ENGINEERING JOURNAL

Article

New Achievement in the Prediction of Highway Accidents

Gholamali Shafabakhsh[a,*] and Yousef Sajed[b]

Faculty of Civil Engineering, Semnan University, University Sq., P.O. Box 35196-45399, Semnan, Iran
E-mail: [a]shafabakhsh@semnan.ac.ir (Corresponding author), [b]yousef.sajed@sun.semnan.ac.ir

Abstract. Most research has been carried out about crash modeling but there is little attention to the urban highways. The candidate's set of explanatory parameters were: traffic flow parameters, geometric infrastructure characteristics and pavement conditions. Statistical analysis is done by SPSS on the basis of nonlinear regression modeling and during the analysis, principal components are identified to assist the principal component analysis method and more important variables recognized that could indicate the best description of crash occurrence on the basis of available logics. Results indicate that the number of accidents per year increase with: length, peak hour volume and longitudinal slope whereas it decreases with radius. The presented models show that the crashes occurrence increase with the increase in each of section length, peak hour volume and longitudinal slope variables whereas it is decreased with the decrease of curvature. The remarkable result in this study was the effect of longitudinal slope variable on the crash occurrence.

Keywords: Highway accidents, regression models, section length, curvature, longitudinal slope, peak hour volume.

ENGINEERING JOURNAL Volume 19 Issue 1
Received 24 April 2014
Accepted 15 September 2014
Published 30 January 2015
Online at http://www.engj.org/
DOI:10.4186/ej.2015.19.1.139

DOI:10.4186/ej.2015.19.1.139

1. Introduction

In recent years, a lot of studies have been conducted about accidents, condition of their occurrence and also the prediction of occurred crashes counts. In these studies, the effects of different effective variables on crashes have been considered in order to prevent or reduce the crashes counts.

The first prediction's model of multilane roads was presented by Persaud and Dezbik in 1993. In this research, the generalized linear models, average daily traffic (ADT) data and traffic volume per hour (VH) were used. The results of these studies show that the rate of crash is increased by the growth of the traffic flow introduced by ADT and VH. Also, accident's hazard is more in 4 lane freeways than over 4 lane freeways at the same situation such as traffic volume, free flow and much maneuver power of road users. Thus, the traffic flow introduced by VH is more suitable than ADT for describing crash phenomenon in congested or free flow traffic because of the difficulty of accurately estimating of VH. Therefore, ADT is usually applied in prediction models [1].

Knuiman (1993) examined the effect of median width of four-lane roads on crash rate using a Negative Binomial Distribution. His model was on the basis of the Log-Linear regression method. The findings indicated that crash rate decreases with increasing median width. Furthermore, wider medians considerably reduce "crossover accidents" involving head-on crashes between opposing vehicles. As a result, a much greater positive effect of medians on severe crashes than on property-damage-only crashes is expected [2].

Hadi (1995) proposed several accident-prediction models with regarding both to multilane roads and two-lane roads of rural or urban designation. The dependent variables were total crash rate or injury crash rate by consideration of Poisson and Negative Binomial regression models. The values of these accident indicators were estimated as a function of AADT and road environmental factors. By examining the effect of traffic flow on the crash rate the conclusions reached were that crash rate increases with increasing AADT on roads having higher levels of traffic, while it decreases with AADT on roads with lower traffic volumes [3].

Milton and Mannering (1996) implemented their studies on the basis of negative binomial regression model. Results presented that the crash occurrence is less in sections with median even with more traffic. Also, they represented that the rate of crash is reduced in the middle sections of road and in different speed limitations, obviously. Thus, it must not increase the cross speed limitation unless geometric infrastructure characteristics, signs, right side guard of road and other limitations would be improved [4].

Golob and Recker (2003) used the linear and non-linear multivariate statistical analyses to determine how the type of accidents is related to traffic flow, weather and lighting conditions. The study approach was based on the Principal Component Analysis (PCA) in order to identify the most significant variables from a set of original traffic flow variables, and a canonical correlation analysis (CCA) was used to relate the identified principal components both to weather and lighting conditions. Variables used consisted of: a) lighting variables (classified sections based on lighting and weather condition), b) crash variables (type of contact, location of crash occurrence and crash severity), c) traffic flow variables (median, middle lane changes, right lane changes, average volume of left lane, volume changes of left lane and right lane) [5].

Golob (2004) evaluated the safety effects of changes in freeway traffic flow, in a subsequent paper. In this study, three crash characteristics were used in the analysis, namely crash type (rear end, sideswipe or hit object, number of vehicles involved); crash location (e.g. left lane, interior lanes, right lane, shoulder); crash severity (injuries and fatalities per vehicle) [6]. Also Hauer (2004) developed statistical road safety modeling by using the Negative Binomial distribution. The dependent variable was the number of accident per year, while the independent ones were geometric characteristics and traffic flow. In this study, the author suggested guidelines for assigning the functional form to each variable in the model, and observed that the model equation should have both a multiplicative and an additive component. The multiplicative component is to account for the influence of variables that have a continuous role along a road (such as lane width or shoulder type), while the additive component is to account for the presence of hazardous points (such as driveways or narrow bridges). The most innovative aspect of this study was the introduction of an alternative tool for measuring the goodness-of-fit of the predictive models, the so-called Cumulative Residuals (CURE) Method. This method consists of plotting the cumulative residuals as a function of the independent variable of interest, a good CURE plot being one oscillating around zero [7].

