ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

The Research of Traffic Accident Hotspots Identification Based on Clustering Ensemble Model

TAO Gang[1, 2, 3], SONG Huansheng[1], Mohsen A. Jafari[4], CHEN Yanxiang [3,5*]
(1.School of Information Engineering, Chang'an University, Xi'an 710064, China;
2. Anhui Keli Information Industry Co. Ltd., Hefei 230088, China;
3. The key Lab of Urban ITS Technology Optimization and Integration, Ministry of Public Security, Hefei 230088, China,
4. Industrial and Systems Engineering, School of Engineering, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901-8554, USA
5. School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract: ***In order to eliminate hidden danger of accident and improve traffic safety, an accident hotspots identification method based on principle component-clustering ensemble model was proposed to analyze and quantify safety levels of different roads, extract principle components, carry out clustering classification for comprehensive evaluation function of principle components through Canopy-Kmeans ensemble clustering algorithm and extract accident hotspots. The hotspots identification experiment on Anhui section of G50 Hu-Yu highway showed that principle component-clustering analysis method can not only be used to carry out scientific accident statistics analysis and effectively identify accident hotspots, but also reflect real traffic safety situation, providing scientific and reasonable basis for improvement of traffic safety decision making performance.***

Keywords: **accident hotspot, principle component analysis, K-Means, Canopy.**

I. INTRODUCTION

With the development of urbanization and the application of scientific technology in the field of traffic and transportation, motor vehicle has been so popularly used all over the world and become an indispensable tool in human's social and economic life, which have greatly facilitated the economic development of the society. However, traffic accidents also followed and became one of the serious public threats in the society. Statistics data shows that every day 140,000 people from the planet were injured, more than 3000 people were killed and more than 15,000 people were disabled for their whole life from traffic accident. Only in 2005, 450,254 accidents happened in our country, causing deaths of 98,738 people, injuries of 469,911 people and direct economic loss of 1.88 billion Yuan. And the mortality per 10 thousand vehicles reached to 7.6[1]. Therefore, it has become an urgent task for the whole society. Researches showed that under the condition of limited capital, number of traffic accidents can be effectively lowered down and management of road traffic safety can be improved through identifying the black-spots and getting the spots managed. Researches from foreign countries on identifying traffic accident hotspots were started earlier and fruitful results had be achieved, such as accident number [2,3,4]; accident rate[5-6], equivalent accident number[7] , quality control[8] and

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

others. These theories provide theoretical basis for traffic accident hotspots identification.

Researches on identifying accident hotspots were started relatively late in our country. However, with the fast development in this field, some theoretical system was established. For example, in 2005 Xian Jiaotong University proposed to establish a normal distribution model of accident black-points through the interaction between accident black-points and proposed curve judgment basis of identifying black-spots [9]. In 2006, Yulong Pei improved quality control method and calculated average accident rate through including the statistics features of accident frequency and adopting gamma distribution in his study, making the identifying result of accident hotspots more accurate [10]. Moreover, Shuang Chen proposed optimized identification method of accident hotspots [11]. Shigui LUO and Wei Zhou proposed conflict judgment method [12].

Since traditional hotspots identification was realized mostly from the perspective statistics analysis of accident data through setting some parameter threshold values, the relationship between close accidents is isolated. In this paper, clustering analysis method was adopted based on using accident statistics data to identify accident hotspots. In previous method of identifying hotspots, multiple variables needed to be established and there were problems of Multi-collinearity, dimension and weight between multiple variables, which caused model to include large amount of redundant information. Therefore, in order to avoid the problem of collinearity between multiple variables, principle component analysis was proposed in this paper to quantify information of different road and extract principle components. After this, a new matrix was formed to carry out clustering analysis after principle component values were obtained. The advantage of this method lies in that theoretical calculation and historical statistics were integrated, which can not only cluster road unit with similar variables, but also reflect individual features of different road units and overcome the coincidence of historical analysis data. The principle is clear and operable.

