Engineering Journal issn: 0125-8281

Article

Hotspot Location Identification Using Accident Data, Traffic and Geometric Characteristics

Yousef Sajed[1,a], **Gholamali Shafabakhsh**[1,b,*], **and Morteza Bagheri**[2]

1 Faculty of Civil Engineering, Semnan University, University Sq., P.O. Box 35196-45399, Semnan, Iran
2 Faculty of Railway Engineering, Iran University of Science and Technology (IUST), Iran
Email: [a]yousef.sajed@semnan.ac.ir, [b]shafabakhsh@semnan.ac.ir (Corresponding author)

Abstract. Determining the criterion for critical limits is always one of the essential challenges for traffic safety authorities. The purpose of identifying accident hotspots is to achieve high-priority locations in order to effectively allocate the safety budgets as well as to promote more efficient and faster safety at the road network level. In recent years, human, vehicle, road and environment have been recognized as the three main effective elements of the road transportation in the occurrence of accidents. In the present study, with combining the parameters related to accidents, geometric parameters of the accident location and traffic parameters, hotspots were identified. At the present research, were used the superior methods of Poisson regression and negative binomial distribution and based on the combined criteria of frequency and severity of accidents and equivalent injury factors by a floating segmentation of road. Then using Time Series Models in ANN, result were compared and validated. The results of ANN models demonstrate that the frequency method of accidents tends toward places with high traffic volume. MATLAB and STATA software were used. Non-native plumbing, curvature, slope, section length and residential area had more significance, and their coefficients indicated the significant effect of these parameters on the occurrence of the frequency and severity of accidents in hotspot locations.

Keywords: Hotspots identification, regression models, frequency and severity of accidents, ANN models.

ENGINEERING JOURNAL Volume 23 Issue 6
Received 3 May 2019
Accepted 26 August 2019
Published 30 November 2019
Online at http://www.engj.org/
DOI:10.4186/ej.2019.23.6.191

DOI:10.4186/ej.2019.23.6.191

1. Introduction

Correct identification of accident prone areas has a significant effect on the reduction of road accidents and damages caused by it. Due to the lack of domestic studies and researches on the identification methods of accident hotspots, the introduction of a comprehensive and localized method for identifying hotspots is required more than before. It is obvious that paying more attention to the topic of the accident hotspots as the first and most important step in all road safety projects and localizing the identification methods for hotspots can reduce the growing problem of road accidents [1]. Suburban accidents constitute a major part of accidents. The statistics show that death in suburban roads accounts for more than 69 percent of the total deaths caused by accidents in the country. The targeted and systematic reduction of accidents requires a comprehensive road safety management. Introducing accident hotspots is the first step in the road safety management process. Accident hotspots are sometimes recognized by definitions such as: hazardous road locations, high-risk locations, accident-prone situations, places requiring improvement and etc. [2]. With the development and implementation of road safety improvements, two major objectives are followed: the identification of accident hotspots and the assessment of areas with the highest potential for reducing accidents [3]. Investigating and studying the accident hotspots in Iran are in low and inadequate level due to the lack of a codified plan to identify and prioritize these spots and the appropriate database in which the identification of country hotspots can be registered and updated after supplying the validity and implementation of the corrective measures of its data. While neither are valid scientific methods used to identify and prioritize them, nor are the effectiveness and reduction of accidents in these spots evaluated after spending cost and securing them [1].

2. Literature Review

Accident Hotspots: There are many definitions for accident hotspots. However, conducted researches has emphasized that there is no comprehensive accepted definition of what is termed "dangerous." Elvik (2007) defined the hotspot as "every spot that has a higher number of accidents than other similar spots due to local risk factors" [4]. This definition refers to this concept that hotspots are the spots where the risk factors of the geometric and traffic design have a lot contribution in accidents and are reduced by the accident engineering strategies [5]. An accident hotspot is a spot in which at least 10 accidents occurred during a three-year period, or at least four accidents occurred during one year [6]. The definition of hotspot in different countries is presented in Table 1.

In general, there is no acceptable and definite definition for accident hotspots. Usually, accidnet hotspots are places in which there is a probability of high risk or accident occurrence. These spots are locations in which the potential for accidents is unacceptably high. The risk of accident occurrence is not the same throughout a road network. In certain situations, the level of risk is higher than the overall levels of risk in adjacent areas that more accidents occur in these situations. Although the term "hotspot" refers to a certain position, it is often used for sections of the road. These spots are usually found in certain areas of the road, such as crowded intersections and sharp curves. According to the Australian Department of Transportation and Economy, the locations are classified as hotspots or black spots after identifying the risk level and the probability of an accident occurrence in each location. As mentioned above, in certain locations, the level of risk is greater than the overall level of danger in the surrounding areas, and the accidents will occur more in these places with a high relative risk [3]. Preventive or observation-based method: that emphasizes on analyzing the physical and functional characteristics of the road for identifying existing road safety problems or road construction projects. This method is called Road Safety Inspection [1].

In identifying accident hotspots, the first method has been emphasized by more researchers, and its ability for identification of correct spots is more. But the first method requires accident information. Unfortunately, in many less developed countries, the importance of accurate recording of accidents for future uses is not explained and their accident database has many shortcomings. In our country, due to the long length of roads and the fewer police presence in roads with the lower importance grade, and most importantly, the failure to record many accidents that do not have plaintiff (generally single-vehicle or damage accidents), as well as the lack of recording geographic coordinates, there is no proper information for identifying accident hotspots with the first method. These issues highlighted the importance of using safety inspections.

Table 1. Definition of accident hotspot in different countries.

Country	Accident Hotspot Definition
Germany	Road sections with the length of 300 m, occurrence of more than 3 similar accidents during one year, occurrence of more than 5 accidents during three years
England	Road sections with the length of 300 m, spot in which the total number of road accidents is more than 12 accidents over three years.
Spain	Road sections with the length of 1km, occurrence of more than 5 injury accidents or 2 fatal accidents during one year, occurrence of more than 10 injury accident or 5 fatal accident during 3 years
Czech	Road sections with the length of 250m, occurrence of at least 3 injury accidents during 1 year, occurrence of at least 3 similar injury accidents during 3 year, occurrence of at least 5 similar accidents during 1 year
Netherlands	Occurrence of at least 10 accidents, totally occurrence of at least 5 accidents with similar properties of analysis period is 3 to 5 year.

Source: Rahimov and Haj Ali (2011)

Table 2. Comparison of Criteria for hotspots in some countries (Source, Astaraki).

Country	Section's length	Frequency
Australia	A short section	During 5 years at least 3 accidents
England	300 meter	During 3 years 12 accidents
Germany	300 meter	During 3 years 8 accidents
Norway	100 meter	During 3 years 4 accidents
Portugal	200 meter	During 3 years 5 accidents
Thailand	variable	During one year at least 3 accidents

In this research, the geometric and traffic characteristics of the study route are extracted based on field surveys and inspections. The recorded spots are reviewed and examined closely as accidents in COM forms of police 114 (Computerized forms of accident registration) in road, and the accurate geographical coordinates are recorded for them to have the ability to exact analysis and transfer to the route aerial map as well as GIS.