Caliendo and Parisi (2004) implemented a comprehensive study about the PCA method and selecting the effective principal variables in crash occurrences. The purpose of this study was the deletion of unnecessary variables in the crashes count prediction models for the exhibition of desirable and communicative model. They advised this method to other researchers because of its simple and

uncomplicated trait in different to Fuzzy logic, etc. The results for curves presented that crashes are increased by increase in section length (L), curvature (1/R), average daily traffic (ADT) and the difference in design speed between tangents and curves. Also for tangents, crashes are increased by increase in section length (L) and average daily traffic (ADT) [8].

Caliendo, Guida and Parisi (2007) discussed in a new paper about variables related to the traffic flow, geometric infrastructure characteristics, pavement surface and rainy condition. In this paper, the crash prediction models were presented for four-lane median-divided rural roads in Italy. The Poisson, Negative Binomial and Negative Multinomial regression models, applied separately to tangents and curves were used to model the frequency of accident occurrence [9].

Anastasopoulos, Tarko and Mannering (2008) applied Tobit Regression for recognition of principal parameters that affect accidents. They withdrew the previous methods such as negative binomial or multinomial methods, etc and exhibited their research with the new above method. The results of this research showed that the factors related to the status and quality of the pavement (side friction coefficient, rough index of pavement, rut of pavement, pavement's degree) and geometric infrastructure characteristics (type and width of median, width of shoulder, counts of ramps, bridges, horizontal and vertical curves) affect crashes counts [10].

The above studies show that there are several variables that affect accidents. In these studies, the attention points of researchers have concentrated on rural roads and there is little attention to the accidents of urban highways. This paper examines the accidents of urban arterial highways in Tehran and it is set up on the basis of the traffic flow parameters and effective geometric characteristics that affect accident occurrence in urban highways. Identifying the most important variables that increase crash counts in the urban arterial highways is the aim of this paper until after identifying each variable on the basis of available data and their effects, the costs of accidents will be reduced. Examining the above studies shows that more important parameters of crash prediction models are the traffic and geometric design parameters. According to the experience of these studies, researchers tried to indicate simpler models with higher probability ratio and they applied various linear and nonlinear models to present logical models with more effective variables on crash occurrence to reach this aim. The aim is identifying more effective variables on crash occurrence of urban highways.

2. Methodology and Materials

In this paper, nonlinear models are applied for modeling because of their more flexibility in adapting estimations on collected data. So the logistic analysis was done on the accident data observed of the three urban arterial highways in Tehran during a 3-year monitoring period extending between 2005 and 2008 until more important causes of crash occurrences and their effects on crashes were identified on such highways.

Crashes statistics were collected from the three urban arterial highways in Tehran (4 or 5-lane median-divided) and their characteristics are shown in Table 1. During the period of observation, crash data, traffic flow and geometric characteristics were collected.

Accidents data were collected and extracted from the official reports of the Tehran's Driving Police (Rahvar) and Tehran's Highway Police. For each accident a variety of details was recorded, including location of accident, horizontal alignment (tangent or curve), vertical alignment (upgrade or downgrade), weather and pavement surface conditions (dry or wet), type and severity of accidents.

Table 1. Characteristics of under study highways.

Highway Name	Curve Sections		Tangent Sections		Total Length (m)
	Total Length(m)	Count	Total Length	Count	
Chamran	5198	22	6727	19	11925
Navab	1620	9	5367	19	6987
Saeidi	376	3	921	4	1297

Some 12,891 accidents were considered in this study, 32 of which were fatal, 1348 were injury accidents and 11,601 were property damage only accidents. So in summary, during the period of observation 1380 casualty accidents (fatal and injury crashes) equal to 10.63% of all crashes and 11,601 property damage only

accidents or 89.37% of all crashes were registered. 34.40% of all crashes occurred on curves and 65.60% on tangents or straight sections. So the chance of crashes occurrence at tangents is 1.91 higher than crashes that occurred at curve sections. Table 2 illustrates the accident count data observed during the 3-year monitoring period.

Table 2. Crash count of under study highways separately for curves and tangents.