II. HOTSPOTS IDENTIFICATION IDEA OF PRINCIPLE COMPONENT CLUSTERING ANALYSIS

- Indicator selection: absolute indicators such as number of injuries, death and accidents and direct economic loss were chosen;
- Data standardization;
- Principle component analysis: carry out principle component analysis after data standardization and establish comprehensive assessment model of principle component through specifying cumulative contribution of variance and referring to principle component factor load matrix;
- Cluster Classifying: carry out clustering analysis of comprehensive score factors by using Canopy-Means clustering, obtain comprehensive average score of clusters and select hotspots clusters of accident through the principle that larger mean value will cause higher possibility of accident occurrence.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

III. PRINCIPLE COMPONENT ANALYSIS

A. Data Source

Accident data of Anhui section of G50 Hu-Yu highway （Xuancheng-Wuhu-Tonglin-Chizhou-Anqing）with a length of 481 km from 2011 to 2013 was chosen as case of accident hotspots identification. Total number of accident is 255; number of death is 16 and number of injuries is 31. It is an enclosed two-way and four-lane road, with its width ranging from 21.5 to 23.5 meters and design speed on the road 100 km/h. The road was divided into 481 section units. Table 1 shows the accident statistics data from some of the road units.

Table 1 Accident statistics table from 2011 to 2013 of some road units

ID	No. of accidents	Death and heavy injuries No.	Light injuries no.	Direct economic loss(from ten thousand)
340	25	2	19	19.3
312	1	1	0	0.3
316	3	8	1	5.5
335	2	0	0	0.8
373	2	0	0	0.4
333	2	0	0	0.4
363	2	1	0	0.7
332	2	1	0	0.5
…	…	…	…	…
354	1	0	1	0.2

B. Indicators Selection

Based on previous researches on accident hotspots identification, four indicators including number of accidents, number of death, number of injuries and direct economic loss were chosen as factors set. These four absolute indicators can not only reflect features of accident hotspots but also simplify calculation in actual operation to make it highly operable.

a. Standardization of Accident Data

Data standardization refers to the process of scaling attribute values according to a specific ratio and make them fall into a relatively smaller section (such as [0, 1])). This method works effectively for data mining algorithm of which the core is distance measure such as nearest neighbor, K-Means, FCM and others.

Through standardization, deviation of data mining result which is caused by variation of attribute values can be eliminated. Since input fields with high value will drive the whole training process and cover the influence of input fields with lower value on the training model. Currently, there are mainly two types of standardization:

- Min-Max standardization (Equation 1)

$$v' = \frac{(v - \min_A)}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \qquad (1)$$

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

This type of standardization can change original data and maintain the relationship between original data.

- b z-score standardization (Equation 2)

$$v' = \frac{v - \overline{A}}{\sigma_A} \qquad (2)$$

$\overline{A}$ and σ_A are average value and standard deviation of A. When the maximum value and minimum value of attribute A is uncertain, this method will be very effective.

b. Steps of Experiment

(1) Data Preprocessing

The research in this paper was based on SPSS Modeler data mining platform which includes five steps of CRISP-DM industry standard: data understanding, data preparation, analysis of model, assessment and deployment. Data preprocessing was carried out through SPSS Modeler. The flowFigure of preprocessing was shown in Figure 2:

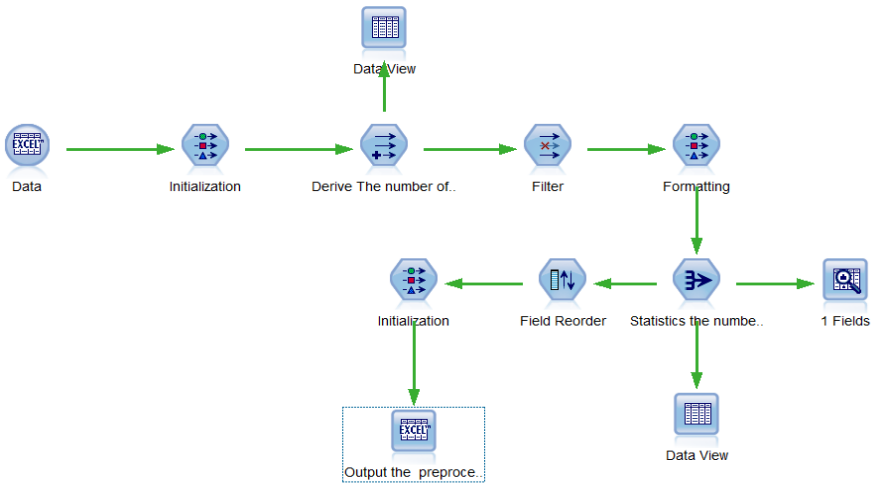

Fig 2 Flow Figure of accident data preprocessing

(2) Principle Component Extraction

Principle component extraction of preprocessed data was carried out through principle component analysis and factor analysis module in SPSS Modeler platform. Table 2 showed the initial Eigen value, extracted factor Eigen value, cumulative contribution rate of variance and the corresponding principle component for Eigen value of cumulative contribution rate of variances which is more than 85%.

