Principal Component Analysis Applied to Crash Data on Multilane Roads.

Caliendo, C.
Associate Professor, Department of Civil Engineering, University of Salerno
e-mail: ccaliendo@unisa.it

Parisi, A.
Ph. D. Student, Department of Civil Engineering, University of Salerno
e-mail: alparisi@unisa.it

Synopsis

During the last few years, a lot of road-accident-predictive models have been developed by using Multiple Linear Regression and Poisson or Negative Binomial Distribution. More innovative methodologies based on fuzzy logic and neural networks have also been used. The application of these methodologies is not easy when a large number of variables is considered. Moreover, the influence of some variables on road accidents might not be equally significant. It would thus appears useful to have an analysis tool primarily in order to remove the redundant variables for accident-predictive models. Even if under-used in crash data, Principal Component Analysis (PCA) may be suitable for this purpose. PCA is a form of analysis used for extracting a reduced number of factors, called principal components, from a set of original variables, discarding as little of the information as possible.

Our objective is to verify PCA potentiality for removing redundant variables in accident analysis. For this purpose a five-year monitoring period was carried out on a four-lane median divided road. A database was subsequently created with the surveys regarding the type and number of accidents, traffic flow, horizontal and vertical alignment, sight distances and pavement surface characteristics. PCA was applied to homogeneous sections having constant horizontal curvature, separated into tangents and curves. By means of the correlation matrix the results indicate that the number of accidents on curves increases with the length (L) of the homogeneous sections, the curvature radius (1/R), the average daily traffic (TGM) and the design speed change (ΔV) between tangents and curves; whereas there is a negative correlation between these crashes and the longitudinal slope (i%), the sight distance (vis) and the pavement friction defined in terms of CAT (Side Friction Coefficient measured by means of a SCRIM equipment). Thus the results obtained prove the knowledges about this subject. Six principal components were found to account for about 90% of the variance in the original eight variables. The multiple correlation coefficient (ρ^2) between the original variables and the principal axes shows that the least significant variable is ΔV. In keeping with the literature, the correlation matrix for tangents indicates that road accidents are positively correlated to the length of the homogeneous sections (L) and the average daily traffic (TGM), and negatively correlated to the pavement friction (CAT) and the longitudinal slope (i%). Four principal components were found to account for over 90% of the variance in the original five variables. The multiple correlation coefficient for tangents (ρ^2) shows that the variables examined are all equally significant.

Principal Component Analysis Applied to Crash Data on Multilane Roads.

Road safety depends mainly on the relationship among the following three components, namely the human factor (speed, perception of road characteristics, driving behavior and psycophysical capabilities), the vehicle (performance and tyre-road interaction), and the environment (horizontal and vertical alignment, sight distance, pavement surface conditions, weather conditions, safety barriers, signals, lighting, traffic flow). Many studies have analysed these parameters with the purpose of identifying the combination that most affects accidents.
Evaluating the combined action of these three components (human factor-vehicle-environment) is most complicated, especially on account of the difficulty in finding sufficient data about the first two parameters. However, in road engineering, a relationship between road safety and environment is more significant. During the last few years, a lot of accident-predictive models relating to traffic flow, infrastructure geometry, pavement surface and weather conditions have been developed by using multiple linear regression and Poisson or Negative Binomial distribution. More innovative methodologies based on fuzzy logic and neural networks have also been used. Many researchers have refined these models showing their operating limits and revealing questions that are still open. However, these models were developed in foreign countries where accident surveys, human behavior, infrastructure and traffic characteristics differ from those in Italy. They also refer to two–lane rural roads, while four–lane median divided roads have been investigated to a lesser degree. Moreover, no model considers all the variables affecting accidents both on account of the difficulty in managing a lot of data and also because some of these are less significant. Therefore we need an analysis tool to verify primarily that all the original variables affect accidents, or disregard redundant ones in looking for accident-predictive models. With regard to this problem, it is becoming ever more acknowledged that the technique of principal component analysis (PCA) appears to offer an appropriate approach for identifying the main variables that cause accidents.
Such is the context wherein the present work is set. This paper studies crashes occurring in our country (Italy) on four–lane median divided roads, as a function of traffic flow, infrastructure characteristics, pavement surface conditions and sight distance. The objective is to identify the most significant variables affecting road accidents. For this purpose, a five-year monitoring period was carried out on a specific infrastructure. The database contains homogeneous sections, having horizontal constant curvature, divided into tangents and curves. Accidents, sight distance, longitudinal slope, design speed difference between consecutive elements (ΔV), pavement friction (CAT) and traffic flow (TGM) were associated with these homogeneous sections. To understand the correlation structure of the variables and to remove unnecessary redundancy from the original set of variables so that we might interpret the results accurately, principal component analysis (PCA) was performed. PCA was applied to extract a reduced number of variables (called principal components), which are a linear combination of original variables, discarding as little of the information as possible. The interest in this approach was due to the need to verify the potentiality of this analysis tool and to address the choice of variables for predictive models more appropriately.