Highway name	Year	Occurred crash count						
		North-South roadway			South-North roadway			Total highway crash
		Tangent	Curve	Total	Tangent	Curve	Total	
Chamran	2005 2006	941	486	1427	946	505	1451	2878
	2006 2007	1032	553	1585	1061	561	1622	3207
	2007 2008	472	237	709	489	269	758	1467
Navab	2005 2006	471	246	717	469	243	712	1429
	2006 2007	514	262	776	512	264	776	1552
	2007 2008	464	194	658	465	279	744	1402
Saeidi	2005 2006	70	41	111	73	39	112	223
	2006 2007	126	66	192	124	66	190	382
	2007 2008	114	61	175	114	62	176	351
Total		4204	2146	6350	4253	2288	6541	12891

The database does not include accidents taking place on the ramps of junctions, on service areas, at tollbooths or on shoulders, since such accidents are not due to traffic flow and geometric design characteristics. Pedestrians and bicycles are forbidden to use this infrastructure. Thus, no pedestrian and bicycle were involved in accidents.

For the purpose of the subsequent statistical analysis, homogeneous road sections were separated. The key characteristics of these three highways were collected form the Traffic and Transportation Organization of Tehran's municipality. These characteristics include longitudinal slope, different segments of curves and tangents, length and width or curvature of each segment, number of lanes, location of junctions and the like. Segments for each carriageway have constant horizontal curvature and longitudinal slope. For these segments the following major variables did not change: width and number of lanes, type and width of shoulders, median width and type. Similar segments of highway are separated by this method.

After investigating previous studies are shown in introduction and identifying the effect amount and importance of most variables, in this research among the different available parameters, we attended to traffic flow parameters and horizontal and vertical dimension characteristic parameters. For example, many complicated calculations and particular statistics are necessary to examine the effect of drivers caution on crash occurrence that it is not included in modeling easily. So a variable such as present junctions count in the path are not included because of low junctions count in the three involved highways. The wet effect variable is not included in modeling because it is not effective on increasing and decreasing crashes counts. If all other conditions are the same, the effect of moisture is to lower the surface skid resistance, so at high speed, and if the pavement surface is already polished, the likelihood of crash occurrence would be higher in wet condition. "Figure 1" shows view of Chamran Highway (for example).

Fig. 1. View of Chamran highway.

2.1. Traffic Flow Parameter

The initial studied parameter in modeling is related to the traffic flow. In this study, peak hour traffic volume (PHV) is used for entering the effect of traffic flow on crash occurrence. The traffic flow was extracted from the traffic file of the Traffic and Transportation Organization of Tehran's municipality for each of the three involved highways, in input and output points and in different sections. The amount of PHV is measured in each divided section of the three highways on the basis of these explanatory statistics of peak hour traffic volume and involved in the analysis.

In many sections, whereas traffic flow volume evaluating was difficult and it was necessary to control the accuracy of vehicles traffic on a highway; it was assumed that the total number of vehicles leaving a carriageway in a given exit point was equal to the number of vehicles entering the other carriageway through the corresponding controlled access point in the opposite direction. Table 3 illustrates the maximum and minimum registered PHV parameter statistics in each of the three highways.

Table 3. The amount of peak hour volume (PHV) of under study highways.

Highway Name	Minimum PHV (vehicles/hour)	Maximum PHV (vehicles/hour)
Chamran	2506	7562
Navab	1885	5190
Saeidi	3416	4823

2.2. Geometric Design Characteristics Parameters

Geometric design characteristics of each of the three involved highways are extracted on the basis of final data that had been collected by Traffic and Transportation Organization of Tehran's municipality. The plans of each of the three involved highways are supplied by Autodesk Inc. Auto Cad 2009 software and the analysis is done for parameters and necessary characteristics. These geometric characteristics that are presented in separate horizontal and vertical dimension characteristics were: tangent length, curve length, curvature and longitudinal slope.

All curves are simple circular form in horizontal alignment and whereas curves are adapted to simple curve forms, all curves in geometric characteristic's plan are defined as simple curves. In vertical alignment, because of the large noticeable differences between each couple consecutive longitudinal slope and the choice of condition of longitudinal slope, vertical curves are ignored.

2.3. Modeling

The aim of Crash prediction models presentation is to identify the more important variables that must be entered in the model by examining the descriptive variables of crash occurrence till the best description of crash is expressed. With these models, the effectiveness amount of each effective principal variable in crash

occurrence can be determined and the risk of crash occurrence can be reduced with decreasing negative effects. The importance of data analysis tools will be identified by this introduction.

In this study, SPSS software was applied to data analysis and necessary outcomes presentation. SPSS software is one of the earlier application software in statistical analysis. It is one of the professional statistic software and proceed to statistic subject about social sciences, psychology and behavioral sciences, etc. so it is a comprehensive and flexible analyzer and a data management system that is able to get data from any file types and uses them to produce applied reports and complicated statistical analysis of composite behavior of data. Before data analysis being cleared which variables have a more explanatory ability in different to the other variables. On the other hand, the required regression model should be identified in order to modeling. These cases are further described in section (2.3.2).