Table 2 Eigen values of variables and cumulative variance contribution rate

component	Initial Eigen value			Eigen value of extracted factors		
	Eigen value	variance contribution rate	Cumulative variance Contribution rate	Eigen value	variance contribution rate	Cumulative variance Contribution rate
1	2.484	62.107	62.107	2.484	62.107	62.107
2	0.926	23.162	85.269	0.926	23.162	85.269

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

3	0.366	9.161	94.429	-	-	-
4	0.223	5.571	100	-	-	-

(3) Steep Order Test of Factor Variables

Figure 4 is the factor variables steep order test screed Figure of sample principle component analysis, with horizontal axis representing principle component serial no. and vertical axis representing Eigen value of common factors. It can be seen from the Figure that Eigen values of the first two components change a lot while from the third component, the changes of their Eigen values are minor, which means that the requirement of analysis of original variables can be met by extracting the first two factor variables.

(4) Comprehensive Assessment of Principle Component Factors

Table 3 showed the load factors matrix after the rotation of two extracted components through variable matrix. It can be seen from Table 3 that the first component represents economic indicator and reflected indicators of light injury no. and accident no. The second component represents indicators of death no. and heavy injury no. Corresponding eigenvectors of Eigen value can be calculated by using extracted principle component and principle component factor load matrix in Table 3. Corresponding Eigen values of eigenvectors $a_1$anda_2 can be calculated through dividing the square root of Eigen value by factor load matrix.

a_1=（0.53，0.267，0.554，0.587），a_2=（-0.343，0.94，0.098，0.02）

Comprehensive factors will be summarized according to variance contribution rate which will be considered as weight as shown in Equation 3.

$$F = \eta_1 F_1 + \eta_2 F_2 \qquad (3)$$

η_1 and η_2 represents variance contribution rates of first and second principle components. Comprehensive scores of 84 road sections will be obtained (some of these were shown in Table 4). Since different indicators were standardized, negative numbers appeared at last. A high comprehensive assessment value represents a high rate of accident occurrence and low value represents low rate of accident occurrence.

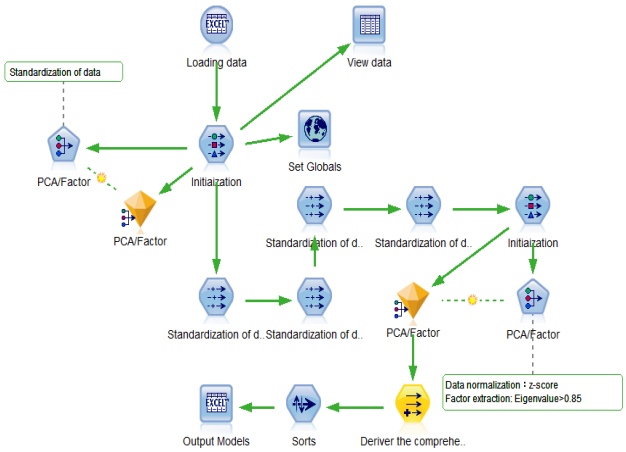

Fig 3 Model establishment Figure of principle component analysis

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

Table 3 Load Matrix of Principle Components

Name of variable	Principle component	
	1	2
No. of accidents	0.834	-0.33
No. of death and heavy injuries	0.419	0.899
No. of light injuries	0.872	-0.094
Direct economic loss (from ten thousand Yuan)	0.924	-0.021