STUDY METHODOLOGY

Our analysis method involved several steps. First, the study started by monitoring a rural motorway. This infrastructure, which is a four-lane median divided road, is 51.6 km long. The horizontal alignment contains tangents and circular curves without transition curves. The vertical alignment consists of gradients and circular curves. Monitoring was conducted between the years 1998 and 2003, during which time accident data, pavement friction and traffic flow were collected. We possess 1916 observed accidents, 1320 of which were crashes occurring on tangents and 596 on curves. Thus, the number of accidents on tangents was twice that occurring on curves. The data base does not cover accidents that took place at junctions, service areas, tollbooths and hard shoulders since such accidents are not due to traffic flow and infrastructure characteristics. The database was organized for each carriageway in function of homogeneous sections with constant horizontal curvature, divided into tangents and curves. The number of accidents per year, section length (L), curvature radius ($1/R \neq 0$ only for curves), longitudinal slope (i%), average daily traffic (TGM), sight distance (vis) on curves, pavement friction (CAT), and design speed difference (ΔV) between tangent and curve were associated with every homogeneous sections. Thus the number of accidents per year and carriageway occurring on these homogeneous sections represents the dependent variable that could be related both to the traffic flow and environmental conditions for accident predictive models.

Since the objective of the successive step is to find the most significant variables both for tangents and curves, principal component analysis (PCA) was duly performed.
First of all, the variables examined were standardised to overcome the different unit of measurement of the original variables obtaining the **Z** matrix:

$$z_{ij} = \left(x_{ij} - \bar{x}_j\right)/ s_j \qquad \text{for i = 1,2,...,n and j = 1,2,...,q}$$

where $\bar{x}_j$ and s_j are, respectively, mean and standard deviation of the generic variable x_j. The elements in **Z** matrix have zero mean and unit variance. The covariance between two variables z_k e z_j (for $k \neq j$ = 1,2,...,n) is the correlation coefficient.
Next the covariance matrix (**R**) was computed, which contains the correlations between the original variables:

$$R = \frac{1}{n} Z'Z$$

where **Z'** is given by the matrix transpose of **Z**.
From the so-called characteristic equation of the covariance matrix:

$$\det\left(R - \lambda I\right) = 0$$

where **I** is the identity matrix containing unit values, the eigenvalues λ_h were calculated.
The variance accounted for by the principal components were computed with the eigenvalues λ_h.
The eigenvectors $\mathbf{v}_h$ were identified by means of the matricial equation:

$$\left(R - \lambda_h I\right)v_h = 0$$

in addition, the **S** matrix was determined, which contains the correlation coefficients between the original variables and the principal axes:

$$S = \frac{1}{n} Z'YL^{-1/2} = \frac{1}{n} Z'ZVL^{-1/2} = RVL^{-1/2}$$

where **V** is the matrix of the eigenvectors and **L** is the diagonal matrix whose elements are the eigenvalues.
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:

$$U = ST$$

where **T** is the rotation matrix.
The multiple correlation coefficients (ρ^2) between original variables and rotated principal axes 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.
Finally the results obtained were plotted and compared by means of the correlation circle. The objective of this last step was to understand more clearly the correlations between the original variables and the rotated principal axes. When the points representing the original variables are closer to the principal axes as well as to 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.

RESULTS

The PCA results are summarized in this paragraph. The analysis was carried out distinctly for curves and tangents to show the separate role they play on crashes.

