Journal of the Eastern Asia Society for Transportation Studies, Vol. 6, pp. 3669 - 3682, 2005

Development of an Accident Prediction Model using GLIM (Generalized Log-linear Model) and EB method: A case of Seoul

KIM, Se Hwan
Graduate Student
Dept. of Civil, Urban and Geo-systems Engineering
Seoul National University
San 56-1 Shillim-dong, Gwanak-gu
Seoul, Republic of Korea
Fax: +82-02-889-0032
E-mail: ksehwan78@hanmail.net

CHUNG, Sung Bong
Research Fellow
Center for Transport Infrastructure Investment
The Korea Transport Institute
2311, Daehwa-dong, Ilsan-gu, Goyang-si
Gyeonggi-do, 411-701, Korea
Fax: +82-31-910-3224
E-mail: bbaqui@koti.re.kr

SONG, Ki Han
Ph. D. Candidate
Dept. of Civil, Urban and Geo-systems Engineering
Seoul National University
San 56-1 Shillim-dong, Gwanak-gu
Seoul, Republic of Korea
Fax: +82-02-889-0032
E-mail: willkh@hanmail.net

CHON, Kyung Soo
Professor
Dept. of Civil, Urban and Geo-systems Engineering
Seoul National University
San 56-1 Shillim-dong, Gwanak-gu
Seoul, Republic of Korea
Fax: +82-02-889-0032
E-mail: chonks@snu.ac.kr

Abstract: The objective of this paper is developing an accident prediction model at four-legged signalized intersections in Seoul City to control random and local characteristics of accident. The first step is to classify and analyze the factors of accidents, and construct raw accidents data as an ordinal category. This step is able to make the structure of accidents data to include random characteristic, and the next step is to make a prediction model using GLIM(Generalized Log-linear models) including the error system having the negative-binomial distribution. Then, EB method (The Empirical Bayesian method) using the cross and time series data of subject elements is supplemented to the basic model, in order to correct the global prediction results. In this paper, the total 145 intersections in 156 intersections in Seoul are used, 80 for calibration and 65 for validation, and 11 intersections are abandoned because of short on data.

Key Words: Accidents prediction model, Generalized Loglinear Model, EB method

1. INTRODUCTION

The development of a transportation system has been the generative power for human beings to have the highest civilization above creatures in the earth. However, this growth has caused safety problems that the transportation systems for our efficiency and comfort rob us of our lives, so efforts to overcome this issue have been and will be made by many a person in various fields. Especially, transportation engineers have made passionate researches to analyze accidents and predict potential accidents in our systems. The objective of this paper is to develop an accident prediction model at four-legged signalized intersection in Seoul City in order to serve a turnover point of these endeavors.

First of all, accidents have the random characteristic in itself, and the general regression models for analyzing accidents has the fatal limit in this point. In this study, the first step for coming over this boundary is to classify and analyze factors of accidents, and construct raw accidents data as an ordinal category on the base of this fundamental task. This step is able to make the structure of accidents data to include random characteristic of our enemy against safety. The next step is to make a prediction model using GLIM (Generalized Log-linear models), which can use the ordinal category structure and include the error system having the negative-binomial distribution.

However, this is just beginning for making a satisfaction to our goal, because this method can cover only random characteristic not local characteristic of accident. In order to get a solution of this limitation, EB method (The Empirical Bayesian method) is supplemented to the basic model, an accident prediction model using GLIM, for this method is able to corrects the global prediction result which does not include local characteristic of accidents, and use cross and time series data of subject elements. This integration model using GLIM and EB method is the final result of our study.

In this paper, total 145 intersections in Seoul from 1998 to 2000 are used. Moreover, past accidents are classified by referring to accident diagrams. Accident of intersection is classified as three banishment of right angle collision, rear-end collision, and broad side collision. The development of an accident prediction model using this algorithm can be the first pace for solving the limit of random and local characteristic of accidents, which is not included clearly in the present models.

2. LITERATURE REVIEW

2.1 Accidents Prediction Models using Multiple Regression

Multiple Linear Regression model has various shortcomings to use for predicting the number of accident. First, dependent variables are assumed to follow normal distribution in this model, but the number of accident is not so. And it is assumed that there is no relation between error and independent variable, but this assumption is not always true in case of accident in actuality. In addition, this model can deduce the negative number that could not appear as the number of accident. Moreover, when accident did not happen in any spot, this method always predicts zero as the number of accident, and this result strains the truth that zero number means that spot absolutely safe.