Table 4 Principle component and comprehensive scores

Road No.	Accident No.	Death and heavy injury No.	Light injury No.	Direct economic loss / (from ten thousand Yuan)	$Fctor-1	$F-factor-2	Comprehensive score
340	5.81321774	1.86438858	8.525561884	6.4726186	7.681029681	1.16512306	5.040323
324	-0.538786	-0.2237266	-0.253016675	-0.530252	-0.483597559	-0.144523016	-0.33382
369	-0.538786	-0.2237266	-0.253016675	-0.530252	-0.483597559	-0.144523016	-0.33382
312	-0.538786	0.82033098	-0.253016675	-0.457306	-0.60372029	0.880175155	-0.17109
335	-0.2741192	-0.2237266	-0.253016675	-0.274939	-0.277853024	-0.182677667	-0.21488
363	-0.2741192	0.82033098	-0.253016675	-0.311412	-0.437415958	0.831748341	-0.07902
353	-0.538786	-0.2237266	-0.253016675	-0.201993	-0.36527695	-0.113706525	-0.2532
332	-0.2741192	0.82033098	-0.253016675	-0.384359	-0.463709427	0.824900232	-0.09693
…	…	…	…	…	…	…	…
354	-0.538786	-0.2237266	0.209013775	-0.493779	-0.301815606	-0.135442107	-0.21882

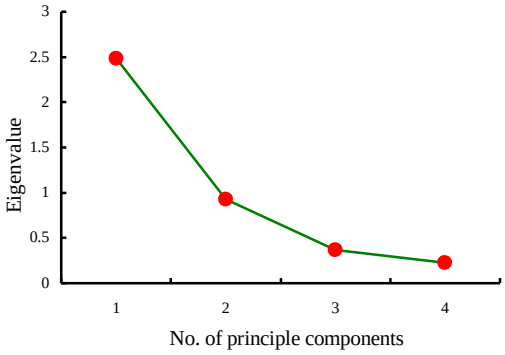

Fig 4 Scree Figure

IV. CLUSTERING ANALYSIS

Clustering analysis for two field, road section and comprehensive score, in Table 4 was carried out through Canopy clustering algorithm to determine the optimum number of clusters, which would be used as levels of accident. K-means algorithm would be used to select final accident hotspots. Figure 5 is the clustering overview Figure, from which we can see three clusters with each cluster representing danger level of accidents were classified among G50 highway sections of accident occurrence. According to the principle, the high value represents high possibility of accident occurrence and thus

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

high possibility of being accident hotspots. The hotspots identification results were shown in Table 5.

In our country, accident occurrence hotspot is defined as a road with its length reaching 2000 meters or bridge or culvert on which more than three serious accidents happened within one year. From Figure 6, it can be known that annual average accidents on hotspots and potential hotspots reached up to 30 and annual death number reached more than 3, which on one hand proved the rationality of accident hotspots identification method proposed in this paper based on principle component clustering analysis and on the other hand reflected the occurrence trend of frequently occurring and easy-to-occur features of accidents. Therefore, traffic management needs to be strengthened to carry out cause analysis of identified accident hotspots and take corrective actions.

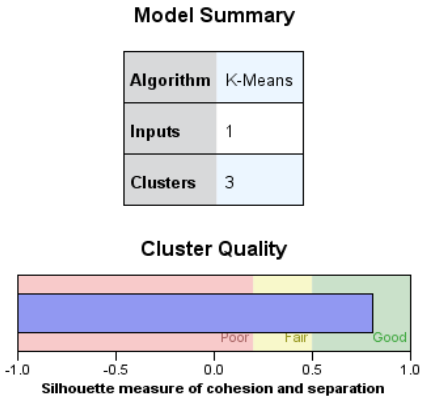

Fig 5 Clustering overview

Table 5 Hotspots identification results

Cluster No.	Road section No.	Comprehensive score	Accident hotspots identification
1	340	5.04	Accident hotspot
2	The rest（75）	-0.103346155	Non-hotspot/good traffic safety
3	567、358、319、400、315、316、322、314	0.844702848	Potential hotspot

