk level obtained in the aforesaid article is an indicator, and the accident unit is clustered. To preselect the appropriate K value, a comprehensive evaluation

Table 3: Results of principal component analysis of accident unit.

Unit	1	2	3	4	5	6	...	53	54	55	56
Comp1 accident risk level	0.25206	0.10617	0.18704	0.17621	0.00126	1.05557	...	0.12078	0.00294	0.00168	0.00797
Comprehensive risk level	0.20583	0.0867	0.15273	0.14389	0.00103	0.86198	...	0.09863	0.0024	0.00137	0.00651

The higher the value of the comprehensive evaluation, the more dangerous the road segment, while the lower the value of the comprehensive evaluation, the safer the road segment.

Table 4: Accident black spot identification results.

Risk levels	Road units
Black spots in the accident/ road safety hazards	23, 9, 6, 25, 17, 7, 20
Potential accident black spots/ road safety conditions in general	12, 16, 14, 15, 52, 22, 40, 19, 35, 1
Nonaccident black spots/road safety conditions	Other road units

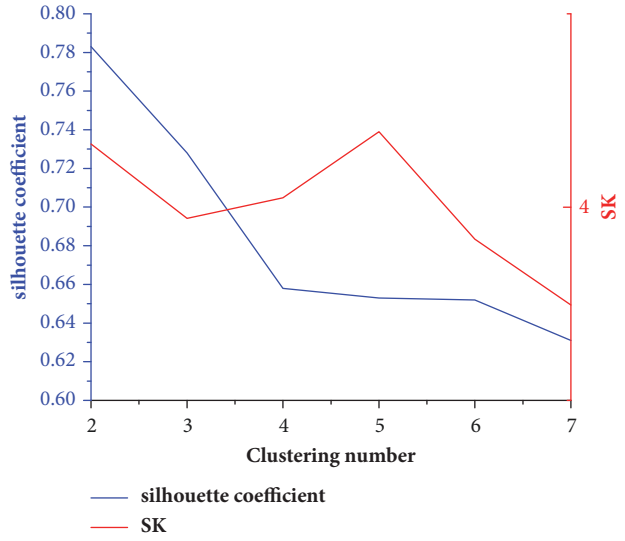

Figure 5: Silhouette coefficient of K and SK broken line figure.

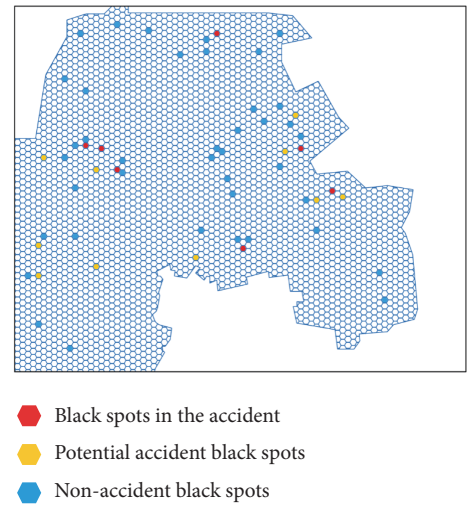

Figure 6: Hierarchical sketch of black spots.

of the analysis results of SK and Silhouette Coefficient shall be conducted which can been seen in Figure 5.

In this experiment, the value of the silhouette coefficient is decreasing. Since the contour coefficient and the clustering quality are positively correlated, the value of K should be prioritized. In addition, the smaller the SK value, the better the quality of the cluster. When K=3 is combined, the contour coefficient is high and the SK fold line has obvious trough shape at this time. Therefore, this K-means cluster analysis has the most suitable K value.

The rural road accident data of this analysis is divided into three clusters. According to the principle that "the higher the score, the higher the possibility and risk of a traffic accident; that is, the higher the probability of becoming an accident black spot", the three clusters correspond to three risk levels (see Table 4).

To enhance the output of identifying the blackspots, Figure 6 abstracts maps into cellular maps, and the accident point is mapped to the corresponding cell.

4.3. Analysis Based on Accident Cause. In the previous article, the relevant parameters were constructed with traffic police record data to identify black spots in rural accidents. Next, the cause of the accident is analyzed to test the identified black spots. In the investigation and record of accident cases, the duty of the Chinese traffic police is the identification of the accident, that is, whether or not the cause of the accident is a human factor. The cause of the accident recorded will not take into account the problem of the marginal blurring of the road factor and the human factor, so that the inherent cause of the accident cannot be exposed.