In general, nine methods for identifying hotspots are presented in various and valid sources. Each of them is briefly mentioned: Accident Frequency Method, Accident Rate Method, Accident Critical Rate method (Qualitative Control), Equivalent Property Damage Only Index (3 EPDO Index), Accident severity Method, Developed Accident Severity Method, Accident Number-Severity Method (Used in Organization of Road and Road Transport), Lost Cost Method (cost of losses + cost of heavy injuries + cost of light and possible injuries + cost of financial damage = lost value), Accident Density Method.

Road Safety Inspection: The principles of this method are based on existing inspections and identifying the weaknesses and shortcomings of the route. Normally, inspection costs will account for less than 0.5% of the total costs of the road construction project, but will result in significant investment return.

The risk is the probability of an accident occurrence, which can be expressed as a rate with a number of measurement values including time, traffic flow, road length, or road user interaction.

In a research by Haghighi, seven commonly applied HSID methods (accident frequency (AF), PIARC coefficient based equivalent property damage only (EPDO), P-value (Islamic Republic of Iran Ministry Roads and Urban development), accident rate (AR), combined criteria, empirical Bayes (EB), societal risk-based) were compared against six robust and informative quantitative evaluation criteria (site consistency test, method consistency test, total rank differences test, total score test, sensitivity test and specificity test). These tests evaluate each method performance in a variety of areas, such as efficiency in identifying sites that show consistently poor safety performance, reliability in identifying the same black spots in subsequent time periods .To evaluate the HSID methods, three years of crash data from the Kerman state were used. Analytical

Hierarchy Process (AHP) method has been used for determination the importance coefficients of evaluation tests and as a result, showed that the total rank differences test is the most appropriate test. The quantitative evaluation tests showed that the EB method performs better than the other HSID method. Test results highlight that the EB method is the most consistent and reliable method for identifying priority investigation locations [7].

Studies conducted over the last few decades show that design elements influence the road safety. Design elements include cross-section features, horizontal path, vertical path, road shoulders, intersection design (level and non-level), lighting, accesses and how to control them, and pavement quality. In many conducted studies in this field and the proposed model, the fit analysis has been used to explore the relationship between accident rate and design elements. Dickon and his colleagues discussed the difference between the accident hotspots and sections [8], as well as in the field of the accident severity assessment for this group they suggested a numerical weight of 9.5 for fatal accidents and severe injuries and they suggested numerical weights of 3.5 for mediate and light injury accidents. Taylor and Thompson presented a method in which the total weight of the eight factors was used as an indicator of being risky [9]. These eight factors include accident frequency, accident rate, accident severity, V/C ratio, distance of vision, collisions, sudden movements and expectations of drivers. McGuigan, like Jurgenson, suggested that for each section or intersection of the network, the difference between the existing accident frequency and the accident frequency expected for that section or intersection should be considered as a capability of improvement for each network component. The higher the capability, the higher the desired intersection or cross-section rank in the final ranking [10]. Sohn used the card method, neural networks and regression analysis to categorize accidents based on the severity in Korea, and showed that seatbelt and helmet are the most important determinants of accident severity [11].

Karlafits studied Indiana crashes using tree regression analysis and found that, the most important causes of road accidents for two-lane suburban roads are daily traffic volume, shoulder width, pavement service and pavement surface friction, respectively. While the most important factors in the occurrence of accidents for multi-lane roads are the importance of daily traffic volume, middle-of-the-road width, friction of pavement surface, road width and pavement service index, respectively [12]. Cheng and Washington used three simple ranking methods based on the accident frequency, confidence intervals, and empirical Bayes methods to identify accident-prone intersections of Arizona State. The results of their work indicated the superiority of the empirical Bayes method relative to the two mentioned methods in identifying high-risk intersections [13]. El- Basyouny and Sayed employed the Multivariate Poisson Log-Normal method to identify the accident prone intersections of the city of Edmonton, Canada. They used two criteria of frequency and severity to identify and prioritize accident prone intersections [14]. Cafiso in a study on the second-class suburban roads (which were constructed on the basis of lower standards, with more sharp and horizontal curves), found that on average, these curves had a larger share of accidents [15].

Shariat Mohaimani and Tavakoli evaluated the severity of injuries caused by accidents in two-lane suburban roads using data mining models. According to their studies, it seems necessary to pay attention to the construction of overtaking bands and the intensification of applying regulations to reduce the hazardous overtaking maneuvers in these roads [16]. Kazemi and Zoghi identified and prioritized the accident hotspots on the suburban roads and provided a software. The results of their studies indicated that the criterion of accident frequency identifies places that have higher traffic volume, the criterion of accident rate identifies places with lower traffic volumes, and the criterion of severity also identifies locations that are outside of the city. None of these criteria is necessarily the best. Each of them highlights the problems from their own point of view. As a result, it is better to use more than one criterion for identification and compare the results [1]. Sadeghi identified and prioritized accident prone sections with the route segmentation approach and data envelopment analysis. Comparison of road sections using linear programming in the data envelopment analysis framework provides a method that can be used to prioritize road sections, intersections, fields, or the entire paths of an area for the road safety organization in terms of the other road safety. In the present study, the relative inefficiency of 154 road sections was obtained. It is a new experience in terms of defining input and output indicators based on the data envelopment analysis method for prioritizing the road sections. With the current method, a number of sections are neglected despite inappropriate (efficiency) performance, while at a low cost, we can achieve better results. As a result, this method yields privileges (inefficiency) that allow the road sections to be properly ranked and prioritized [5]. In the another research, It is attempted to identify and prioritize the accident prone points (black spots) in "Iraanshahr-Sarbaaz-Chabahar" road located in Baluchistan, Iran, without no use of accident data but rather using Analytic Hierarchy Process (AHP), which is the enhanced procedure of road safety audit technique [17]. Another paper aims at presenting a novel

approach, capable of identifying the location as well as the length of high crash road segments. It focuses on the location of accidents occurred along the road and their effective regions. In other words, due to applicability and budget limitations in improving safety of road segments, it is not possible to recognize all high crash road segments. Therefore, it is of utmost importance to identify high crash road segments and their real length to be able to prioritize the safety improvement in roads [18].

Vosoughifar examined how to identify hotspots using the Geographic Information System (GIS). According to the results of this study, arches, intersections and roadside installations play a significant role in increasing the risk of road accidents [19]. Saffarzadeh using statistics and information related to 580 km of the country's main roads, using a SPSS software, presented a mathematical model in which the effect of ADT on the risk index is significant [20]. Nassiri and Moshfegh presented two relationships for the risk index with the combination of different components [21]. Safety visits is a method that was investigated in the study of Sajed and Azimi. In order to study the case, 14 tunnels of suburban roads of Ardabil province were selected and then important factors affecting the safety of tunnels, such as tunnel lighting and tunnel placement in archways were determined and Tunnel Safety Index (SI) were calculated for their safety assessment and compared with the accident statistics of tunnels. Finally, the Risk levels (RL) of the tunnels were proposed to prioritize the safety measures [22].