2.3.1. Selecting principal variables

In this study, the principal variables of modeling are extracted according to the PCA method. Thus, it is possible to select the principal variables of model by this method because of avoiding complicated methods such as sever arithmetic logics for example fuzzy logic and such. There are consecutive steps in this method that the principal variables leading to crash can be obtained after carrying out these steps.

A 3-year monitoring period extending from 2005 to 2008 was perused on the three urban arterial highways. The geometric design of these three highways is assumed as a set of tangent and simple curve (without link curve) segments and the other necessary data were collected in addition to the crash data and geometric design characteristics during the study period. Crash data are categorized separately in similar sets on the basis of the similarity of tangent and curve segments for each direction after collecting.

By categorizing data, the final variables are: Y (occurred crash count), L (curve section length whether tangent), 1/R (curvature), n (number if lane). Y as occurred crash count in each section is identified as a dependent variable on the basis of the other independent variables.

The collected data must be standardized. So, the Z standard matrix for the primary variables is introduced.

$$Z_{ij} = (x_{ij} - x_j) / S_j;\ i=1, 2,\ldots, n\ \&\ j=1,2,\ldots,q \quad (1)$$

where x_j and S_j are mean and standard deviation of the generic variable of x_{ij}, respectively. The elements in Z matrix have zero mean and 1 variance. The covariance between two variables z_k and z_j (for $k \neq j = 1$, 2,..., n) is the correlation matrix.

Next the covariance matrix (R) was computed, which contains the correlations between the primary variables:

$$R = 1/n\ Z'Z \quad (2)$$

where Z' is given by the transpose matrix of Z.

From the so-called characteristic equation of the covariance matrix:

$$\text{Det}\ (R - \lambda_h I) = 0 \quad (3)$$

where I is the identity matrix containing unit values, the Eigen values λ_h were calculated.

The variances accounted for the principal variables were computed with the Eigen values λ_h. The Eigen vectors v_h were calculated by means of the metrical equation:

$$(R - \lambda_h I)\ v_h = 0 \quad (4)$$

In addition, the S matrix was determined, which contains the correlation coefficients between the axes of the primary variables and the principal variables:

$$S = 1/n\ Z'\ Y\ L^{-1/2} = 1/n\ Z'\ Z\ V\ L^{-1/2} = RVL^{-1/2} \quad (5)$$

where V is the matrix of the Eigen vectors and L is the diagonal matrix whose elements are the Eigen values. Next the principal components were rotated by the Quartimax method in order to assist in interpreting the results.

This method permits the maximizing of the sum of the fourth power of elements contained in U, which is:

$$U = S.T \quad (6)$$

where T is the rotation matrix.

The multiple correlation coefficients ρ^2 between the axes of primary variables and rotated principal variables were computed as the addition of the square of the elements for each row contained in U. Higher ρ^2 are associated with the most significant original variables.

3. Results

3.1. Preliminary Results

Using the PCA method, the results obtained were plotted as correlation circle diagram. The correlation's deal of each variable is identified by these diagrams. When the points representing the original variables are closer to the principal axes as well as the circumference a high correlation is shown. Furthermore, if the original variables are closer to each other a close correlation between these variables is revealed.

Rotated parameters matrix is shown in Table 4.

Table 4. Rotated components matrix in tangents.

	Component				ρ^2
	1	2	3	4	
ACC	0.688	0.523	0.262	0.031	0.816
L	0.024	0.961	-0.044	0.059	0.930
PHV	0.915	-0.087	-0.171	0.130	0.891
L.S	-0.049	-0.003	0.982	-0.018	0.967
n	0.147	0.071	-0.018	0.986	0.999

U matrix exhibits ρ^2 values as shown in Table 4 that it is possible to consequence the principal variables on the basis of its results. Also Table 5 as score matrix of variables gives a noticeable help to principal variable choice. The values of ρ^2 are between 0.816 up to 0.999 on the basis of the multiple correlation coefficients as shown in Table 4. Minimum effect on crashes is applied by PHV. Since there is a little difference between ρ^2 for this variable and the other variables, it is withdrawn the omission. Then variable stays with minimum change in rotation process. This variable is omitted to remove its destructive effect, because of the great correlation between this variable and the other variables that it was shown previously.

Table 5. Score coefficients matrix of tangents components.

	Component				ρ^2
	1	2	3	4	
ACC	0.688	0.523	0.262	0.031	0.816
L	0.024	0.961	-0.044	0.059	0.930
PHV	0.915	0-.087	-0.171	0.130	0.891
L.S	-0.049	-0.003	0.982	-0.018	0.967
n	0.147	0.071	-0.018	0.986	0.999

The parameters score matrix as shown in Table 5 illustrates that L.s variable has the best correlation with crash counts while it has the least correlation with the other variables. The dependence of L.s variable to crash occurrence is 93.3%.

Figure 2 is presented on the basis of the two first parameters of U matrix which represents the correlation's amount of each variable to the axes and to each other. This diagram specifies which variables should stay in the modeling.

