Relating Freeway Traffic Accidents to Inductive Loop Detector Data Using Logistic Regression

Junhyung Park
Graduate Student Researcher
Dept. of Transportation Systems Engineering
Hanyang University at Ansan
1271, Sa-3 dong, Sangnokgu, Ansan-si
Kyunggi-do, Korea, 426-791
Fax: +82-31-406-6290
E-mail: gaiakail@hanyang.ac.kr

Cheol Oh, Ph.D.
Assistant Professor
Dept. of Transportation Systems Engineering
Hanyang University at Ansan
1271, Sa-3 dong, Sangnokgu, Ansan-si
Kyunggi-do, Korea, 426-791
Fax: +82-31-406-6290
E-mail: cheolo@hanyang.ac.kr

ABSTRACT

This study proposes a methodology for utilizing loop detector data to understand the mechanisms of accident occurrence on freeways. Loop detector data were processed to derive useful traffic parameters representing spatio-temporal variations of traffic conditions, in order to contextualize which lead to accident events. Binary logistic regression modeling was conducted to correlate traffic variables with the accident data that were collected from the Seohaean freeway during the three years, from 2004 to 2006. Outlier filtering and data imputation were also performed to prepare for more reliable traffic variables. The outputs of the model, which are actually probabilistic measures of accident occurrence, can be utilized not only in designing warning information systems but also in evaluating the effectiveness of various traffic operation strategies in terms of traffic safety.

INTRODUCTION

Traffic safety research to prevent accidents and to reduce injury severity on the road is a significant topic in the field of transportation engineering since it aims to save human lives. Huge effort has been expanded to conduct other studies, focused on identifying accident causes and on applying better countermeasures to prevent traffic accidents. Most studies have analyzed historic data, which has been aggregated over a long term period such as annual average daily traffic (AADT) or hourly volume, to derive explanatory variables to model the relationship between traffic safety and traffic conditions. (Zegeer et al., 1988; Vogot and Bared, 1998; Council, 1998; Anderson et al., 1999; Park et al., 2007).

Since the late 1990s, active development and operation of intelligent transportation systems (ITS) related businesses such as the freeway traffic management system (FTMS) has allowed transportation researchers to collect real-time traffic data, enabling new research on traffic safety. The development and use of a model that estimates accident likelihood through the analysis of relationships between traffic accident data and real-time vehicle data from a detector is the representative case. Research in this area has been conducted with the assumption that the disruptive traffic conditions leading to traffic accidents should be identifiable prior to the actual accident occurrence. Figure 1 shows a conceptual illustration of accident occurrence as predicted by traffic conditions.

Oh et al. (2001) reported that the standard deviation of speed data during the five minutes leading up to an accident occurrence is may represent disruptive traffic conditions. The group focused on analyzing the traffic accident and loop data from a detector on the California I-880 freeway. Data collected from the loop detector closest to the traffic accident location was used.