Poisson models deal with discrete data so that they have most of desirable characteristics to describe vehicle collisions of positive number and random attribution. However, these models may produce wrong coefficients and wrong standard errors if data has excessive variance, and it has been problem to apply this model because variables that explain the number of accident were categorized data.

Usually, Log linear models are considered as a basic method to analyze effects of categorized data. Basically, this model can be expressed as follows:

$$\ln Y_i = \beta(X_i) \qquad i = 1, 2, \ldots, n \tag{1}$$

Where, Y_i: Number of accident for combination i

X_i: Independent variable i

β Parameters in formula measure relation in X_i and display degrees of explanation power about accidents.

Reason, why this model is widely used, is that elements affecting in accident are categorized data. Moreover, because accidents are discrete essentially, expression of difference about accident reaction is most efficient in expressed data system by categorized style. And log linear method make it possible to test significance of categorized data as fixed quantity. In addition, it can handle positive characteristic of accident by Poisson distribution. However, weighted least squares method (WLSA) that use for exact calculation in this model usually displays high residual and needs large sample generally.

2.2 Empirical Bayesian Method

Application of Empirical Bayes (EB) through multivariate analysis technique was tried in earlier researches for measurement to find risk sections in roads. Standards used for estimation of some object's risk degrees depends on the object's characteristic and past incident records. When it used only one standard, it has various problems. If it wishes to deduce the number of accident using only one characteristic of object, it has a serious problem to choose reference group has an object. First, if it tries to find characteristics as like the object has, because the number of sample is decrescent, estimation becomes difficult. Then, some other elements that can contribute in accident are abandoned and analyst can intervene in selection of characteristics.

Moreover, if it only uses data of past accident, a problem will be caused by random characteristics of accident, and it will be not easy to find average of reference group including the object. Thus, development of accident prediction models used in before and after test of improvement achievement or comparison test with other targets can be impossible. Therefore, EB method makes estimation model of the number of accident using multivariate analysis technique, and it produces the corrected number of accident using past accident.

EB method assumes that the number of accident depends on not only characteristics of object but also past accident records. When the number of accident is estimated, two factors should be used. If x, the number of accident, follows Poisson distribution and belongs to reference group, which has E(y) and VAR(y), estimation equation of this object is as follows:

$$\alpha E(y) + (1-\alpha)x \tag{2}$$

$$\text{Where, } \alpha = \frac{E(y)}{E(y) + \mathrm{Var}(y)} \tag{3}$$

Where, y: the number of accident

Variance of level of risk is as follows:

$$\alpha(1-\alpha)E(y) + (1-\alpha)^2 x \tag{4}$$

3. DEVELOPMENT OF AN ACCIDENT PREDICTION MODEL

3.1 Data Collection and Variable Selection

Data sampled 156 four-legged signalized intersections is given in the help of 'Road Traffic Safety Authority' in Korea. However, 145 intersections are extracted from raw data and 11 intersections are abandoned because of their inconsistence with other intersections, and total 80 intersections are used to calibration and 65 remainders to validation.

Set of accident data is arranged about three occasions that are rear-end collision, broadside collision and right angle collision that occur frequently in all accident patterns of basic accident diagram. Broadside collision includes minor collision, turning left broadside collision, turning right broadside collision and minor collision of road alteration, and right angle collision includes broadside right angle collision, the head of vehicle right angle collision.

First of all, variables that need in construction of model formula are selected as many as possible on the basis of earlier researches and possible variables such as total volume, number of major lane, number of minor lane, intersection angle, median, signal control pattern, intersection location, intersection square measures, visual noise level for the entering approach etc. are arranged. Then, we investigated whether selected intersections are affected by other variables that are not selected and decided variables in preceding process, and all the selected variable are analyzed statistically, and last variables are fixed. Content of selected variable is as following Table 1. We organize given variables as category to use in the model because GLIM needs categorized data based on static analysis.

Traffic volume is used as offset of this model, and it is calculated by uniting discharged traffic by entry direction in intersection because the number of accident is dependent variable in intersections. Two-dimensional geometric structure is mainly used as variables of this model because it supposed that driver's condition and traffic operation are same in these intersections.