Regarding the proposed models, it can be assertively stated that the results of each model is reliable within the range of conditions in which the parameters of the model are taken into account. For example, weather conditions, driving behavior strongly influence the model results. Therefore, regardless of the changes in the factors involved, the direct use of a model and its results leads to the inconsistent results. For developing countries, the pattern of accidents is strongly influenced by human factors. What is important is the fact that the pattern of accidents in one section is influenced by the selected parameters and its nature is formed based on the spatial and temporal conditions of the selected variables. To this end, the process of accident model for Iran's roads has been reviewed and presented with the aim of identifying the parameters and factors involved in the rate of accidents and their relationship.

For a contribution, in addition injury and fatality, non-injury (only monetary) accident records are considered in the modeling in this research. In addition, the presented models can be identified high risk spots and segments on the road based on accident severity and frequency. In this research, a floating segmentation method with a fixed length of 2 km were used that has advantages in competition with fixed segmentation method.

3. Methodology

The accident prone location refers to the location in which the accident indicator exceeds a critical value (critical criterion). It should be noted that the location can be a spot or a road section of or a range. In this way, it is essential to know the types of indicators and related criteria for the identification of accident prone locations. In scientific sources, accident prone locations are often referred to as black spots, high accident areas or hotspots.

Whenever an accident index exceeds a certain limit, then the critical condition for a spot or section is created. Accordingly, that spot and section are identified as an accident prone location or black spot. Therefore, determining the criterion for critical limits is always one of the essential challenges to traffic safety authorities. The purpose of identifying high-accident locations is to reach priority locations in order to allocate optimally and effectively the safety budgets as well as to promote more efficient and faster safety at the level of the road network. Obviously, a suitable criterion for communities depends on factors and parameters such as annual safety budgets, technology levels, the amount of trained personnel, community operating strengths, and safety strategic plans and projects. Therefore, it is not possible to prescribe a definite and stable criterion for different communities. In recent years, human, vehicle, road and environment have been recognized as the three main and effective elements of road transportation in the occurrence of accidents.

3.1. Conceptual Model of Research and General Structure of Research

In this research, a risk model based on the accident index (the number calculated for an accident hotspot in a road and in the study time period) for accident hotspots (a spot in a road or road segment with a maximum length of 1 km on which at least one fatal or injury accident has occurred) is determined to identify and prioritize accident hotspots of the suburban roads.

DOI:10.4186/ej.2019.23.6.191

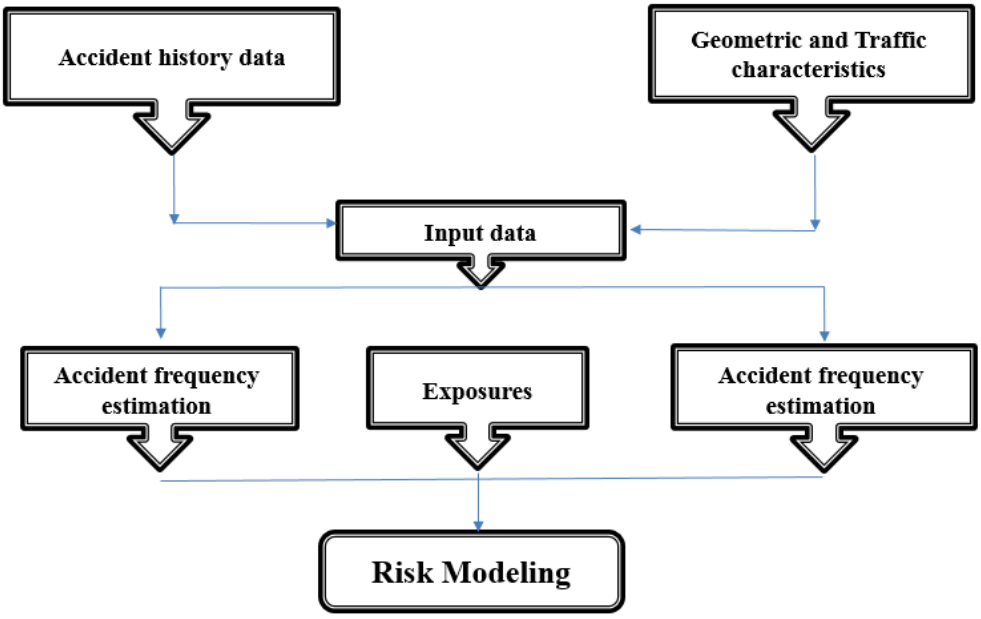


Fig. 1. Conceptual model of research (Identification process of the accident hotspots and segments).

In this research, a general approach is done in three phases (Fig. 1):

1. Examining and evaluating the current conditions on the studied roads (safety inspections through field surveys as well as compliance with geometric design standards), which include:
   1.1. Traffic characteristics of the road including: annual average daily traffic volume, traffic speeds, traffic combination and traffic performance of the route, and etc.
   1.2. The geometric characteristics of the road including: the length of the segment (section length), the width of the roadway, the vertical and horizontal curves, the maximum slope and curvature*slope.
   1.3. Road safety and warning signs including: Vertical and horizontal warning, limiting and route warning signs, type and conditions of safety equipment for the sides.
   1.4. Personal information of drivers and users of the route including: age, gender, level of education and familiarity with the route (lic).
2. Examining the history of accidents occurring in the site, including: the number and severity of accidents, the incident time, the cause of the accident and the location of occurrence and their statistical analysis (these statistics should have proper accuracy and dispersion in the country). For statistical analysis and validation of models, statistical software such as SPSS and STATA as well as MATLAB (for neural networks) will be used.
3. Providing an appropriate risk model for identifying and prioritizing spots based on new combinational methods and validating those using Regression Models Time Series Models based on the type, the amount and the dispersion of data.

Figure 2 illustrated that the major problems of segmentation with the sequential fixed length. Table 3 shows the percentage of accidents covered by accident-prone segments along the route.

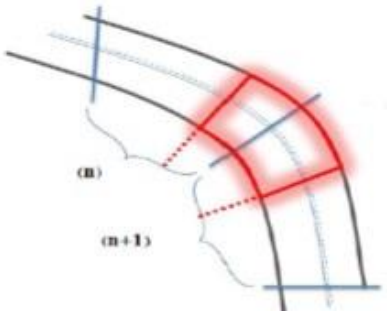
The usual approach in investigating the factors affecting road accidents is that the road network is firstly segmented, and high-accident segments in the road network are identified by selecting a suitable performance criterion and during the network screening process. Then, the relationship between various factors of traffic, human, route geometry, or a combination of them with an occurred accident in high-accident segments of the route is established using a mathematical model, and the effect of each factor in the occurrence of accidents is examined. The mathematical models used in previous researches are divided into three general categories of regression models, multi-criteria decision models, and pattern recognition models. In this research, suitable regression models and pattern recognition have been used.