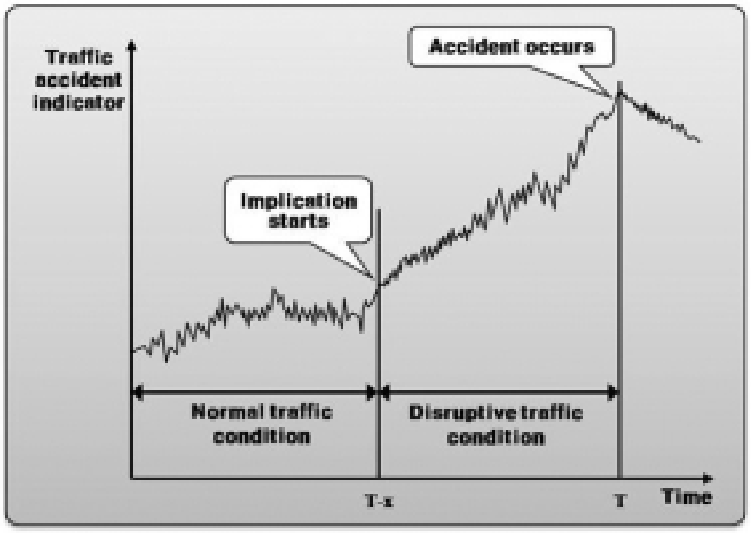

Figure 1: Traffic accident indicator

Lee et al. (2002, 2003) developed a model that predicts accident likelihood by using the traffic accident and loop data from a detector on a freeway in Ontario, Canada over a period of 13 months. They analyzed, the deviation of the speed data collected from the accident location, either upstream and downstream of the detector. Golob et al. (2001) analyzed the effect of 30-second loop data segments from the detector prior to the time of accidents in southern California. They also assessed the effects of weather and lighting conditions on accident occurrence. Abdel-Aty et al. (2004) analyzed data collected in Orlando, Florida and concluded that the average occupancy during the five to 10 minutes prior to the accident as recorded at location upstream of the traffic accident in the context of a downstream detector's Coefficient of variation (CV) for speed data within a five-minute window can predict accident. Moreover, Adel-Aty and Abdalla(2004) analyzed the relationship between real-time traffic data, highway geometric parameters and traffic accidents, using 337 traffic accidents that occurred in 1999. If the change in speed over 15 minutes was high, they reported that the traffic accident likelihood was evaluated 0.5 miles downstream of the applicable detector.

Our study presents a methodology for relating inductive loop detector data with traffic accidents that on the Seohaean freeway in Korea, during 2004~2006. Our major research problem is how to distinguish abnormal traffic conditions that may lead to traffic accidents. A logistic regression technique was usedto model the relationship, which is capable of estimating accident likelihoods. In addition, an example application of the proposed relationship modeled by the logistic regression is introduced and further discussed for the field implementation.

This paper is organized as follows. Chapter 2 describes how to process loop detector data and thereby establish the dataset for our analyses. Chapter 3 introduces binary logistic regression analysis, which is our chosen method for model development. Chapter 4 discusses technical issues to ensure successful field implementation. The last chapter summarizes the major findings and contributions of this study. We also suggest limitations and directions for future research.

Table 1: Accident statistics   Unit : number of cases

Year	Number of traffic accidents	Types of accidents		Daytime vs. nighttime		Direction towards Seoul		Direction towards Mokpo	
		Collision and rear-end collision	Others	Daytime	Nighttime	Daytime	Nighttime	Daytime	Nighttime
2004	348	69	279	224	124	122	72	102	52
2005	311	44	267	200	111	113	61	87	50
2006	280	47	233	199	81	106	37	93	44
Total	939	160	779	623	316	341	170	282	146

DATA PREPARATION

This study classified traffic conditions into accident-free traffic conditions and disruptive (abnormal) traffic conditions that may lead to accidents. In order to derive useful independent variables for modeling purposes, a large amount of loop detector data had to be systematically processed.

Loop Detector Data Processing

We collected Seohaean freeway loop data from detectors and traffic accident data over three years - 2004, 2005 and 2006. As shown in Figure 2, this study used traffic volume, average speed and average occupancy to represent traffic conditions for up to 15 minutes (t-15) before, and traffic conditions between 15 minutes (t-15) and 30 minutes (t-30) before the time of an accident (t). Data were collected from two detectors at neighboring upstream and downstream stations to extract traffic conditions that may contribute to accidents.

As shown in Figure 2, "up" represents upstream of the accident location, and "dn" represents downstream. t1 and t2 signify 15 minutes and 30 minutes before the time of the accident, while "O", "V" and "S" signify occupancy, volume and speed respectively. "D" in front of the O, V and S indicates the value of the difference between the average value of the data from the detector collected during a single day and the value of the data from the detector at each time instant and each postmile. Differences were calculated in terms of the degree of deviation from the day's average value. Figure 3 shows the data processing procedure.