Consequences show that the crash count variable (ACC) is almost independent of first and second axis. PHV and L variables have maximum dependence to the first and second axes respectively. While L.s and n variables show dependence to second axis slightly. Consequently, the best correlation with crash counts is belonged to L.s that it is confirmed in previous ratiocination. In Tables 6 and 7 are presented respectively, rotated parameters matrix and score coefficients matrix in curves.

DOI:10.4186/ej.2015.19.1.139

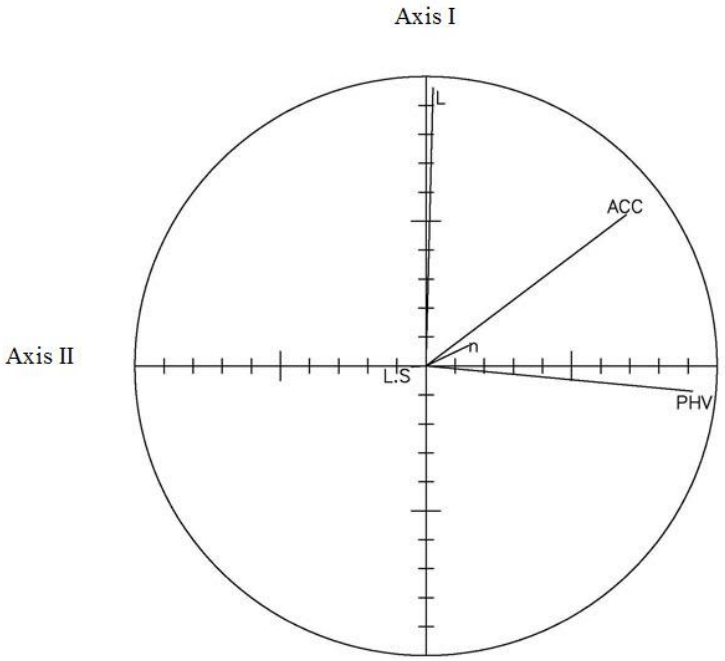

Fig. 2. The correlation circle diagram of tangent components.

U matrix exhibits ρ^2 values as shown in Table 6 that it is possible to consequence the principal variables on the basis of its results. Also, Table 7 as score matrix of variables gives a noticeable help to principal variable choice. The values of ρ^2 are between 0.813 up to 0.992 on the basis of multiple correlation coefficients as shown in Table 6. Minimum effect on crashes is applied by 1/R. Since 1/R is the best introducer for curves, it is withdrawn from its omission. The n variable stays with minimum change in rotation process. This variable is omitted to remove its destructive effect, because of great correlation between this variable and the other variables which is shown previously.

Table 6. Rotated components matrix in curves.

	Component					ρ^2
	1	2	3	4	5	
ACC	0.876	-0.151	0.169	0.185	0.086	0.860
L	0.035	0.024	0.982	-0.010	0.108	0.978
1/R	0.747	0.436	-0.191	-0.109	-0.127	0.813
PHV	0.047	0.962	0.042	-0.022	0.056	0.933
L.S	0.094	-0.035	-0.009	0.988	-0.012	0.986
n	0.000	0.043	0.110	-0.012	0.989	0.992

The parameters score matrix as shown in Table 7 illustrates that L.s variable has the best correlation with crash counts while it has the least correlation with the other variables. The dependence of L.s variable to crash occurrence is 93.2%.

Figure 3 is presented on the basis of the two first parameters of U matrix which explains the correlation's amount of each variable to the axes and to each other. This diagram identifies which variables should stay in modeling.

Consequences show that the crash count variable (ACC) is tended to the second axis. The L.s variable is slightly dependent to this axis. The PHV variable is dependent to the first axis while n variables show dependence to this axis slightly. The L and 1/R are almost independent from two axes. Also there is dependence between L and 1/R variables.

Table 7. Score coefficients matrix of curves components.

	Component				
	1	**2**	**3**	**4**	**5**
ACC	0.697	-0.272	0.105	0.024	0.076
L	-0.012	0.058	0.980	-0.030	-0.142
1/R	0.539	0.248	-0.171	-0.160	-0.081
PHV	-0.138	0.886	0.074	0.114	0.007
L.S	-0.074	0.099	-0.029	0.992	-0.012
n	0.017	0.001	-0.142	-0.014	1.009