The main objective of this research is to investigate the causes and factors of road accidents based on various traffic and geometric parameters of the route and present a risk model to identify accident hotspots. For this purpose, the network of roads is first segmented and screened with the aim of identifying high-

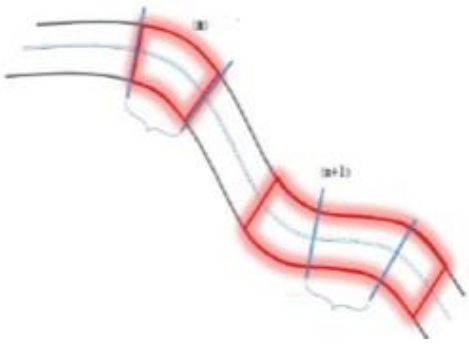
accident segments. The usual methods of segmentation due to the inherent weakness of the fixed assuming the length of the segments (static method) are not able to correctly identify the boundary of the road segments. Hence, in this study, dynamic segmentation method (with fixed length, but floating) has been used (Fig. 2).

After segmentation and screening of the network of roads and identification of the high-accident segments, the effect of human factors, traffic characteristics and road characteristics (as the most important factors affecting the accident occurrence) are determined by collecting and measuring a set of visible variables in high-accident segments and using the proposed risk model. The method is that after identifying the factors affecting the occurrence of accidents, measurable variables of each of the effective factors are collected and prepared. and the relationship between these factors with each other and how they affect the occurrence of accidents are determined based on regression modeling and pattern recognition (Artificial Neural Network) with a partial least squares approach (based on a set of theoretical principles reflecting the cause and effect relationship of factors affecting the occurrence of accidents with each other).

The main advantage of the proposed model for identifying the hotspots is its applicability based on traffic and geometric information of the route and accident data with the specific spatial coordinates based on a dynamic segmentation method that includes both the severity and the number of accidents. Although the coefficients presented by the Ministry of Roads and Urban Development have been used to determine the Equivalent Property Damage Only Index, in the modeling and in spite of the simple identification and prioritization method presented in this instruction (which just consider the Equivalent Property Damage Only Index of accident severity and road performance and segment length), using this information, a comprehensive risk model for identifying hotspots has been proposed that with the maximum coverage of the high-accident segments, it can provide a suitable practical approach for prioritizing and budgeting and modifying those spots.



a. Failure to coordinate the accident-prone segment position with the defined segments.



b. Failure to coordinate the length of the accident-prone segment with the length of the defined segments

Fig. 2. Major problems of segmentation with the sequential fixed length.

In the floating segmentation method with a fixed length of 2 km, the highest percentage of accident coverage was obtained for segments with sequential and floating fixed lengths of 3 and 5. Due to the segmentation of the accident hotspots based on the above-mentioned method, in the adjacent space units we can also determine the boundary and length of the high-accident segments of the route better than the

previous methods. Therefore, the prioritization of the high-accident segments of the route by using this method to allocate the route improvement credit is more optimal than other methods and with a specific amount for improvement of the route, more percentage of accidents will be covered.

Table 3. Percentage of accidents covered by accident-prone segments along the route.

Segmentation method	With floating fixed-length segmentation			With sequential fixed-length segmentation		
Accident percentage	5	3	2	5	3	2
	72	79	85	69	77	81

3.2. Regression and ANN Models

The most common statistical models in the discussion of road safety are the Poisson regression model and negative binomial distribution. These methods are used to model discrete, independent and positive events. These models are employed to select variables and also for modeling accidents [23]. Due to the fact that the accident data has an over dispersion (in the sense that the data variance is greater than the mean), a negative binomial model is used to overcome this problem. For such data, the Poisson method should not be used. For example, if the Poisson model is used to estimate the number of expected accidents, a great difference is created between occurred and estimated accidents [24]. A negative binomial model is the modified Poisson model for solving problems of data with over dispersion. This model is based on the assumption that the Poisson parameter has a gamma distribution. The model is derived from a closed equation and the mathematical rules for solving the relationship between the mean and variance are almost straightforward. The negative binomial model is obtained by rewriting the Poisson parameter for each observation i as follows:

$$\lambda_i = EXP(\beta X_i + \varepsilon_i) \tag{1}$$

where: EXP) ε_i (is the gamma distribution error with the mean 1 and the variance α. Adding this equation allows the variance to differ from the mean as follows:

$$VAR[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2 \tag{2}$$

Poisson's regression model is a mode of a negative binomial model in which α reaches zero. It means that the choice of one of these two models depends on α. α is usually called the over dispersion parameter. The Poisson Gamma model is the most commonly used model for estimating the frequency of accidents.

Then using Time Series Models (TSM) in ANN, result were compared and validated. The results of artificial neural network models demonstrate that the frequency method of accidents tends toward places with high traffic volume and in addition, the severity of the accident is not considered in this method. MATLAB-2015a and STATA software were used.

3.3. Case study

In the present study, with combining the parameters related to accidents (including accident time, accident cause and accident severity), geometric parameters of the accident location (including: road width, shoulder width and radius of horizontal and vertical curves, road surface conditions), and traffic parameters (including: average daily traffic volume, heavy traffic percentage and average speed of route in accident day) were combined. There for, three-year statistics and information (2015-2017), which have taken place over 130 km from the main two-lane Ardebil-Sarcham (Fig. 3) suburban road and inserted in the police accident registration forms and all the information along with the geographical coordinates of the spots on the route has been reviewed, the route accidents in different sections have been modeled and evaluated using Poisson regression models and negative binomial distribution of the number and severity of accidents in the STATA software. Figure 4 displays the distribution diagram of accidents along the route and per kilometer.

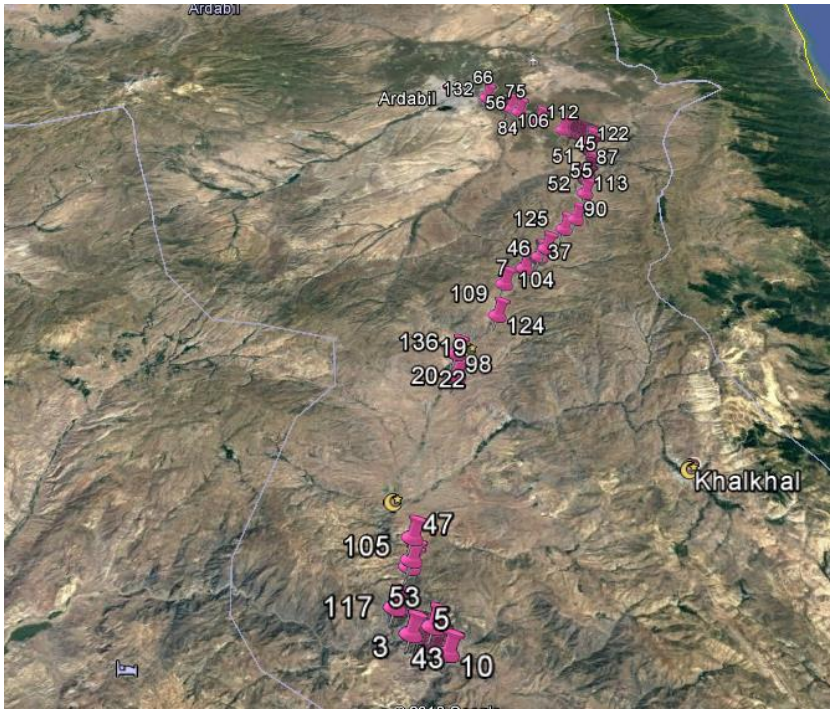


Fig. 3. Global Positioning Systems of accident location (GPS\ Ardabil-Sarcham road).