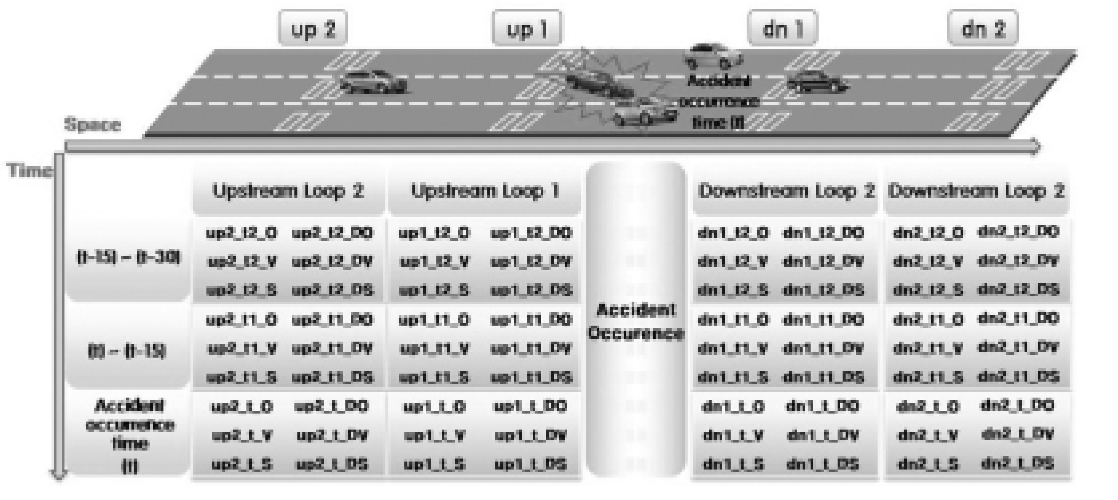

Figure 2: Conceptual illustration for loop detector data processing

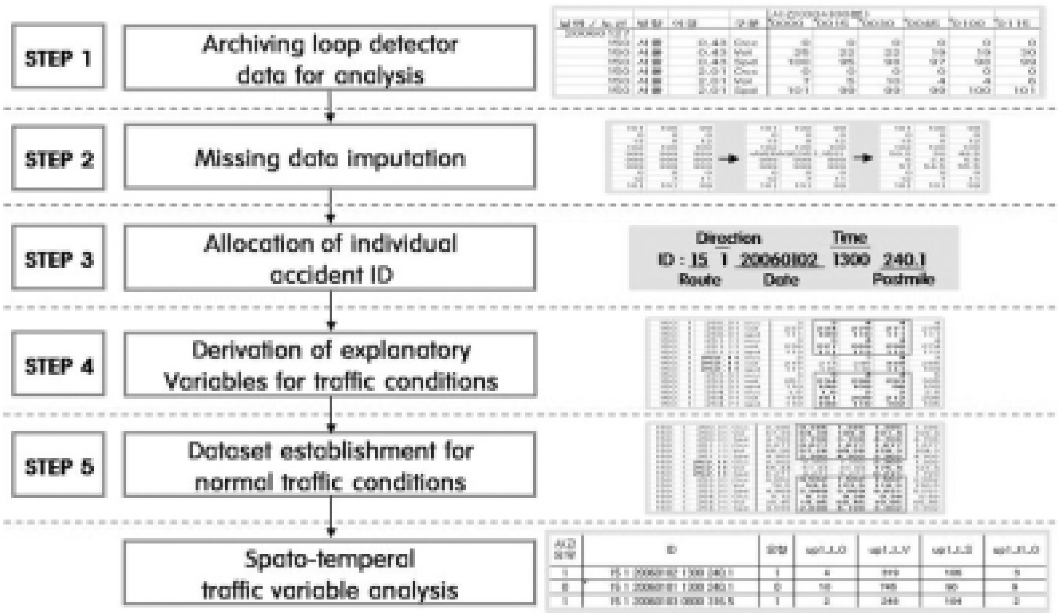

Figure 3 : The proposed loop data processing procedure

[Stage 1] Archiving loop detector data for analysis

This study used traffic accident data obtained from the Seohaean freeway. The first step in the analysis was to archive loop detector data for the Seohaean freeway.

[Stage 2] Imputing missing data

The accuracy and completeness of the loop detector data significantly affects the reliability of the analysis in this study. A neighboring location reference technique was used to impute all missing loop data which were recorded as "-999" in the database. Upstream and downstream data from the detectors were averaged to recreate the missing data. In cases where there were missing data points in the corresponding upstream and downstream datasets, the 15-min before/after data were averaged for imputation.

To evaluate the reliability of our data given missing information, we divided all accidents into three classes(0, 1 and 2) according to the amount of imputation needed. The standard that classified into three cases is as follows.

- Type "0" : complete data without any missing data points
- Type "1" : data imputed by a neighboring reference method
- Type "2" : incomplete data because of too many missing points

[Stage 3] Allocating individual accident IDs

Traffic accident data is used to allocate individual IDs that reflect data such as route, direction, date, time, postmile and others, for each traffic accident. The direction towards Seoul is indicated by "1" while the opposite direction is coded as "0". We used 24-hours time.