Curves

The matrix **X** of original variables was obtained by using data from the monitored infrastructure, which contains: the number of accidents per year and carriageway (n.acc./year*carr) occurring on curves, the

length (L) of curves, the curvature radius (1/R), the average daily traffic (TGM), the pavement friction (CAT), the longitudinal slope (i%), the sight distance (vis) and the design speed difference (ΔV) between tangent and curve. In order to overcome the different unit of measurement of the original variables, the standardised matrix (**Z**) was computed. Then the covariance matrix (**R**) presented in Table 1 was derived:

Table 1: Covariance Matrix (R) for Curves

	n.acc./year*carr	L	1/R	TGM*10^{-4}	CAT	i (%)	vis	ΔV
n.acc./year*carr	1	0.23	0.10	0.15	-0.08	-0.05	-0.06	0.04
L	0.23	1	-0.27	0.09	0.05	-0.01	0.13	0.21
1/R	0.10	-0.27	1	-0.43	-0.08	-0.01	-0.35	-0.63
TGM*10^{-4}	0.15	0.09	-0.43	1	-0.15	0.00	0.15	0.32
CAT	-0.08	0.05	-0.08	-0.15	1	-0.05	0.09	0.04
I (%)	-0.05	-0.01	-0.01	0.00	-0.05	1	-0.04	-0.05
vis	-0.06	0.13	-0.35	0.15	0.09	-0.04	1	0.23
ΔV	0.04	0.21	-0.63	0.32	0.04	-0.05	0.23	1

The covariance matrix coincides with the correlation matrix as the variables are standardised. The diagonal contains unit elements that represent the correlations of each variable with itself. The remaining elements are the correlations between the different variables: positive values indicate a direct proportionality between variables, and negative values show an inverse relationship between them. In the first column of Table 1, a positive relationship is observed between the number of accidents (n.acc./year*carr) and: the length (L) of curves, the curvature radius (1/R), the average daily traffic (TGM) and the design speed difference (ΔV). These crashes are, in contrast, negatively related to the pavement friction (CAT), longitudinal slope (i%) and sight distance (vis). The results obtained confirm our knowledge with regard to the way in which the number of accidents on curves increases or decreases in accordance with the variables examined.

By using the above-mentioned characteristic equation of the covariance matrix, the eigenvalues λ_h were found. Then the eigenvectors $\mathbf{v}_h$ were calculated by means of the relative matricial equation. Finally both the **S** matrix and **U** (rotated matrix of S) were computed.

Table 2 contains the calculated eigenvalues λ_h, the variance accounted for ($\lambda_h/\Sigma\,\lambda_h$) and the cumulative percentage of the variance accounted for ($\Sigma\,(\lambda_h/\Sigma\,\lambda_h * 100)$).

Table 2: Eigenvalues λ_h, variance accounted for ($\lambda_h/\Sigma\,\lambda_h$), cumulative percentage of variance accounted for ($\Sigma\,(\lambda_h/\Sigma\,\lambda_h * 100)$) and multiple correlation coefficients (ρ^2) between original variables and rotated principal axes for curves

	λ_h	($\lambda_h/\Sigma\,\lambda_h$)	$\Sigma\,(\lambda_h/\Sigma\,\lambda_h * 100)$
λ_1	2.222	27.77	**28 %**
λ_2	1.256	15.70	**43 %**
λ_3	1.115	13.94	**57 %**
λ_4	0.984	12.30	**70 %**
λ_5	0.800	10.00	**80 %**
λ_6	0.740	9.25	**89 %**
λ_7	0.584	7.30	96 %
λ_8	0.300	3.75	100 %

	u_1	u_2	u_3	u_4	u_5	u_6	ρ^2
n.acc./ year*carr	0.105	0.352	-0.270	-0.091	-0.039	-0.848	**0.94**
L	-0.265	-0.067	-0.839	-0.119	0.228	-0.165	**0.87**
1/R	0.881	0.028	0.031	-0.040	-0.194	-0.072	**0.82**
TGM*10-4	-0.603	0.501	0.247	0.089	-0.012	-0.304	**0.78**
CAT	-0.105	-0.883	0.175	0.035	0.043	-0.399	**0.98**
i (%)	0.024	0.027	-0.176	0.979	-0.067	0.036	**1.00**
Vis	-0.221	0.009	0.234	0.024	0.943	0.026	**0.99**
ΔV	-0.845	-0.009	-0.085	-0.083	0.008	-0.018	**0.73**

Six principal components were found to account for about 90% of the variance in the original eight variables. The so-called “loading factors” (elements of the **U** matrix), which represent the correlation coefficients between the original variables and the rotated principal axes, are also listed in Table 2.