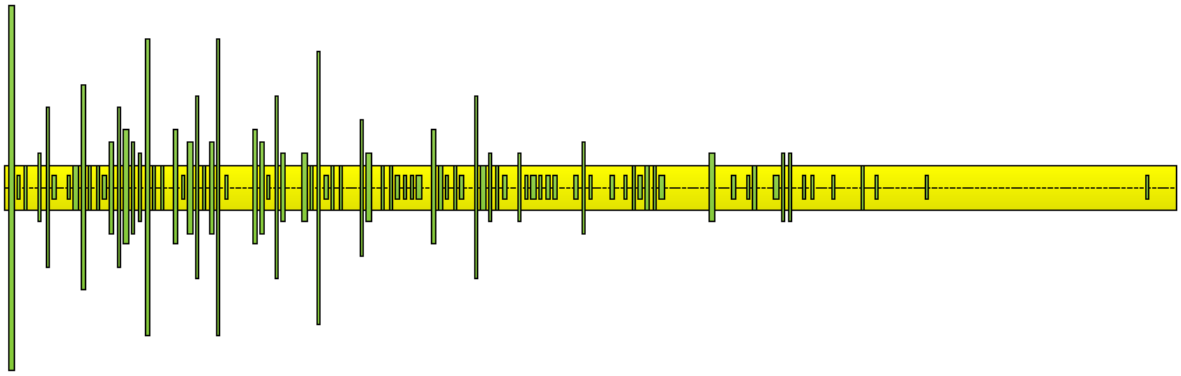


Fig. 4. Distribution diagram of accidents along the route (130 kilometer).

3.4. Data and Modeling

Table 4 indicates the sample and the names of independent and dependent variables and their values, as well as the categorization of quantitative and qualitative variables that are prepared for modeling in the Excel software environment and entered into the statistical software STATA and MATLAB. Table 5 shows the sample of road segmentation that sorted by accident frequency (this table sorted by 18 to 1 accident for 1 or 2 kilometer segments) along the route (130 kilometer). Table 6 shows the names and characteristics of hidden and observable variables as well as their indexes for human factor parameters, geometric characteristics of the road, traffic characteristics, and accidents Also, Table 7 displays the route data sample calculated according to Tables (4 to 6) for two segments (30-31) and (2-3), which have the highest accidents.

3.4.1. Regression Modeling

Data from three years of Ardabil-Sarcham suburban road accidents were processed in EXCEL environment and entered into the software environment of STATA for regression modeling. The output of the model and the definition of the parameters are given in Tables 8 and 9.

Table 4. Samples and names of independent and dependent variables and their values.

Variable name	**Specifications**	**Variable type/value/unit**
A.S.	Dependent variable (with coefficient)	Fatal\injury\damage (equaled)
A.F.	Dependent variable (accident frequency)	Number of accidents
S.C.	segment length	1 or 2 km
Volume	Hourly traffic volume	Vehicle in hour
AADT	Annual average daily traffic	Vehicle in hour
P.C.H	Traffic volume of passenger cars	passenger cars in hour
N.P.C.H	Traffic volume of non- passenger cars	Vehicle in hour
Av. Speed	Average speed at the time of accident	Km/h
Weekend	Days of week	Weekend 1, rest 0
Lic	Native , Non-Native license plate	Native 0, Non-Native 1
Light	Lighting conditions	Day 1, night 0
C.T.	Collision Type	One or more vehicle-overturning-with pedestrian- etc
WC	Weather conditions	Clear 1 etc 0
A.C.	Accident cause	Nine causes of accidents
GC	Geometric condition of accident location	P-straight\ P-intersection\ P-Residential\P- Slope\ P-horizontal curve
GPS	-	Global Positioning Systems of accident location

Table 5. Sample of road segmentation that sorted by accident frequency (this table sorted by 18 to 1 accident for 1 or 2 kilometer segments).

Segment	**kilometer**		**Accident frequency**
1	30	31	18
2	2	3	17
3	20	21	17
4	1	2	16
5	21	22	16
6	2	2	16

3.4.2. Time Series Modeling by ANN

3.4.2.1. Describing the data and statistical analyzes on the data

In this section, statistical operations were performed on the accident data and the correlation between the data was examined. A wavelet neural network integrated model was used to do this. The main objective of this section is to identify the most important factors affecting the number and severity of accidents using data mining. In this chapter, after analyzing the data in the form of charts and tables, the most important factors

were identified using MATLAB software. In this section, 2642 accident scenes involving a total of 2,297 people from 2015 to 2017 years are considered for this study. The data used in the study of driving accidents are generally of a discrete and qualitative type (accident type, collision type, vehicle type, lighting status, accident location, etc.). Some data such as driver's age, shoulder width, radius of the horizontal arcs and the longitudinal slope of the route, as well as the traffic volume and the average traffic speed and etc. can have small values. By examining the above-mentioned various items and components, the risk index was obtained and the graph was depicted as follows. 70% of the data was allocated to the training and evaluation of the network and 30% was allocated to the result. AI data was entered into the neural network of Black Box and was used to predict the model using 5 neural neurons. The coefficient of efficiency was obtained E=0.46 for the neural network method. The next method is the neural-wavelet network, with an efficiency coefficient of E =0.84, which indicates the superiority of this method over the previous method. The Wavelet neural network integrated model first filters the data and then enters the neural network (Fig. 5).

3.4.2.2. The criteria for choosing the best model

In order to choose the best model, there are various criteria in both discrete selection and data mining methods that can be applied. In this study, Log Likelihood, AIC and AICc criteria were used to select the best discrete selection model. On the other hand, the AUC, Accuracy, and Error Criteria have been used to select the best data mining model.

3.4.2.3. Modeling based on artificial neural network (ANN) method

At this stage, based on the data processing in the Excel software environment, input and output matrices were prepared for the data. As the number of accidents were considered as output and all the discrete and continuous variables, that the statistical analyzes were performed on them, were considered as the input matrix of the model.

Table 6. Indicators used to define hidden variables.

Hidden variable	Observable variable	Indicator
Human factor	Age	18-30
		30-45
		45-60
		60 and more
	Gender	Male
		Female
	Local (native)	Native
Geometric characteristics of the road	Maximum slope	|Maximum slope| = Max (slope^{+}, |slope^{-}|)
	Curvature	$CCR_{sec}\frac{\sum_{i=1}^{n}\|Yi\|}{L}$
	Slope and curvature concurrency	Maximum curvature x Maximum slope
Traffic characteristics	Average speed of vehicles	Average speed of vehicle in accident hours
	Heavy vehicles	Traffic of heavy vehicles / hourly
Accidents	Hourly rates of accidents	$R_{JH}=\frac{F_i \times 10^3}{HT}$
	Daily rates of accidents	$R_{JD}=\frac{F_i \times 10^3}{DT}$

Table 7. The sample of route data calculated according to tables (4 to 6) for two segments (30-31) and (2-3).