3.2 Construction of Model Formula

In this study, we develop estimation model of the number of accident at four-legged signalized intersections of Seoul City using general linear model (Generalized linear model) as basic model for development of final estimation model. Because the number of accident and dependent variables are discrete data and follows Poisson distribution, we use general

linear model and Maximum Likelihood Estimation (MLE) method. Log function is used by link function because occurrences of accident follow Poisson distribution. This research is using general log-linear model (Generalized log-linear model) conclusively. Model formula of general log-linear model (Generalized log-linear model) is expressed as following formula:

$$\ln \mu - \ln F = \beta_0 + \beta X^T \quad (5)$$

Where, $\mu : E(y)$

F : Offset

y : The number of accident in total intersections

β_0 : Dummy variable

β : Vector of parameter

X : Vector of independent variable

Table 1. Content of Applying Variable in Model

Variable	Sign	Category	Description	Variable	Sign	Category	Description
Total volume	F	Positive number	Total entry volume	Safety Zone	S	1	O
						2	X
Number of major lane	LJ	1	LJ ≤ 4 lanes	Building Turning Left lane	L	1	O
		2	LJ > 4 lanes			2	X
Number of minor lane	LM	1	LM ≤ 2 lanes	Building Turning Right lane	R	1	O
		2	LM > 2 lanes			2	X
Intersection angle	A	1	$A < 15^\circ$	Turning Left Prohibition	LP	1	O
		2	$15^\circ \leq A < 30^\circ$			2	X
		3	$A > 30^\circ$				
Median	M	1	O	Channelization	E	1	O
		2	X			2	X

3.3 Calibration of Basic Model

Basic model, which has the largest significance in candidate models, is selected after given candidate models are constructed through analysis of main factors and earlier researches using selected variables for development of basic model formula. We judged goodness of fit using SAS 8.e and observed P value in estimation of parameter for judgment of significance. Then, likelihood ratio is examined through sequential test and partial test, and analyze residual degrees is tested. A final model is selected based on these results in accordance with common sense.

First, value of deviance, pearson x^2, log likelihood ratio are standards for goodness of fit. As given deviance and pearson x^2 are larger than value of criterion, it is better because null hypothesis that is statistic follows distribution x^2 should be rejected. There is not serious problem if value of goodness of fit (log likelihood ratio) does not differ greatly because it is relative value between comparison models.

Moreover, P value should be low in each parameter. In addition, Goodness of fit is examined through sequential test and partial test to know effect of each variable, and we examined influence of each variable for model. Sequential test is examination method that adds variable one by one, and partial test is method that examines goodness of fit when model is constructed by each variable. Most suitable model was decided after we test significance of model and examine existence of odd point analyzing residual degree finally.

Table 2. Candidate Model Formula

Number of model	Model formula
1	ln Y = ln F + A + LJ + LM + L + LP + S + M + R + E
2	ln Y = ln F + A + LJ + LM + M + R
3	ln Y = ln F + A + LM + R + M
4	ln Y = ln F + A + LM + M + R + E
5	ln Y = ln F + A + LM + L + M + R

3.3.1 Rear end collision model

Model 4 showed that deviance and pearson x^2 are much greater than others. Although log likelihood is low, model 4 is selected as basic formula because difference is not significant when model 4 compared with other models. Given basic model formula is as follows.

$$\ln Y = \ln F - 6.7224 + 0.2925 \times A(1) + 1.0678 \times A(2) - 0.0864 \times LM(1) - 0.0312 \times M(1) - 0.0156 \times R(1) - 0.0310 \times E(1)$$

This model proves that intersection angle affects mostly on safety, and when the number of minor lane is low and turning right lane is built and channelization is applied, intersections becomes more safe in the case of rear end collision.

Intersection angle gives positive influence to rear end collision because of sight distance and in case of the number of minor lane, negative effect is expected because more small as the number of lane in a road section is, less possibility of accident occurrences as it has. Building median, turning right lane and applying channelization can also reduce possibility of accidents itself.