The traffic police define the accident such that the responsible person cannot be determined as "Other reasons (road users without obvious fault)". Before the cluster analysis of the cause of the accident, this article removed the accident

Table 5

Cause sequence		Clustering center		
		FREQ	CDR	EQUIV
Main factors	Turning vehicles do not allow straight-through vehicles and pedestrians Nonmotorized vehicles violate traffic rules	32.21	0.06	12.39
Minor factors	Deliberately impeding the safe driving behavior of others	7.5	0.21	5.74
Inducing factors	The sidewalk does not allow parking	8.28	0.13	3.82
Hidden factors	Illegal reversing Violating traffic lights Violating the marking line Unsafe travel distance	2.04	0.46	1.21
Negligible factors	Illegal reversing Reverse driving	3.54	0.003	1.25

Table 6

Cause sequence	Clustering center		
	FREQ	CDR	EQUIV
"Other reasons (road users without obvious fault)	114	0.04	91.25

data of "other reasons", because the proportion of "other reasons" is too large (more than 50%), which will have a negative impact on the identification of the cause of the accident and the accuracy of the black spot of the accident. The analysis concerning this cause will be conducted after clustering.

This paper selects the number of accidents (cases), fatality rate, and traffic accident equivalent as variables. **Freq** is the variable which represents the number of accidents. The calculation formula of the traffic accident equivalent of this analysis is **Equiv** $= \alpha \times$ Death $+ \gamma \times$ Injure $+ \mu \times$ Proportion, where α, γ and μ are accident weights. Referring to Cao [22] and the accidental black spot verification experience in Zhejiang Province, the parameters are selected in the traffic death equivalent study, the equivalent ratio of death and injury is 3:1, the value of α is 1, and the value of β is 0.33 and μ is 1/5000.

The mortality rate is calculated as **Crude Death Rate (CDR)** = Death/Accidents. The result of clustering is shown in Figure 7.

The three-dimensional eigenvectors $F_i = (Freq_i, CDR_i, Equiv_i)$ of the causal features are constructed, and the fuzzy clustering is performed for each reason. According to Figure 6, sort out the factors. The factors, "drunk driving", "fatigue driving", and "unlicensed driving", are not recorded in Table 5, because they have the subjectivity of drivers and have little to do with road environmental factors. It is worth noting that the frequency of these three types of accidents is low, but CDR of the cluster which "fatigue driving" and "drunk driving" belong to is as high as 0.463, so it is recommended to carry out strict law enforcement supervision. Table 5 shows the result of FCM and Table 6 shows the eigenvectors of "other reasons (road users without obvious fault)".

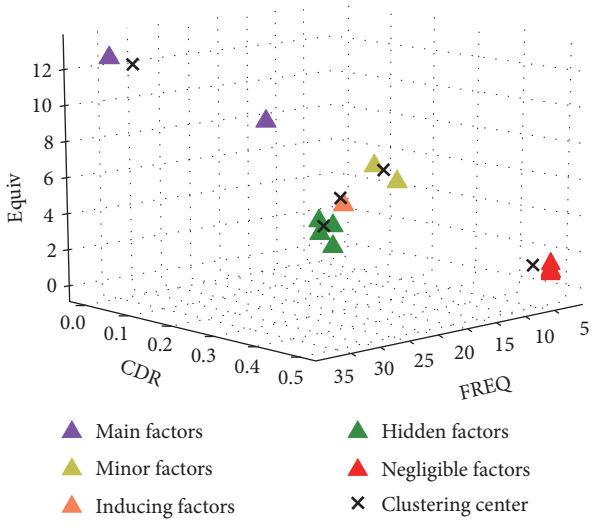

Figure 7: Diagrammatic sketch of cause partition.

This paper selects the Xie-Beni indicator to evaluate fuzzy clustering. The Xie-Beni indicator refers to for the data set X, in the given number of clusters K and the membership matrix T, the parameter definition values are

$$S(T, k) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} \tau_{ij}^{2} \left\| x_j - c_i \right\|^2}{n \min_{1 \le p q \le k} \left\| c_p - c_q \right\|^2} \quad (15)$$

Table 7: Cluster validity test of fuzzy clustering.

	Xie-Beni	Xie-Beni after adding noise
FCM	2.128	4.520
FCM based on weighted entropy and squared sum improvement	0.916	1.360

Table 8

		Accident black spots	Nonaccident black spots	Total proportion
Main reason	***Other reasons (road users have no obvious fault) R1***	***0.615***	***0.455***	***0.533***
	Turning vehicles do not allow straight-through vehicles and pedestrians R2	0.173	0.145	0.159
	Nonmotorized vehicles violate traffic rules R3	0.096	0.109	0.103
Total		***0.885***	***0.709***	***0.794***

The meaning of each parameter in the formula is the same as the definition of the aforesaid fuzzy clustering. The numerator represents the degree of density of the cluster, that is, cluster tension. The denominator expresses the minimum distance of the cluster center, that is, the degree of separation. The minimum point of the value is the optimal solution.