Road : Ardabil-Sarcham	From (km)	To (km)	Accident number	Hourly rates of accidents	Daily rates of accidents	Absolute value of maximum segment- slope	Maximum curvature	Native	Non native	Male	Female	Age group 18-30	Age group 30-45	Age group 45- 60	Age group more than 60	Heavy traffic rate at the accident hour	Average Daily Traffic
	30	31	18	11.92	0.512	6.07	15	10	8	17	1	3	11	2	2	0.27	35117
	2	3	17	9.82	0.422	0.5	0	9	8	15	2	4	9	3	1	0.20	40295

Table 8. The output of the Poisson model at the final stage (8) for AF.

```
Poisson regression                              Number of obs   =        385
                                                LR chi2(8)      =     499.31
                                                Prob > chi2     =     0.0000
Log likelihood =  -1206.734                     Pseudo R2       =     0.1714
```

AF	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
maxslope	10.33675	1.115896	9.26	0.000	8.149638	12.52387
Curvature	-.0494246	.0052047	-9.50	0.000	-.0596256	-.0392235
SectionLen~h	.3543231	.0426996	8.30	0.000	.2706335	.4380128
MaxSC	.960281	.1543305	6.22	0.000	.6577988	1.262763
Light	-.1661208	.0399315	-4.16	0.000	-.244385	-.0878566
singlemult~e	.1425403	.0396759	3.59	0.000	.064777	.2203036
PStraight	.170746	.058784	2.90	0.004	.0555314	.2859606
PED	-.1933975	.1004298	-1.93	0.054	-.3902363	.0034414
_cons	1.238712	.0930708	13.31	0.000	1.056296	1.421127

3.4.2.4. Sensitive analysis and validation of models

In addition to the tests for the neural network model, one of the most commonly used methods for evaluating models created for accidents is the comparison between model output and observed values. Typically, this comparison is performed in two stages, once using the data used in the modeling stage and again using a series of independent data not used in the modeling process. The results of the comparison of the first stage show the good fit of the model with the data used in modeling, while the results of the comparison of the second stage indicate the generalizability of the model results (Figs. 7 to 9). The value of R^2 obtained from the model fitting with the observed value and without considering the computational error according to Fig. 8 is equal to 0.61, indicating the appropriate fit of the model in the estimation of accidents. According to the above figure, the value of R^2 obtained from fitting with considering the computational error is equal to 0.97 (MSE=0.25, R^2 =0.97) which indicates the appropriate fitting of the proposed model using the proposed method in order to achieve the research objectives (Fig. 7).

Table 9. The output of the Poisson model at the final stage (11) for AS.

Poisson regression — Number of obs = 385; LR chi2(11) = 793.23; Prob > chi2 = 0.0000; Log likelihood = -1948.8507; Pseudo R2 = 0.1691

AS	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Curvature	.0291933	.0025306	11.54	0.000	.0242335	.0341531
PSlope	-.5880085	.0473349	-12.42	0.000	-.6807833	-.4952338
PResidential	.8863764	.0751831	11.79	0.000	.7390202	1.033733
PCH	.0022383	.0002407	9.30	0.000	.0017665	.0027101
Weekend	.3954599	.0377714	10.47	0.000	.3214293	.4694904
lic	.334926	.0371214	9.02	0.000	.2621694	.4076827
Overturn	-.2098296	.0373931	-5.61	0.000	-.2831186	-.1365405
PStraight	-.3670012	.0686491	-5.35	0.000	-.501551	-.2324514
maxslope	-5.14553	1.007254	-5.11	0.000	-7.119711	-3.171349
PHorizantal	.1674909	.0436667	3.84	0.000	.0819057	.253076
PED	.2199638	.0721572	3.05	0.002	.0785384	.3613892
_cons	1.809395	.0507271	35.67	0.000	1.709971	1.908818

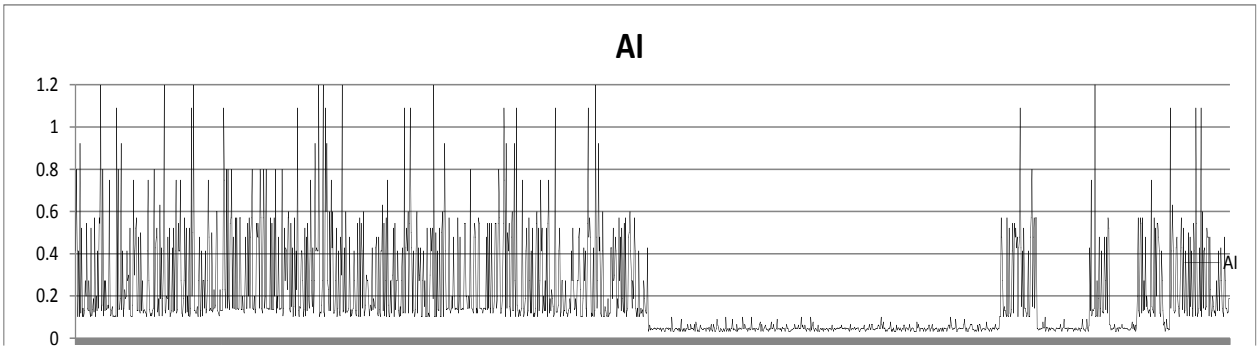


Fig. 5 The result of the accident severity index model based on traffic and geometric parameters.

For each accident record, all the columns of the record are independent variables that are related to a single segment. Therefore, in some segments of one and two kilometers the record is zero, and in some of them the number of records sometimes reaches ten. On the other hand, the criterion of being accident prone location according to the instructions of the Ministry of Roads and Urban Development in 2015 is as follows: Accident hotspot means an accident spot where the accident index number of that spot is greater than or equal to the numerical mean of this index at the network level of the country's roads. Furthermore, during a period of three years, those spots on which at least two fatal accidents, or three injury accidents, or one fatal accident, plus two injury accidents are occurred should be listed as accident hotspots.

Hence, in the model, some segments of the route that has at least two fatal accidents, or three injury accidents, or one fatal accident, plus two injury accidents, have been used in the risk model and other components are not included in the model.

In the following, various diagrams for training data, testing, the results of the study and the output of the neural network are described in Fig. 6.