Table 3. Result of Statistics Analysis

	1	2	3	4	5
Deviance	3.3426	3.4466	3.4284	3.4809	3.355
Scaled Deviance	3.3426	3.4466	3.4284	3.4809	3.355
Pearson x^2	3.1833	3.3132	3.2908	3.3372	3.1306
Scaled Pearson x^2	3.1833	3.3132	3.2908	3.3372	3.1306
Log likelihood	1051.461	1041.4988	1040.3598	1040.4185	1044.384

3.3.2 Broad Side Collision Model

Model 2 showed that standard statistic value is not the best among any model formulas. Although model 1 includes all variables, there is little gap compared with model 2. Moreover, it is the most reasonable to explain broad side collision using model 2. Given basic formula is as follows:

$$\ln Y = \ln F - 6.7755 + 0.3831 \times A(1) + 0.7919 \times A(2) - 0.0097 \times LJ(1) - 0.2067 \times LM(1) - 0.0104 \times M(1) + 0.1071 \times R(1)$$

In the same way, we can know that effects of intersection angle is the most influential factor because of sight distance and that accident decreases as the number of major road and minor road decreases because there is a lot of accident occasions by lane alteration. Median gives

negative influence because it reduce opportunity of accident, and building turning right lane gives positive influence by sudden lane changes.

Table 4. Result of Statistics Analysis

	1	2	3	4	5
Deviance	3.0379	2.9531	2.9072	2.948	2.9519
Scaled Deviance	3.0379	2.9531	2.9072	2.948	2.9519
Pearson x^2	2.9216	2.931	2.8849	2.915	2.9362
Scaled Pearson x^2	2.9216	2.931	2.8849	2.915	2.9362
Log likelihood	979.4161	976.0103	976.0036	976.1699	976.0472

3.3.3 Right Angle Collision Model

Result of model 5 is the best when it compare with result of other models. Model 5 is selected because it is the most reasonable as possible model formula. Therefore, model 5 was selected conclusively and basic model is as follows:

$$\ln Y = \ln F - 7.1282 + 0.0053 \times A(1) + 0.2716 \times A(2) + 0.1048 \times LJ(1) + 0.0149 \times LM(1) + 0.0752 \times M(1) - 0.3246 \times R(1)$$

In this model, effects of intersection angle have the largest influences because sight distance is an important factor in the case of right angle collision, and the number of lanes also has much effects on this type of accident because right angle collision happen easily in narrow space. Then, installation of median gives positive influence in accident because median disturbs sight distance in the process of rotation, and installation of turning right road gives negative effect because it reduce opportunity of right turn collisions.

Table 5. Result of Statistics Analysis

	1	2	3	4	5
Deviance	2.3443	2.2498	2.2192	2.2517	2.2542
Scaled Deviance	2.3443	2.2498	2.2192	2.2517	2.2542
Pearson x^2	2.4441	2.3382	2.2988	2.3284	2.3384
Scaled Pearson x^2	2.4441	2.3382	2.2988	2.3284	2.3384
Log likelihood	326.0944	324.3828	324.237	324.323	324.2434

3.4 Basic Model Validation

The basic model is tested using Paired comparison method based on data which are not used in calibration, and the result deduces significance in 1% level. In conclusion, it is validated that given basic model formula can be transferable to other intersections.

In case of rear end collision, it seems to be accurate compared with other accident models, and the maximum value of t is 18.77. Considering in accident frequency and models by a practical manner, range of the estimate is usually from 0.015 at least to 10.7 at maximum. Moreover, it can be 0 in the real accident frequency, but it's not in a model. It's a not 0 but positive value, this is a reasonable result.

Table 6. Result of Validation

Model	Average	Standard deviation	t-value	Intercept		
Rear end collision Model	0.82	1.73	18.77	Level of Significance	1%	5%
Broad side collision Model	0.33	2.46	5.22		2.64	1.99
Right angle collision Model	0.63	1.79	13.82			