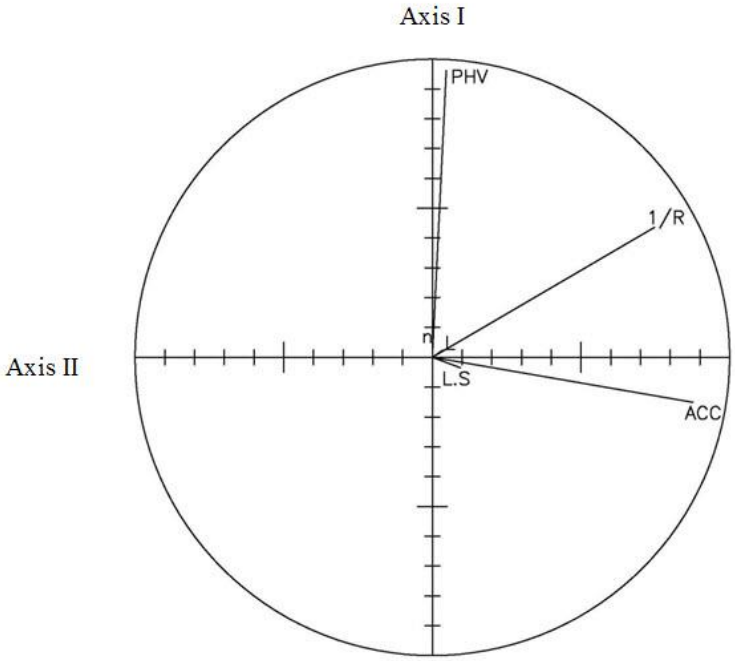

Fig. 3. The correlation circle diagram of curve components.

The remarkable result is that the L.s variable has the most effect on the occurred crash counts at tangents and curves among the existing explanatory variables. There are several reasons to affect crash occurrence by longitudinal slope:

1. The maximum amount of longitudinal slope and length of a section are effective on the amount of vehicles speed.
2. Controlling of vehicles is more severe in steeper slopes because of the longitudinal friction coefficient and pavement condition and consequently the risk of crash occurrence is increased.
3. The sight distance in convex vertical curves (at collision location of two consecutive longitudinal slopes) is decreased and consequently the risk of crash occurrence is increased.

On the basis of discussed ratiocinations, the effective and included variables in urban arterial highways accidents modeling were specified in each tangent and curve sections. So the parameters of crash prediction model for tangent sections are section length (L), peak hour volume of traffic (PHV) and longitudinal slope (L.s). Also, the parameters of the crash prediction model for curve sections are section length (L), peak hour volume of traffic (PHV) and longitudinal slope (L.s) and curvature (1/R).

3.1.1. Selecting the regression model

The next significant step in data modeling is selecting one of the linear or nonlinear regression models to present an adapted model for collected data that it can presents best adapted estimates on statistic. In the linear regression models, used approximations are coefficients of a linear equation which contains one or

more independent variables which express the best estimated value of independent variable. Also, for non-linear regression, coefficients of a nonlinear equation such as exponential, power, logarithmic and etc. is presented which show a good approximation of nonlinear variable.

In this research, after testing the linear regression models, we did not discover good consequences. Therefore, on the basis of nonlinear models which lead to better results, the crash counts was modeled. Also, nonlinear regression models were used in modeling test and the final model was selected as a combination of some models.

The other noticeable point in regression model is the distribution type which data are matched in them. For data distribution, there are many data distribution method which can use the best one referring to this point that used data are matched in one of these methods. Some of these usual mathematical distributions are Normal distribution, Gamma distribution, Binomial distribution, Multinomial distribution, Poisson distribution and other like them.

For recognizing the used data matched in such type of distribution, we can specify the type of distribution by Q-Q diagram in SPSS software which expresses the type of data distribution.

3.1.2. Presentation of prediction crash model

The registered crashes data during the 3-year monitoring period extending from 2005 to 2008 was perused on the three urban arterial highways and was analyzed for presentation of prediction crash model. These statistics contain curves and tangents accident data separately. According to the last parts, modeling is carried out for curves and tangents separately.

3.2. Model Results

3.2.1. Crash prediction model for tangents

Tangents accidents data in 42 parts of tangents (84 parts in couple directions of roadway) are considered in the three urban arterial highways with 13,015 meters total length. With regard to the conducted analysis, the principal variables of the model are section length (L), Peak hour volume of traffic (PHV) and longitudinal slope (L.s). A lot of nonlinear regression models are applied to present the final model, but the model used for crash count estimation of tangent parts in discussed highways has the best result.

$$ACC = x_0 \{ Exp[x_1 Ln(L)] + Exp(x_2 L.s) + Exp (x_3 PHV^{x4}) + x_5 \} \quad (7)$$

Data analysis for presentation of model coefficients is done with SPSS software in 25 steps. The results of analysis show that we have more accurate estimations in comparison with other models. The value of R^2 coefficient was increased up to 37.8% which is more accurate in comparison with the previous models. The results of data analysis of crashes at tangent parts are shown in Table 8 that express the estimation of selection regression coefficient data.

Then, for using model coefficients with more certainty, we have to reduce the value of R^2. In this research, for approaching better values in regression coefficients in comparison with the last step coefficients which shown in Table 9, this restriction has been reduced by Min Residual Method.