[Stage 4] Deriving explanatory variables for hazardous traffic conditions

Once the traffic accident ID had been identified, spatio-temporal traffic variables were extracted from the Seohaean freeway loop data from the detectors. Data from 15 minutes before the time of the accident collected from the two detectors on the rear-side direction, two detectors ahead of the accident, and the traffic volume, speed, and average occupancy data from 15 minutes and 30 minutes before the accident were extracted from the loop detector closest to the accident location. We extracted a total of 72 traffic variables, composed of the difference values pertaining to each traffic variable for the observed value and daily averages from each detector. The difference value is the absolute value of the value that is deducted from the data for each time in the context of the 24 hour averaged data from each loop detector.

[Stage 5] Deriving explanatory variables for normal traffic conditions

Each traffic variable was extracted under accident-free traffic conditions to draw out traffic variables related to accidents, according to the occurrence or non-occurrence of accidents. Seventy-two traffic variables were extracted, up until one day before the accident occurred. When an accident had occurred on the day before the accident under normal traffic conditions, the data from two days before the accident was extracted. Accident data is indicated by "1", and the data from the day before the accident is annotated "0".

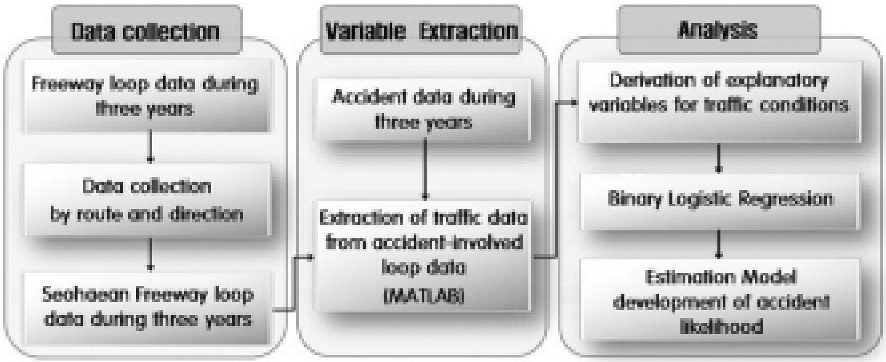

Figure 4 : Overall procedure for data analysis and model development

METHODOLOGY AND MODEL DEVELOPMENT

The overall procedure for data processing and model development is presented in Figure 4.

Binary Logistic Regression Analysis

Binary logistic regression can be applied to our binary classification problem if the traffic conditions that can lead to accidents are a dependent variable of the model set equal to 1, and the accident- free traffic conditions are identified by 0.

This study applied the binary logistic regression technique to the loop data from detectors based on an accident likelihood probability model. As the output of the model is a probabilistic measure between 0 and 1, the accident likelihood can be directly predicted. The dependent variable "1" signifies that the probability of an accident is "1". Accident probability may assume diverse values within the range of 0 to 1, but it cannot be outside this range. Logistic regression assumes that the relationship between the independent variable and the dependent variable between 0 and 1 has an S shape, as shown in Figure 5. The accident likelihood estimation function is shown in Equation 1.

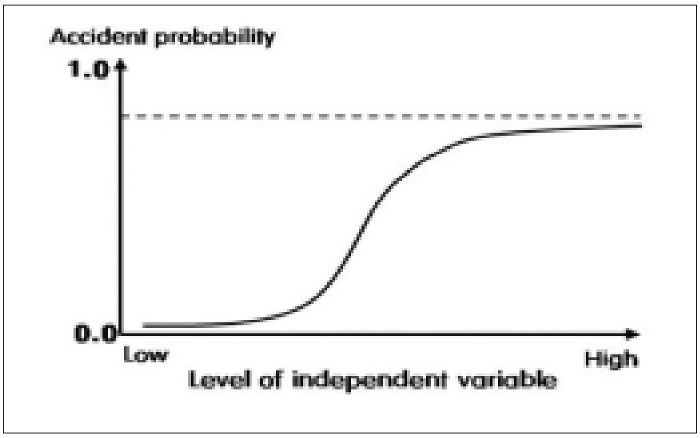

Figure 5 : Logistic relationship between independent variable and traffic

$$\Pr(ACC_i = 1 \mid X_i) = \frac{\exp[f(X_i, \beta)]}{1 + \exp[f(X_i, \beta)]} \quad (1)$$