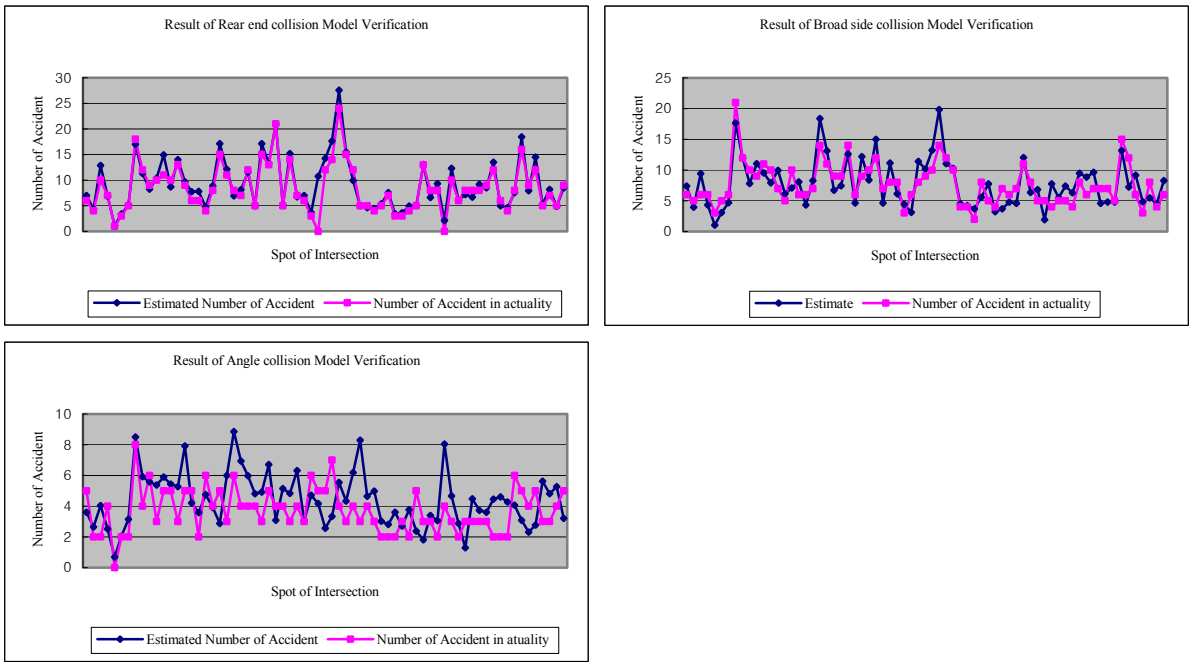

Figure 1. Results of Model Verification

In case of broad side collision, its accident frequency is similar to rear end collision. However, the theoretical value by an estimated model is the most inadequate of 3 different accident models. The mean value is the smallest in difference between the real accident frequency and the estimated value, but the standard deviation is so large that the value of t is the smallest, and value of t is satisfied in significance level 1%. The maximum gap is 14.35, and the minimum is 0.234.

In case of right angle collision, 5.289 is maximum difference and 0.034 is minimum difference between actuality value and estimate because the number of accidents of right angle collision is smaller than that of two accident types. Moreover, t value is suitable well in actuality number of accident next to rear end collision model by 13.82.

4. REVISION USING EB(EMPIRICAL BAYESIAN) METHOD

EB method should be applied to intersections of similar characteristic that is not used in model construction using accumulative past accidents data, but EB method randomly applied to intersections that are used to construct model by limit of data collection. $E(y)$ is available through multivariate model, but there is no preceding research for k in formula that is $VAR(y) = \frac{1}{k}E(y)^2$ to get $VAR(y)$. Therefore, we applied k value as it is same value in previous research.

Accident estimation and prediction model developed in this paper is more suitable than before because comparison with before EB correction and after EB correction can reflect factors that are not considered in basic models by accident types. However, more accident data and researches about k to prove significance of this model are needful because data used in model construction was applied to EB method again. Comparison with last estimation value using EB method and number of accident in actuality for each model is as follows:

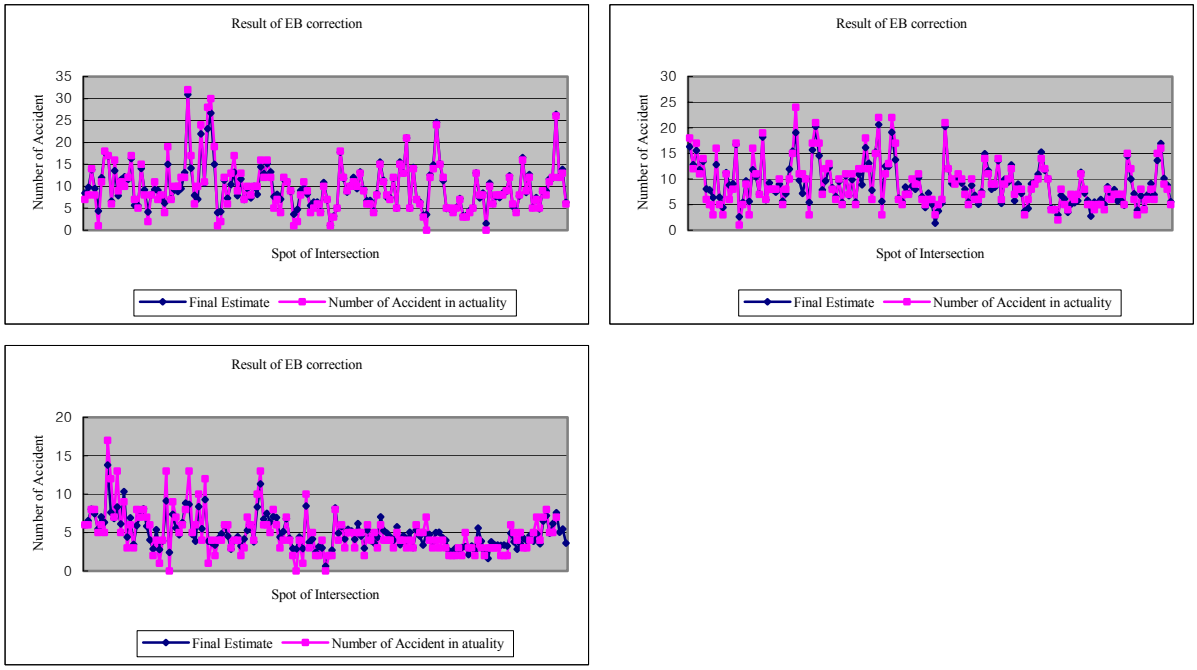

Figure 2. Results of EB correction

5. CONCLUSIONS AND FUTHER STUDY

In this paper, we develop not only a simple Multivariate Model but also the EB method, which can reflects the characteristics of object that are usually unrecognized. The outcome based on this mixed model can be useful to analyze effect of before-and-after for accident prevention policy and to classify between safe and dangerous roads.

Some factors that cause accidents are discrete and continuous, but these factors should be integrated as category variables in order to construct an accident prediction model more than continuous variables, because continuous variables can be changed to discrete variables easily based on reasonable standards, but the reverse is almost impossible. However, linear models or Poisson models are not able to deal with the categorical factors, so in this paper a generalized log linear model, one of categorical data analysis, is applied, and the result is successful.

Furthermore, this model reacts properly on accident occurrences, because best model are selected as each type of accident, and previous prediction models of accidents are almost impossible to explain all types of accidents simultaneously. This trend causes the duplicated investment, but through this research that reflects each type of accidents, every variable for the accident can be measured and duplicating investment can be avoided.

However, accidents can happen in connectors of an intersection as well as the intersection itself, but the condition in the connection part of intersections is not considered in this paper. It is necessary to develop a model that can clarify the relation between intersections and connection parts of intersections in further study. In this research, there is a limit that we use the k-value deduced from preceding researches that is one of important factors in EB method. In further study, the value should be investigated based on more data.

REFERENCES

a) Books and Books chapters

Road Traffic Safety Authority (1996), **Development of Accident Risk Index Model of Intersections**, Korea.

Road Traffic Safety Authority (1996), **Study for analysis and countermeasure of accidents in signalized intersections**, Korea.

Toronto Univ. (Ontario) (1990), **Empirical Bayes Approach to the Estimation of 'Unsafety' : The Multivariate Method,** United States of America.

Do-sup, K. (1993), Regression Analysis for Sociology Sceience, Bobmunsa Publishers, Korea.

Alan Agresti (1990), **Categorical Data Analysis**, A wiley-Interscience Publication.

David Ronald Dickson McGuian, **An Examination Of Relationships Between Road Accidents And Traffic Flow**, Newcastle University Library,

b) Journal papers

F. F. Saccomanno and C. Buyco, Generalized Loglinear Models of Truck Accident Rates, **Transportation Research Record,** Vol. 1172, 23-31.

Shaw-Pin Miaou, Patricia S. Hu, Tommy Wright, Ajay K. Rathi, And Stacy C. Davis, Relationship Between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach, **Transportation Research Record,** Vol. 1376, 10-18.

E. Hauer and B. N. Persaud, How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effects of Warning Devices, **Transportation Research Record**, Vol. 1114, 131-140.

c) Other documents

National Police Agency, Road Traffic Safety Authority (1998), Plan for Basic Countermeasures of risk roads in Seoul.

National Police Agency, Road Traffic Safety Authority (1999), Plan for Basic Countermeasures of risk roads in Seoul.

National Police Agency, Road Traffic Safety Authority (2000), Plan for Basic Countermeasures of risk roads in Seoul.