The multiple correlation coefficients (ρ^2) between the original variables and the six rotated principal axes are to be found in the last column. The ρ^2 values are between 0.73 and 1.00. The lowest multiple correlation coefficient is associated with variable ΔV. Therefore this last one is the least significant variable affecting accidents on multilane roads, and may be ignored when finding predictive models. The redundancy of ΔV appears be due to the great influence on accidents of the variable 1/R to which ΔV is linked.

Figure 1 presents the original variables in the plane of the first two principal components that make a greater contribution to the variance accounted for.

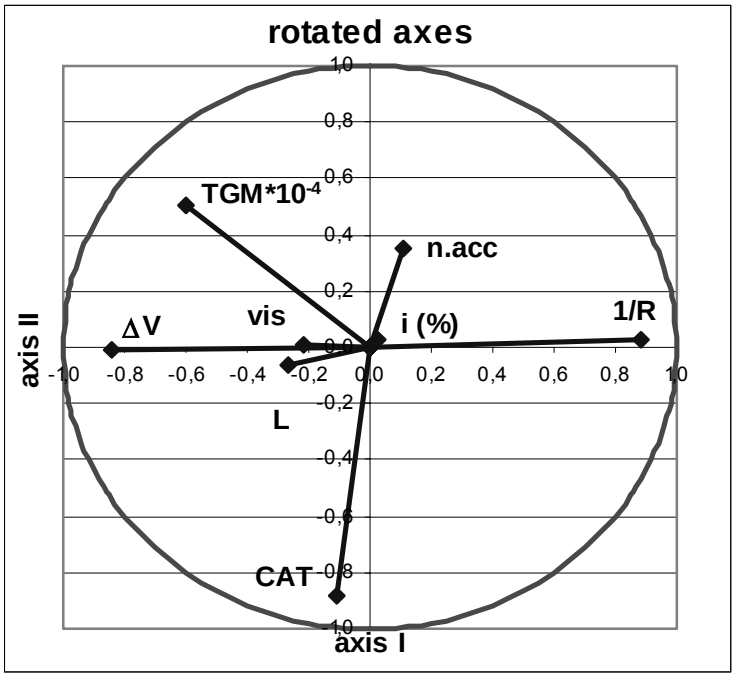

Figure 1: Graphical representation of the original variables on the principal plane and correlation circle for curves

Both the curvature radius (1/R) and the design speed difference (ΔV) are closely correlated to the first principal axis. A lower correlation between this axis and the length (L) of curves, the sight distance (vis) and the longitudinal slope (i%) is instead shown. The variables L and vis are closer to each other, and a close correlation between these ones is registered.
The number of accidents on curves (n.acc./year*carr) and the pavement friction (CAT) are closely correlated to the second principal axis. A less close correlation exists with the average daily traffic (TGM).
Thus the geometric variables appear to be represented by the first principal axis, while the remaining ones by the second principal axis.

Tangents

The matrix **X** of the original variables contains the following: the number of accidents per year and carriageway (n.acc./year*carr) which occurred on tangents, the length (L) of tangents, the average daily traffic (TGM), the pavement friction (CAT) and the longitudinal slope (i%).
The standardised matrix (**Z**) and the covariance matrix (**R**) were calculated. The elements of the matrix **R** are contained in Table 3.

Table 3: Covariance Matrix (R) for Tangents

	n.acc./year*carr	L	TGM*10^{-4}	CAT	i (%)
n.acc./year*carr	1	0.56	0.35	-0.07	-0.05
L	0.55	1	0.11	0.11	0.00
TGM*10^{-4}	0.32	0.10	1	-0.17	-0.01
CAT	-0.08	0.07	-0.26	1	-0.02
i (%)	-0.02	0.01	0.00	-0.07	1

In the first column of Table 3, a positive relationship is shown between the number of accidents (n.acc./year*carr) and the length (L) of tangents and average daily traffic (TGM). Accidents, in contrast,

correlate negatively to the pavement friction (CAT) and longitudinal slope (i%). In keeping with the literature, the results confirm the trend of accidents with these variables.
Table 4 contains the calculated eigenvalues λ_h, the variance accounted for ($\lambda_h/\Sigma\,\lambda_h$) and the cumulative percentage of the variance accounted for ($\Sigma\,(\lambda_h/\Sigma\,\lambda_h * 100)$).