- $Pr(ACC_i)$: Accident likelihood according to the given traffic variable i
- X_i : Traffic variable that affects traffic accident occurrence
- $f(X_i, \beta)$: Function comprised of X_i and parameter β

Table 2 : Binary Logistic Regression R-Square Data

CASE	-2 Log likelihood	Cox & Snell R-square	Nagelkerke R-square	Number of traffic accident cases
Case 1 : accident traffic condition and traffic conditions the day before the accident	2304.647	0.017	0.022	939 cases
Case 2 : accident occurring during daytime and traffic conditions the day before the accident	1505.103	0.037	0.049	623 cases
Case 3 : accident occurring at night and traffic conditions the day before the accident	786.013	0.000	0.000	316 cases
Case 4 : collision and rear-end collision and traffic conditions the day before the accident	264.329	0.226	0.301	142 cases
Case 5 : other accident types and traffic conditions the day before the accident	1904.125	0.018	0.024	779 cases
Case 6 : daytime accident in the direction towards Seoul. Non-accident traffic condition	792.243	0.053	0.070	341 cases
Case 7 : daytime accident in the direction towards Mokpo. Non-accident traffic condition	697.217	0.046	0.062	282 cases
Case 8 : nighttime accident in the direction towards Seoul. Non-accident traffic condition	153.471	0.254	0.340	113 cases
Case 9 : nighttime accident in the direction towards Mokpo Non-accident traffic condition	325.187	0.173	0.231	146 cases
Case 10 : when there is no missing data or when impeccable imputation was possible	1692.932	0.026	0.035	666 cases
Case 11 : when there was extensive missing data such that Imputation failed	587.525	0.035	0.047	273 cases

Model Development Results

This research developed a model for the 11 cases shown in Table 2. Each case is categorized according to the data filtering method, and also according to the degree of loss based on the time of day (whether daytime or nighttime), accident type, direction, and amount of missing data. The R-square value describes the effectiveness of the regression mapping.: the legitimacy increase as this value increases. When conducting our binary logistic regressions, the regression equation was modified to uses only the variables that were related to when the accident occurs. The R-square value of the regression equation was measured according to the binary logistic regression results, as shown in Table 2. Case 8 (nighttime direction towards Seoul) revealed the highest metric, with an R-square of 0.340. Case 4 (collision and rear-end collision) was the second-highest at 0.301. This demonstrates that the variables related to Case 4 and Case 8 are more significant in predicting accident occurrence than are the accident elements of other cases, in terms of influence power. Nighttime collisions and the traffic conditions in the direction towards Seoul are more closely linked to accidents.However, the night traffic environment was unrelated to accident rates, as shown by Case 3's R-square value (0.000). We note that Case 2 and Case 3 have R-square values of 0.049 and 0.000. Thus, there is a difference between daytime and nighttime data. However, when we classify daytime and nighttime with direction at the same time, R-square values for daytime Cases 6 and 7 are 0.07 and 0.06 respectively, and the R-square values for night Cases 8 and 9 are 0.34 and 0.23, each. Thus, R-square values at night were more correlated with accidents than daytime scenarios for both travel directions.

Table 3 : Classification performance

CASE	Observed		Predicted								
			Accident occurrence		Percent correct (%)	Accident occurrence		Percent correct(%)	Accident occurrence		Percent correct(%)
			0	1	0.5	0	1	0.7	0	1	0.8
Case 4	Question of accident	0	89	25	78.070	108	6	94.737	112	2	98.246
		1	37	83	69.167	77	43	35.833	82	38	31.667
	Overall Percentage %				73.504			64.530			64.103
Case 8	Question of accident	0	62	17	78.481	93	9	91.176	98	4	96.078
		1	25	38	60.317	66	40	37.736	82	24	22.642
	Overall Percentage %				70.423			63.942			58.654