To test the improved FCM's noise immunity, a noise comparison was added to the original 3D data set. Yan [23] believes that in the antinoise test of FCM, the analysis of key variables can meet the accuracy requirements. In this paper, vector EQUIV represents the accident risk more comprehensively. Referring to Figure 6, EQUIV mostly floats around 2, and fluctuation magnitude of 0.1 is reasonable. Hence, add Gaussian noise of EQUIV with a mean of 2 and a variance of 0.1 to the three-dimensional point cloud. Data volume is equal to original data. Calculate the Xie-Beni indicator for FCM and improved FCM, respectively (see Table 7).

According to Table 7, the improved FCM has better clustering effectiveness with adding and without adding noise. At the same time, after adding Gaussian noise data points, the improved FCM clustering effectiveness changes are less and more stable. This test can be concluded that the improved FCM based on weighted entropy and squared sum has better accuracy and noise immunity.

The clustering effectiveness of this fuzzy clustering: Xie-Beni value is 0.916, and the segmentation result is ideal. The "turning vehicles do not allow straight-through vehicles and pedestrians" "nonmotor vehicles violate traffic rules" as the main factor cluster for analysis. The indicator vector representing the main cause of the cluster center represents the number and severity of accidents much higher than the centers of other clusters. It indicates that the two genes belonging to the main cause cluster have largely led to the formation of black spots in the accidents. On the other hand, the main cause of the cluster center represents a lower CDR vector for mortality, only 0.058, which shows that although these three reasons lead to a large number of accidents and losses, the fatality rate is relatively low among many reasons. The main factors of statistics are the proportion of the total accidents shown in Table 8.

The main factors which accounted for 88.5% of the accidents were caused in accident black spots, which was higher than the main factors that accounted for 79.4% of the total accidents. However, R1, R2, and R3 are all the main clusters, the ratio of R2 and R3 at the accident black spot is not too different (10% or less).

Notably, the other reasons are represented by R1 (the road user has no obvious fault); the performance in the identified black spots of the accident and the performance in the nonaccident black spots are significantly different. Considering the large number of R1 accidents, the data will avoid the contingency and error, so the problem presented by R1 is more reliable. Conduct Analysis of R1 with practicalities. In the record of accidents, traffic policemen prefer to perform the duty of defining responsibility. Therefore, in the absence of obvious driver responsibility, the traffic police will not record the cause. This means that R1 usually occurs where road environment is the dominant factor in accidents. In the case that the road user has no obvious fault and does not consider the failure of the vehicle itself, R1 can be expressed as a defect in the road environment. Among the 7 road black spots, road environmental problems caused 64 accidents, accounting for 61.5% of the total 104. Among the remaining 49 nonblack point road segments, there were 110 accidents, of which 50 accidents accounted for 45.6% of the total for other reasons. The comparison found that the accident caused by the black spot of the accident was caused by the fact that the proportion of defects in the road environment was significantly higher than that of the nonaccident black spot. It shows that the road black spots identified above, compared with other road units, have obvious problems in the road environment; that is, roads with poor road environmental conditions are identified.

The comprehensive clustering validity is ideal, the main cause of the accident black spots identified is a higher proportion, and the actual analysis of R1, the FCM-based accident causes identification verification of the effectiveness

of the principal component-cluster identification black spot method.

5. Conclusion

This paper presents a methodology to identify high density accident black spots on rural highways conducted with the combined use of grid clustering and principal component clustering. This method is mainly targeted at some areas with rapid development of rural roads, e.g., China.

The gridding-based clustering tool quantifies an object space into a finite road segment which simplified the analysis of rural roads with vague accident location and complex network. This paper, when analyzing rural roads, replaces the units segmented by the established steps with intersections. This is more suited to rural areas with incomplete geographic information in developing countries. The segment of accident blackspots in road safety still remains a theme worthy of research. This paper provides a simple train of thought for dealing with data in less developed areas of traffic management.

Segments can lead road safety professionals to a better understanding, not only of the location of blackspots but of their circumstances. PCA and improved K-means are combined to identify accident black spots, which consider the rapid development of rural roads in China. To some extent, it allows partial data loss and circumstance changes.

However, this paper does not explain the statistics significance of generated clusters. This is an area of research which is worthy of further study. In addition, we test the recognition results on the basis of accident causes recorded in police reports, combined with reality to conduct analysis by using the widely accepted method-FCM. Nevertheless, before-and-after test is proven to be more convincing for traffic safety testing. In the future, we hope to further cooperate with the Chinese government in implementing measures to improve the black spots. Before-and-after contrast test can conduct with the data of improved black spots. This paper adds significant value to the research on the segmentation of accident road section and the method of how we identity the rural highway black spots.