Table 4: Eigenvalues λ_h, variance accounted for ($\lambda_h/\Sigma\,\lambda_h$), cumulative percentage of variance accounted for ($\Sigma\,(\lambda_h/\Sigma\,\lambda_h * 100)$) and multiple correlation coefficients (ρ^2) between the original variables and rotated principal axes for tangents

	λ_h	$(\lambda_h/\Sigma\,\lambda_h)$	$\Sigma\,(\lambda_h/\Sigma\,\lambda_h * 100)$
λ_1	1.711	34.19	**34 %**
λ_2	1.183	23.64	**58 %**
λ_3	0.999	19.96	**78 %**
λ_4	0.722	14.43	**92 %**
λ_5	0.389	7.77	100 %

	u_1	u_2	u_3	u_4	ρ^2
n.acc./year*carr	-0.854	-0.194	0.028	-0.135	**0.79**
L	-0.903	0.167	-0.043	0.040	**0.85**
TGM*10^{-4}	-0.213	-0.710	0.008	-0.663	**0.99**
CAT	0.003	0.819	0.027	-0.566	**0.99**
i (%)	0.029	-0.019	-0.999	0.007	**1.00**

Four principal components were found to account for 92% of the variance in the original five variables.
The "loading factors" (correlation coefficients between the original variables and the rotated principal axes) are also reported in Table 4.
The multiple correlation coefficients (ρ^2) between the original variables and the four rotated principal axes are in the last column. The ρ^2 values are contained between 0.79 and 1.00. In this case all variables are significant because of the higher values of this parameter.
Figure 2 presents the original variables in the plane of the first two principal components that give the higher contribution to the variance accounted for.

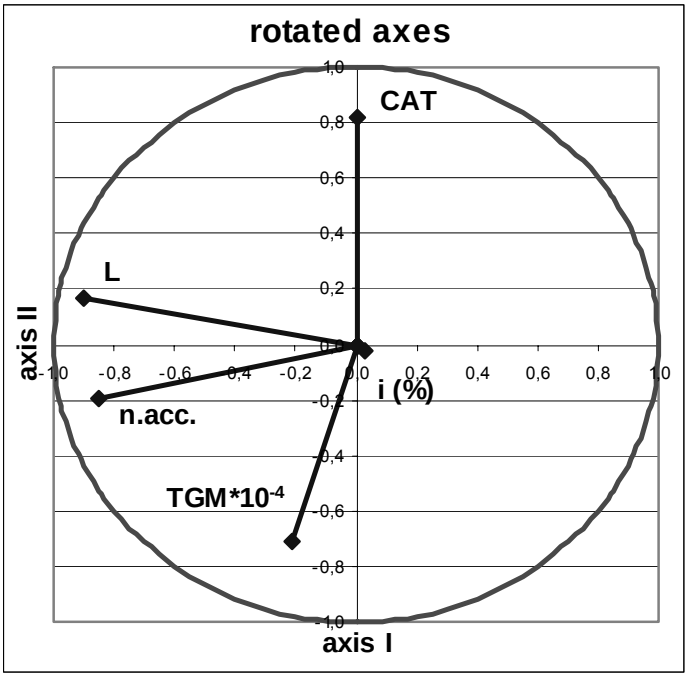

Figure 2: Graphical representation of the original variables on the principal plane and correlation circle for tangents

The number of accidents on tangents (n.acc./year*carr) and the length (L) of tangents are closely correlated with the first principal axis, whereas there is a lower correlation between this axis and the longitudinal slope (i%). Furthermore, the number of accidents and the length of tangents are closer to each other, and also a close correlation between these variables is shown.
The pavement friction (CAT) and the average daily traffic (TGM) have a close correlation with the second principal axis.
As already seen, the geometric variables are represented by the first principal axis, while the remaining ones by the second principal axis.