The regression model of tangent accident prediction is obtained by Min Residual Method, after 50 steps processing by software, with reducing the value of residues:

$$ACC = 44.927\{Exp [0.347 Ln(L)] + Exp (0.610 L.s) + Exp (1.910 PHV^{0.333}) - 22.391\} \quad (8)$$

For example, the crash amount in a tangent section is up to 35 accident in a year with L=0.490Km length, L.s=0.60 longitudinal slope and PHV=$3.898\times10^{+3}$ veh/h Peak hour volume of traffic.

The described model for crashes count prediction shows that accident occurrence is increased at tangent sections of the urban arterial highways with regard to the value and the sign of coefficients of the model variables with increasing in the section length (L), Peak hour volume of traffic (PHV) and longitudinal slope (L.s) parameters.

Table 8. The estimated coefficients of nonlinear model of tangent crashes.

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
x_0	57.197	59.066	-59.144	173.537
x_1	0.232	0.341	-0.440	0.904
x_2	0.547	0.044	0.461	0.633
x_3	2.105	0.938	0.258	3.951
x_4	0.288	0.189	-0.085	0.661
x_5	-35.974	64.793	-163.593	91.646

Table 9. Primary coefficients of tangents regression.

Parameter	Estimate
x_0	57.197
x_1	0.232
x_2	0.547
x_3	2.105
x_4	0.288
x_5	-35.974

3.2.2. Crash prediction model for curves

Curve accidents data in 34 parts of tangents (68 parts in couple directions of roadway) are considered in the three urban arterial highways with 7194 meters total length. With regard to the analysis, the principal variables of the model are section length (L), Peak hour volume of traffic (PHV), longitudinal slope (L.s) and curvature (1/R).

A lot of nonlinear regression models such as curves analysis section were used to present the final model, but the model used for crash count estimation of curve parts in discussed highways has the best result.

$$ACC = x_0 [Ln (x_1 L) + x_2 L.s + x_3 L.s^2 + x_4 (1/R) + x_5 PHV] \quad (9)$$

Data analysis for presentation of model coefficients is done with SPSS software in 7 steps. The results of analysis express that we have more accurate estimations in comparison with other models. The value of R^2 coefficient was increased up to 38.2% which is more accurate comparing with the last values.

The results of data analysis of crashes in curve parts are shown in Table 10 that express the estimation of the selection regression coefficient data.

Table 10. The estimated coefficients of nonlinear model of curve crashes.

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
x_0	5.810	0.956	3.925	7.695
x_1	422.417	305.375	-179.807	1024.640
x_2	0.685	0.182	0.325	1.044
x_3	-0.516	0.072	-0.659	-0.373
x_4	4.367	0.564	3.255	5.478
x_5	-1.224	0.426	-2.064	-0.383

Then, for using model coefficients with more certainty, we have to reduce the value of R^2. In this section like last section, for approaching better values in regression coefficients in comparison with the last step coefficients which shown in Table 11, this restriction has been reduced by Min Residual Method.

By Min Residual Method, after 50 step processing by software, with reducing the value of residues, the regression model of curve accident prediction is obtained:

$$ACC = -1.333[\ Ln(\ 5.949\times10^{-6}\ L\ ) + 0.972\ L.s - 0.239\ L.s^2 + 3.967\ (1/R) - 0.233\ PHV] \quad (10)$$

For example, the crash amount at a curve section is up to 23 accident in a year with L=0.139 Km length, L.s=0.59 longitudinal slope, 1/R=1.61 curvature and PHV=$3.898\times10^{+3}$ veh/h Peak hour volume of traffic.

The described model for the crash count prediction at curve sections of urban arterial highways show that accident incident is increased referred to the value and the sign of coefficients of model variables with increasing in the section length (L), Peak hour volume of traffic (PHV) and longitudinal slope (L.s) parameters and decreasing in the curve radius (R) parameter.

Table 11. Primary coefficients of curves regression.

Parameter	Estimate
x_0	5.810
x_1	422.417
x_2	0.685
x_3	-0.516
x_4	4.367
x_5	-1.224

4. Conclusions

In this study, the accidents which are occurred in urban arterial highways of Tehran are studied for the first time until a logical and trust worthy model of occurred crashes behavior in these highways is presented. The purpose was to consider the different effective factors in crash occurrence such as traffic flow and geometric design parameters of these highways in the development of a model which improves road safety and reduces damages through the study of parameters effectiveness. Necessity of distribution of this research is to increase safety on urban highways by identifying the important causes of accident. Level of each variable's effectiveness can be determined in terms of the reduction in the number of crashes and their cost by using crash prediction models.

In this research the crash-prediction models for the three urban arterial highways in Tehran were set up on the basis of accident data observed during a 3-year monitoring period extending between 2005 and 2008. These statistics contain separate data of crash occurrences for each tangent and curve sections of highways. For each accident, a variety of details was recorded, including location of accident, horizontal alignment (tangent or curve), vertical alignment (upgrade or downgrade), weather and pavement surface conditions (dry or wet), type and severity of accidents. Finally, the parameters which are screened to study accidents quantitatively are: crash occurred count (ACC), section length (L), peak hour volume (PHV), longitudinal slope (L.s), and curvature (1/R).