Table 4 : Modeling results

CASE	Variable	Beta	S.E.	Wald	Sig.	Exp(B)
case 4 collision and rear-end collision	up1_t1_O	0.63607	0.29560	4.63014	0.03142	1.88905
	up1_t2_V	-0.00701	0.00336	4.35420	0.03692	0.99302
	up2_t1_O	-0.54779	0.24958	4.81708	0.02818	0.57823
	up2_t2_V	0.00671	0.00286	5.49887	0.01903	1.00673
	up1_t1_DO	-1.23492	0.48501	6.48305	0.01089	0.29086
	up1_t2_DO	1.11977	0.48792	5.26702	0.02173	3.06416
	up2_t1_DS	0.06987	0.03676	3.61161	0.05738	1.07236
	dn2_t1_DV	0.00816	0.00279	8.54909	0.00346	1.00820
	dn2_t1_DS	0.20489	0.05862	12.21552	0.00047	1.22739
	dn2_t2_DS	-0.12514	0.05715	4.79472	0.02855	0.88238
	Constant	-1.13773	0.29770	14.60587	0.00013	0.32054
case 8 Seoul direction at night	up1_t1_O	1.89637	0.83917	5.10678	0.02383	6.66169
	up1_t2_O	-3.14948	0.94933	11.00636	0.00091	0.04287
	up2_t2_O	1.33663	0.60218	4.92682	0.02644	3.80620
	dn1_t2_O	-1.41153	0.53297	7.01414	0.00809	0.24377
	dn1_t2_V	0.02321	0.00860	7.28865	0.00694	1.02349
	dn1_t2_S	-0.14279	0.05124	7.76556	0.00533	0.86694
	up1_t1_DO	-2.72418	1.28526	4.49253	0.03404	0.06560
	up1_t1_DV	-0.05901	0.02548	5.36282	0.02057	0.94270
	up1_t2_DO	2.73222	1.28032	4.55401	0.03284	15.36690
	up1_t2_DV	0.06750	0.02528	7.13085	0.00758	1.06983
	up2_t1_DV	0.09030	0.02822	10.24061	0.00137	1.09450
	up2_t2_DV	-0.08813	0.02751	10.26567	0.00136	0.91564
	dn1_t2_DV	-0.02210	0.01122	3.87991	0.04887	0.97814
	dn2_t1_DO	-1.57296	0.65590	5.75126	0.01648	0.20743
	dn2_t2_DV	0.02428	0.01337	3.30142	0.06922	1.02458
	Constant	13.90168	5.13050	7.34200	0.00674	1089986

The classification chart differentiated the classification cutoff value at the time of the regression to carry out the analysis, and organized the cases that yielded significant results in Table 3. Classification cutoff valueswere set at 0.5, 0.7 and 0.8 at the time of regression. The results show that the correct classification rate decreased when the classification cutoff value was set high, as shown in Table 3. In Case 4, a correct classification rate of73.5% resulted when the classification cutoff value was set to 0.5. However, when the classification cutoff value was set to 0.8, the correct classification rate was 64.1%. The results of the significant variables for the 11 cases are indicated in Table 4 as a result of binary logistic regression for the data from detectors at the traffic accident location over the three years from 2004 to 2006. In Case 4 (collisions and rear-end collisions), six difference related variables and variables pertaining to up1_t1_O, up1_t2_V, up2_t1_O and up2_t2_V were statistically significant. Variables that can indicate collision and rear-end collision accident likelihood are graphed in <Diagram 6>. In Case 8, 15 variables were calculated, and the variables, up1_t1_O, up1_t2_O, up2_t2_O, dn2_t2_O, dn1_t2_V, dn1_t2_S that included difference related variables were selected. Among the results of regression analysis on Case 4 and Case 8, four variables were calculated with up1_t1_O, up1_t1_DO, up1_t2_DO, up2_t1_DV.

APPLICATION AND TECHNICAL ISSUES

Our model can be used to provide warning information to drivers about traffic conditions where there is a high risk of accidents. Moreover, the results can be applied to traffic management strategies such as variable speed limit scenarios (VSL). For example, it is possible to consider a so called real-time traffic safety management system that expresses the variable speed limit (VSL) along with appropriate warning information if the accident likelihood is above a specific level and when the accident likelihood can be continuously recalculated while monitoring traffic conditions that are collected in real-time at the traffic management center.

Various technical issues need to be resolved to implement such a system. The traffic condition variables extracted in this research using data from detectors aggregated over 15 minutes may not be sensitive enough to dynamic real-time traffic changes. In order to react more sensitively to changes in the transportation environment, it is necessary to make this time slice shorter. As the time slice becomes shorter, the scope of the average value used to calculate the difference value can be applied in more diverse ways as well.