Data Availability

This research has achieved support and cooperation from Taizhou Public Security Bureau. Data are based on the standard police manual reports. Data included in this study are available upon request by contact with the author (Dr.Shen, shenlinglynn@qq.com). Accident personnel name, the official accident ID, accident location, etc., which are independent of research, will not be provided for privacy and political factors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 51478110). The authors would like to thank Taizhou Public Security Bureau for providing the primitive data.

References

[1] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives," *Accident Analysis & Prevention*, vol. 43, no. 5, pp. 1666–1676, 2011.

[2] N. K. ChikkaKrishna, M. Parida, and S. S. Jain, "Identifying safety factors associated with crash frequency and severity on nonurban four-lane highway stretch in India," *Journal of Transportation Safety & Security*, vol. 9, pp. 6–32, 2017.

[3] M. Lamr and J. Skrbek, "Searching for traffic accident clusters to increase road traffic safety," in *Proceedings of the 24th Interdisciplinary Information Management Talks: Information Technology, Society and Economy Strategic Cross-Influences, IDIMT 2016*, pp. 425–432, September 2016.

[4] M. A. Aghajani, R. S. Dezfoulian, A. R. Arjroody, and M. Rezaei, "Applying GIS to Identify the Spatial and Temporal Patterns of Road Accidents Using Spatial Statistics (case study: Ilam Province, Iran)," *Transportation Research Procedia*, vol. 25, pp. 2131–2143, 2017.

[5] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359–364, 2009.

[6] J. de Oña, G. López, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks," *Accident Analysis & Prevention*, vol. 51, pp. 1–10, 2013.

[7] C. F. Tsai and S. C. Huang, "An effective and efficient grid-based data clustering algorithm using intuitive neighbor relationship for data mining," in *Proceedings of the 2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, pp. 478–483, IEEE, 2015.

[8] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 5, pp. 291–305, 2010.

[9] M. G. Augeri, P. Cozzo, and S. Greco, "Dominance-based rough set approach: An application case study for setting speed limits for vehicles in speed controlled zones," *Knowledge-Based Systems*, vol. 89, pp. 288–300, 2015.

[10] P. Liu, L. Yang, Z. Gao, S. Li, and Y. Gao, "Fault tree analysis combined with quantitative analysis for high-speed railway accidents," *Safety Science*, vol. 79, pp. 344–357, 2015.

[11] C. Ding, X. Wu, G. Yu, and Y. Wang, "A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 225–238, 2016.

[12] M. A. Elliott, C. J. Baughan, and B. F. Sexton, "Errors and violations in relation to motorcyclists' crash risk," *Accident Analysis & Prevention*, vol. 39, no. 3, pp. 491–499, 2007.

[13] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: the fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[14] A. E. C. Mondragon, E. S. Coronado, and C. E. C. Mondragon, "Defining a convergence network platform framework for smart grid and intelligent transport systems," *Energy*, vol. 89, pp. 402–409, 2015.

[15] J. Ahn, E. Ko, and E. Y. Kim, "Highway traffic flow prediction using support vector regression and Bayesian classifier," in *Proceedings of the International Conference on Big Data and Smart Computing, BigComp 2016*, pp. 239–244, IEEE, January 2016.

[16] J. Benedek, S. M. Ciobanu, and T. C. Man, "Hotspots and social background of urban traffic crashes: A case study in Cluj-Napoca (Romania)," *Accident Analysis & Prevention*, vol. 87, pp. 117–126, 2016.

[17] M. B. Ulak, E. E. Ozguven, L. Spainhour, and O. A. Vanli, "Spatial investigation of aging-involved crashes: A GIS-based case study in Northwest Florida," *Journal of Transport Geography*, vol. 58, pp. 71–91, 2017.

[18] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[19] T. C. Coburn, *Statistical and Econometric Methods for Transportation Data Analysis*, Chapman & Hall/CRC, 2003.

[20] G. Lin, Z. Jibiao, D. Sheng, and Z. Shuichao, "Urban road traffic accident analysis based on improved k-means algorithm," *China Highway Journal*, vol. 31, no. 4, 2018 (Chinese).

[21] H. Frigui and O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognition*, vol. 37, no. 3, pp. 567–581, 2004.

[22] C. Jianjun, "Analysis and treatment of traffic accidents involving mass casualties based on the equivalent number of deaths," *Journal of Sichuan Police Academy*, vol. 4, pp. 75–81, 2013 (Chinese).

[23] H. Gallagher, J. T. C. Kwan, and D. R. W. Jayne, "Pulmonary renal syndrome: a 4-year, single-center experience," *American Journal of Kidney Diseases*, vol. 39, no. 1, pp. 42–47, 2002.