Some 12,891 accidents were considered in this study, 32 of which were fatal, 1348 were injury accidents and 11,601 were property damage only accidents. So in summary, during the period of observation 1380 casualty accidents (fatal and injury crashes) equal to 10.63% of all crashes and 11,601 property damage only accidents or 89.37% of all crashes were registered. 34.4% of all crashes occurred on curves and 65.6% on tangents or straight sections. So the chance of crashes occurrence at tangents is 1.91 higher than crashes that occurred at curve sections.

The effective principal variables in crash occurrence are selected with the PCA method. The variables are considered which are more independent and effective and lesser correlated in comparison with the other variables. Final analyses show that the principal variables of tangent crashes are: section length, peak hour volume and longitudinal slope, whereas for crashes at curve section they are section length, peak hour volume and longitudinal slope and curvature.

After screening the principal variables, data analyzed by SPSS software with nonlinear regression models and final crash prediction model is presented for each tangent and curve sections separately.

Also R^2 values are observed at 37.8% and 38.2% respectively at tangent and curve sections that they are suitable and comparable with many similar models. Min residual method is applied to decrease the R^2 values and better assurance to have a suitable model.

Presented models show that crash occurrence is increased with the increase in each of section length, peak hour volume and longitudinal slop variables whereas it is increased with the decrease of curvature.

The remarkable result in this study was the effect of longitudinal slope variable on the crash rate. This variable has the least dependence to the other variables and also the best score to describe the crash occurrence. With simple explanation, L.s has the best effectiveness in the crash occurrence of urban arterial highway. The first reason that L.s has the most effect on the occurred crash count is the relation between speed variation and longitudinal slope. Obviously, the traffic congestion is changed by speed variation that is a clear reason for the relation between longitudinal slope, decreasing safety and consequently decreasing the chance of crash occurrences. The second reason is that the effect of longitudinal friction amount and pavement condition in steeper longitudinal slope is increased and the chance of crash occurrence is increased too. The third reason is that the sight distance in convex vertical curves (at collision location of two consecutive longitudinal slopes) is decreased and consequently the risk of crash occurrence is increased.

Driver's behavior is one of the effective variables on crash occurrence, but it was convenient to include in the present study due to the limited time available. So it is recommended to other researchers in the future opportunities to notice the parameters and causes related to driver's behavior, environment condition and the other variables which are related to transportation infrastructures.

The above presented results indicate some of more important causes lead to crash occurrence in urban highways. The effectiveness amounts of the effective crash variables can be determined by showed models and it can reduce the amounts of their negative effects and consequently crash counts too. Moreover, the amounts of these variables can be noticed in new roads designing. So we hope that this research can be applied as a reference for engineers to design and repair urban highways.

References

[1] B. Persaud, and L. Dzbik, "Accident prediction models for freeways," *Transportation Research Record*, vol. 1401, pp. 55–60, 1993.

[2] M. W. Knuiman, F. M. Council, and D. W. Reinfurt, "Association of median width and highway accident rates," *Transportation Research Record*, vol. 1401, pp. 70–82, 1993.

[3] M. A. Hadi, J. Aruldhas, L. Chow, and J. A. Wattleworth, "Estimating safety effects of cross-section design for various highway types using negative Binomial regression", *Transportation Research Record*, vol. 1500, pp. 169–177, 1995.

[4] J. C. Milton and F. L. Mannering, "The relationship between highway geometrics, traffic related elements, and motor vehicle accidents," Department of Transportation, Washington State, FHA, Final Report, 1996.

[5] T. F. Golob and W. W. Recker, "The Relationship among urban freeway accidents, traffic flow, weather, and lighting conditions," *Transport Engineering*, vol. 129, pp. 342–353, 2003.

[6] T. F. Golob, W. W. Recker, and V. M. Alvarez, "Toll to evaluate safety effect of changes in freeway traffic flow," *Journal of Transportation Engineering*, ASCE, vol. 130, no. 2, pp. 222–230, 2004.

[7] E. Hauer, "Statistical road safety modeling," in *Proceedings of the 83rd TRB Annual Meeting*, Washington, DC, USA, 2004.

[8] C. Caliendo and A. Parisi, "Principal component analysis applied to crash data on multilane roads," Department of Civil Engineering, University of Salerno, 2004.

[9] C. Caliendo, M. Guida, and A. Parisi, "A crash-prediction model for multilane roads," *Accident Analysis and Prevention*, vol. 39, pp. 657–670, 2007.

[10] P. Anastasopoulos, A. Tarko, and F. Mannering, "Tobit analysis of vehicle accident rates on interstate highways," *Accident Analysis and Prevention*, vol. 40, no. 2, pp. 768–775, 2008.