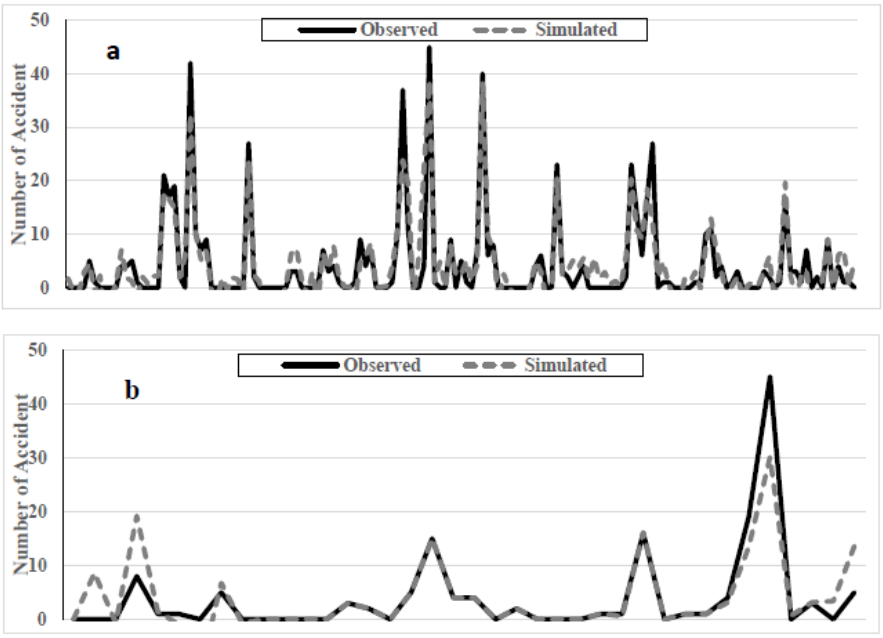


Fig. 6. The observed and estimated values of accident number in the neural network model for two stages of training and testing.

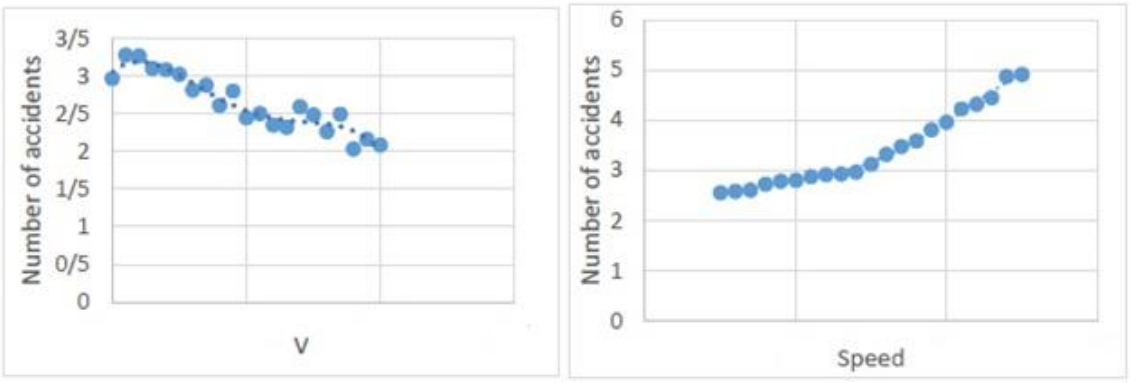


Fig. 7. The effect of volume and speed parameters on accident number.

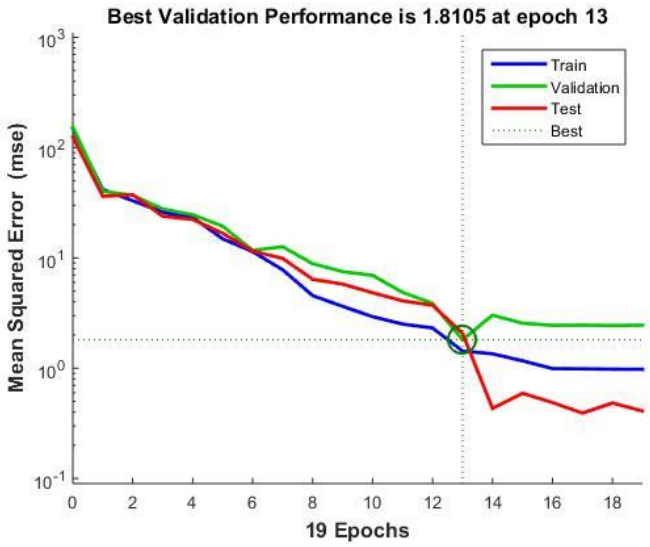


Fig. 8. The best validation performance in Lev. Model.

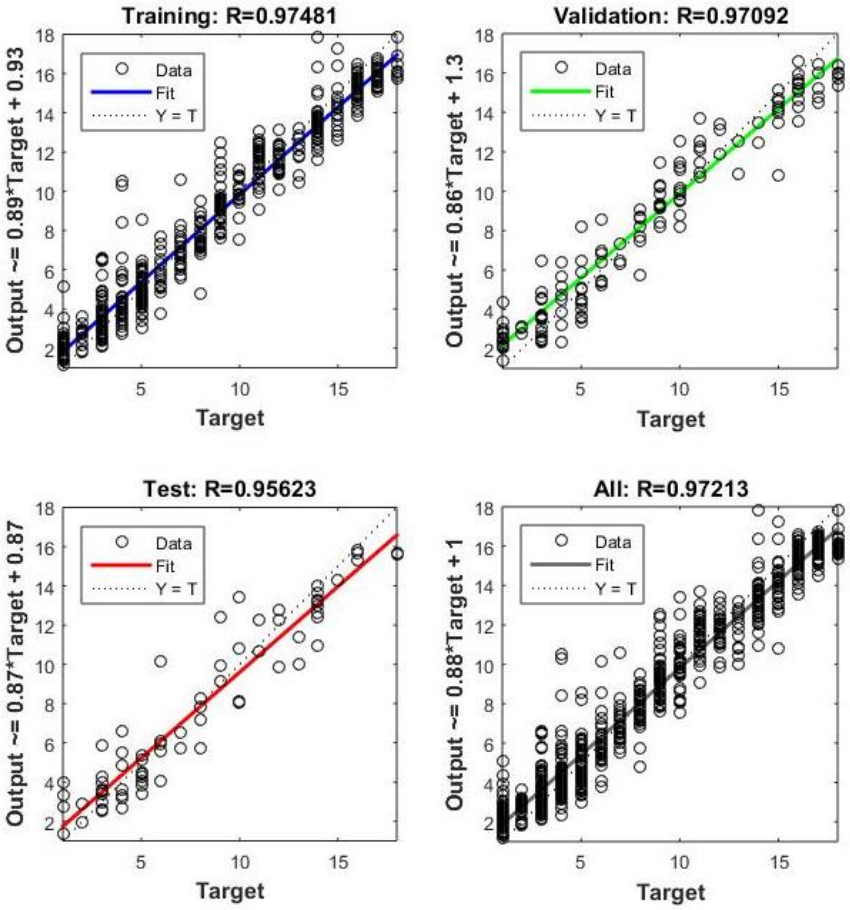


Fig. 9. Model fitting with the regression analyses for training, Validation and test stages.

4. Conclusion

In this chapter, two risk models based on the regression method and pattern recognition were presented to identify accident hotspots based on the severity and number of accidents and the dynamic segmentation method with a floating fixed length of 2 kilometers. To confirm the results of the models, the sensitivity analysis was used which indicated that the increase in accidents was accompanied by an increase in average speed and reduction of traffic volume of the route. Meanwhile, in the neural network method, validation of the model is performed at the testing stage, which the fit coefficient obtained confirms the appropriateness of the method for modeling.