CONCLUSION

The objective of this paper is to remove unnecessary redundancy in accident analysis in order to better address the choice of the most significant variables for predictive models.
Considerable progress has been made in recent years in techniques for establishing the relationships between accidents, traffic flow and road geometry. It has been generally acknowledged that the use of the Negative Binomial distribution is more appropriate than the Poisson distribution or conventional multiple linear regression.
More innovative methodologies such as fuzzy logic and neural network have also been used. The application of the approaches mentioned is not easy when a large number of variables is considered. Moreover, the influence of certain variables on road accidents might not be equally significant. Principal Component Analysis (PCA), even if underused in crash data, is now being considered as a suitable tool for verifying primarily whether all variables affect accidents or disregard redundant ones.
In using crash data of a monitored four-lane divided road, PCA has been successfully applied in this work.
The results of the analysis carried out distinctly for curves and tangents prove the knowledges on this subject.
For curves, the correlation matrix reveals that the number of accidents per year and carriageway (n.acc./year*carr) occurring on them increases with: the length (L) of curves, the curvature radius (1/R), the average daily traffic (TGM) and the design speed difference (ΔV) between tangents and curves. These accidents decrease, in contrast, with the pavement friction (CAT), longitudinal slope (i%) and sight distance (vis). Six principal components were found to account for about 90% of the variance in the original eight variables. The multiple correlation coefficients (ρ^2) between the original variables and the six rotated principal axes were computed to be between 0.73 and 1.0. The ΔV, having the lowest ρ^2, was identified as the least significant variable and may be ignored when looking for predictive models.
For tangents, the correlation matrix shows that the number of accidents per year and carriageway (n.acc./year*carr) on them increases with the length (L) of tangents and average daily traffic (TGM). The number of accidents decreases with the pavement friction (CAT) and longitudinal slope (i%). Four principal components were found to account for over 90% of the variance in the original five variables. The multiple correlation coefficients (ρ^2) was computed between 0.79 and 1.00. The higher values of ρ^2 found for tangents indicate that the five examined original variables are all significant.
Although the results indicate that, for the case studied herein, only one variable may be ignored out of the set of eight original variables, we believe that we have demonstrated that the use of PCA is appropriate for removing redundant variables in accident analysis. Thus, it is hoped that in the future researchers will show greater interest in this methodology.

REFERENCES

ABDEL-ATY, M.A., RADWAN, A.E. (2000), "Modeling Traffic Accident Occurrence and Involvement", *Accident Analysis and Prevention*, volume 32, issue 5, pp. 633-642.
CALIENDO, C., LAMBERTI, R. (2001), "Relationship Between Accident and Geometric Characteristics for four Lanes Median Separated Roads", *Proceedings of the 12th International Conference on Three Continents*, Moscow, 19-21 September.
CALIENDO, C., PARISI, A., VILLANI, P. (2003), "Analisi dell'incidentalità sulle strade a carreggiate separate in relazione alle caratteristiche dell'infrastruttura e dell'intensità di pioggia", *XIII National Conference S.I.I.V, Padova 30-31 October 2003*.
GOLOB, T.F., RECKER, W.W. (2003), "Relationship Among Urban Freeway Accidents, Traffic Flow, Weather, and Lighting Conditions", *Journal of Transportation Engineering*, volume 129, No. 4, July 1, pp. 342-353.
GOLOB, T.F., RECKER, W.W., ALVAREZ, V.M. (2004), "Tool to Evaluate Safety Effects of Changes in Freeway Traffic Flow", *Journal of Transportation Engineering*, volume 130, No. 2, March 1, pp. 222-230.
GOLOB, T.F., RECKER, W.W., ALVAREZ, V.M. (2004), "Freeway Safety as a Function of Traffic Flow", *Accident Analysis and Prevention*, volume 36, issue 6, pp. 933-946.
KENDALL, M.G., STUART, A. (1968), *The Advanced Theory of Statistics*, Charles Griffin & Company Limited, London.
MARDIA, K.V., KENT, J.T., BIBBY, J.M. (1979), *Multivariate Analysis*, Academic Press, London.
MENSAH, A., HAUER, E. (1998), "Two Problems of Averaging Arising in the Estimation of the Relationship Between Accidents and Traffic Flow", *Transportation Research Record 1635*, pp.37-43.
PERSAUD, B., DZBIK, L. (1993), "Accident Prediction Models for Freeway", *Transportation Research Record 1401*, pp.55-60.
PERSAUD, B., RETTING, RA., LYON, C. (2000), "Guidelines for Identification of Hazardous Highway Curves", *Transportation Research Record 1717*, pp.14-18.
WILEY, J. & S. (1984), *Handbook of Applicable Mathematics*, Lloyd Emlyn, New York.