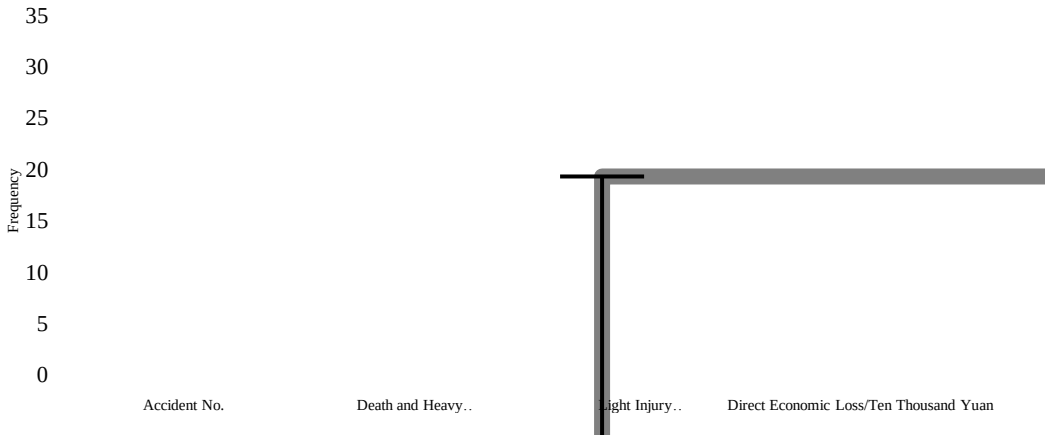

Fig 6 Average annual distributions of different indicators of accident hotspots and potential hotspots from 2011 to 2013

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 1, January 2016

V. CONCLUSION

The principle component clustering analysis method eliminated the defects that multiple variables were needed for establishing mathematic model and avoided the collinearity problem of multiple variables. The safety of different road sections was quantified through scores of principle components. The principle is clear and highly operable and highway accident hotspots can be well identified. Theoretical calculation and historical statistics value were integrated in this algorithm, which can contribute to better clustering for road units with similar variables, reflect individual features of different road units and overcome the coincidence of historical statistics data. Moreover, the proposed algorithm in this paper can also be referred to for identifying hotspots for other roads of different classes. Only modest adjustment of some indicators is needed in actual application.

ACKNOWLEDGEMENT

This work is partially supported by key projects of Anhui Province science and technology plan （15czz02074）, and Anhui Province Nature Science Foundation of China (1408085MKL76).

REFERENCES

[1] Wenjie CHEN. Analysis of road traffic accident black spots [J]. Journal of Chinese People's Public Security University, 2004, 2(2):83-88.

[2] ITE Technical Council Committee. Road safety Audit: A New Tool for Accident Prevention.ITEJournal1995, 66(2):1-6.

[3] RA. Krammes. Interactive Highway Safety Design Model: Design Consistency Module. Public Roads. 1997, 77(5):12-17.

[4] S. Oppe. Development of Traffic Safety Global Trends and Incidental Fluctuations. Accident Analysis and Prevention. 1991, 23(1):58-60.

[5] C. C. Wright , C. R. Abbess , D. F. Jarrett. Estimating the Regression-to-Mean Effect Associated with Road Accident Black Spot Treatment: Towards a more Realistic Approach. Accident Analysis & Prevention. 1988, 20(3):199-214.

[6] Ziad A. Sawalha. Traffic Accident Modeling Statistical Issues and Safety Applications. A thesis to the University of British Columbia for the degree of doctor of philosophy. 2002:119-129.

[7] Larsen. Lotte. Methods of Multidisciplinary in-depth Analyses of Road to Add Traffic Accidents. Journal of Hazardous Materials. 2004, 111(3):115-122.

[8] Jake Kononov. Road Accident Prediction Modeling and Diagnostics of Accident Causality a Comprehensive Methodology. A thesis to the University of Colorado for the degree of doctor of philosophy. 2002:57-80.

[9] Yuzeng LIU. Research on Intelligent investigation and Countermeasures of traffic accident black spots[D], 2005.

[10] Pei Yu long, Ding Jianmei. Improvement in the quality control method to distinguish the Black spots of the road. The 6th Conference of the Eastern Asia Society for Transportation Studies. 2006:2106-2113.

[11] Shuang CHEN. Comparative study on identification of road accident black spots [J]. Shandong traffic science and technology. 2005, (3):7-9.

[12] LUO Shi-gui, ZHOU Wei.Survey way of road traffic conflict technique [J]. Journal of Changan University (Natural Science Edition).2001, 18(1):65-68.