In the floating segmentation method with a fixed length of 2 km, the highest percentage of accident coverage was obtained for segments with sequential and floating fixed lengths of 3 and 5. Therefore, the prioritization of the high-accident segments of the route by using this method to allocate the route improvement credit is more optimal than other methods and with a specific amount for improvement of the route, more percentage of accidents will be covered. In this research, were used the superior methods of Poisson regression and negative binomial distribution and based on the combined criteria of frequency and severity of accidents and equivalent injury factors. Then using Time Series Models in ANN, result were compared and validated. The results of artificial neural network models demonstrate that the frequency method of accidents tends toward places with high traffic volume and in addition, the severity of the accident is not considered in this method. MATLAB-2015a and STATA software were used. Non-native plumbing (lic), curvature, slope, section length and residential area had more significance, and their coefficients indicated the significant effect of these parameters on the occurrence of the frequency and severity of accidents in hotspot locations.

In a general view, this study can be classified as a road safety study that aims to determine the strengths and weaknesses of existing risk index models and propose a developed model to identify and prioritize high-

risk spots and segments of the suburban roads. As mentioned, a significant number of accidents are concentrated in specific spots of the road identified as accident hotspots. Obviously, with regard to credit constraints, the construction of new roads is not feasible and requires a great deal of cost and time; therefore, road safety management by improving the safety of these spots can be the most important action in reducing road accidents and road casualties with the greatest impact. Definite opportunities in this approach including preventive measures to improve safety through the improvement of hazardous situations in the existing network of roads with the aim of preventing accidents and reactive measures with the aim of correcting spots that are identified as accident hotspots can be effective in reducing accidents and damages resulting from it.

The results of artificial neural network models demonstrate that the frequency method of accidents tends toward places with high traffic volume and in addition, the severity of the accident is not considered in this method. Traffic and the nature of the accident are not considered in the method of Equivalent Property Damage Only Index, and the deviation tends towards high-speed locations in the suburban roadways. Therefore, considering both issues can provide more accurate results. However, in the proposed method, the accidental nature of accidents is considered and due to considering the nature of "return to mean in accident data," the accuracy of estimation is increased and in comparison with other methods, it is the most appropriate method for determining the accident hotspots of suburban roadways.

References

[1] A. Kazemi and H. Zoghi, "Identify and prioritize black spots on suburban roads and software," in *Eleventh Conference of Transportation Engineering and Traffic Iran, Transportation and Traffic Organization of Tehran*, Deputy transport and traffic of Tehran Municipality, Iran, 2011

[2] A. Montella, "A comparative analysis of hotspot identification methods," *Journal of Accident Analysis and Prevention*, vol. 42, pp. 571–581, 2010.

[3] G. Astaraki, A. Rassafi, F. Momeni, and B. Amini, "Application of multi-criteria decision in identified hotspots: Using data envelopment analysis and analysis of the Summit," *Transportation Engineering Journal*, 2013.

[4] R. Elvik, "State-of-the-art approaches to road accident black spot management and safety analysis of road networks," Report 883, Institute of Transport Economics, Norvegian Centre for Transport Research, 2007.

[5] A. Sadeghi and I. Ayati, "Identification and prioritization of accident-prone parts of the segmented approach path and data envelopment analysis," *Transportation Engineering*, 2012.

[6] K. Rahimov and D. Haj Ali, "Providing a model for identifying hotspots on the rural roads using multi-criteria decision-making," in *6Th National Congress of Civil Engineering, 6 and 7 May*, Semnan University, Semnan, Iran, 2011.

[7] F. Haghighi and E. Karimi Maskooni, "Evaluation and statistical validation of black-spots identification methods," *International Journal of Transportation Engineering (IJTE)*, vol. 6, no. 1, pp. 1-15, 2018.

[8] J. A. Deacon, C. V. Zegeer, and R. C. Dean, "Identification of hazardous rural highway locations," *Transportation Research Record*, vol. 543, pp. 16-33. 1975.

[9] J. I. Taylor and H. T. Thompson, "Identification Hazardous Locations," Report FHWARD-77-81, US Department of Transportation, 1977.

[10] D. R. D. McGuigan, "The use of relationship between road accidents and traffic flow in black spot identification," *Traffic Engineering and Control*, pp. 448-453, 1981.

[11] S. Y. Sohn, and H. Shin, "Pattern recognition for road traffic accident severity in Korea," *Ergonomics*, vol. 44, no.1, pp. 107-117, 2001.

[12] G. Karlafits Matthew and G. Silvestro, "Effects of road geometry and traffic volumes on rural roadway accident rates," *Journal of Accident Analysis and Prevention*, vol. 34, no. 3, pp. 357-65, 2002.

[13] W. Cheng and S. P. Washington "New criteria for evaluating methods of identifying hot spots," *Transportation Research Record*, vol. 2083, pp. 76-85, 2008.

[14] K. El-Basyouny and T. Sayed, "Collision prediction models using multivariate Poisson-lognormal regression," *Journal of Accident Analysis & Prevention*, vol. 41, no. 4, pp. 820-8, 2009.

[15] S. Cafiso, A. Di Graziano, G. La Cava, and B. Persaud, "Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables," *Journal of Accident Analysis & Prevention*, vol. 42, no. 4, pp. 1072-9, 2010.

[16] A. Shariat Mohaymani and A. Tavakoli Kashani, "Analysis of the severity of accident injuries on two-lane suburban roads using data mining models," *Journal of Transport*, vol. 7, no. 2, 2010.

[17] M. Keymanesh, H. Ziari, S. Roudini, and A. Nasrollahtabar Ahangar, "Identification and prioritization of black spots without using accident information," *Modelling and Simulation in Engineering*, vol. 2017, pp. 1-9, 2017.
[18] A. M. Boroujerdian, M. Saffarzadeh, H. Yousefi, and H. Ghassemian, "A model to identify high crash road segments with the dynamic segmentation method," *Journal of Accident Analysis and Prevention*, vol. 73, pp. 274–287, 2014
[19] M. Vosoughifar, R. Kiamehr, and A. Medqalchy, "Investigatetion of how to identify black spots using GIS, (Case study: Zanjan-Khoramdeh Road)," in *The First Regional Conference on Civil Engineering with Sustainable Development Approach*, Islamic Azad University of Bandar Gaz branch, Bandar Gaz, 2011.
[20] M. Saffarzadeh and A. Fakhro, "Assessment of techno-economic model for the construction and operation of freeways and expressways," *Journal of Transportation*, no. 1, 2006.
[21] H. Nassiri and M. Moshfeg Mojarrad, "Evaluation the risk index of accidents on country roads in Iran," vol. 22, no. 35 (Special of civil engineering), pp. 23-33, 2006.
[22] Y. Sajed and J. Azimi, "Evaluation of tunnel safety of rural roads by AHP method (case study: tunnels of Ardabil province roads)," M.Sc. Thesis, Islamic Azad university of Ahar branch, Ahar City, Iran, 2016.
[23] D. Lord and L. F. Miranda-Moreno, "Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective," *Safety Science*, vol. 46, no. 5, pp. 751-770, 2008.
[24] E. Hauer, "Over ispersion in modelling accidents on road sections and in empirical Bayes estimation," *Journal of Accident Analysis & Prevention*, vol. 33, no. 6, pp. 799-808, 2001.