Table 3 showed how the probability of accident and non-accident occurrence prediction changes as we adjust the classification cutoff value. In Case 4, the correct classification rate decreases from 73% to 64% when the classification cutoff value is increased from 0.5 to 0.8. This suggests that the probability of detecting accident-prone traffic conditions using this likelihood estimation model decreases. Among the many traffic accident conditions, adjusting the classification cutoff value can help adjust the frequency to better express warning messages after detecting only the most dangerous traffic conditions. More research is required on dynamic adjustment of the parameters and speed limit.

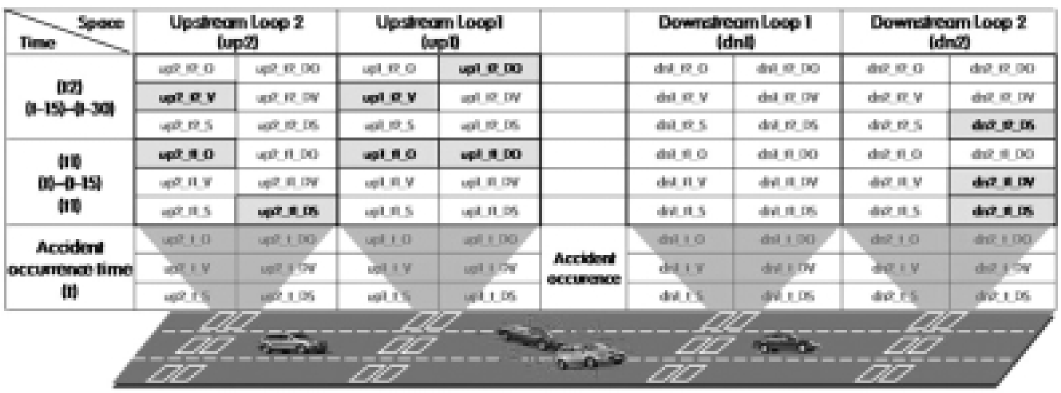

Figure 6 : Traffic variables related to accident likelihood in the case of collisions and rear-end collisions (Case 4)

CONCLUSION

We assessed data collected over three years, from 2004 to 2006, from detectors installed on the Seohaean freeway in light of accident occurrences. Our goal was to identify correlations. The data was subjected to the binary logistic regression to measure the probability of accident likelihood. The results show that collisions, rear-end collisions, and accidents in the direction towards Seoul at night are significantly correlated with traffic accidents. In the case of collisions and rear-end collisions, R-square was 0.301 with a correct classification rate of 73.5%. In the case of the direction towards Seoul at night, R-square was 0.340, and the correct classification rate was 70.4%.

As traffic conditions change, accident likelihood information should be given to drivers so that they can understand the risk associated with current traffic conditions and, focus on decelerating or defensive driving. If the risk level of the accident is greater than a threshold, a warning should be given and the speed limit should changed. However, research regarding a system that can decrease speed appropriately to suit the transportation environment is necessary. In relation to the application of a variable speed limit (VSL), research on the reduction time and changing bounds is particularly important.

In order to ensure that the model developed in this research can be used in the field later, there are various technical issues to resolve. First, we need a classification cutoff value that serves as the standard for classifying whether a model is appropriate given the accident hazardous conditions. Application times and frequencies that are judged to provide correct classification rate should be further explored.

Second, the model must react sensitively to the traffic conditions. Traffic accidents occur during a very short period of time due to various direct and indirect elements. In order to survey changes taking place in these instantaneous and dynamic elements in a more precise manner, it is recommended to use raw data collected over shorter intervals than the 15 minutes that we used in this study. This may have implications for the storage and management of historical data at traffic management centers.

Significant results were obtained from our binary logistic regression for collisions, rear-end collisions and in the direction towards Seoul at night. However, it was impossible to provide a warning message that would convey the likelihood of various accident types. This requires analysis of more data from other detectors, and researchers should further pursue other correlations with accidents. Moreover, evaluation is required to study how the accident rates decrease when the alerts system is actually deployed.

The real-time accident likelihood estimation model proposed by this research encourages drivers to engage in defensive driving and to become aware of accidents, so that spontaneous traffic accidents can decrease.A guidance system is required so that the drivers can become safer on the road.

REFERENCES

1. Abdel-Aty, Nizam Uddin, Anurag Pande, M. Fathy Abdalla, and Liang Hsia, Predicting Freeway Crash from Loop Data from detector by Matched Case -Control Logistic Regression. In Transportation Research Record: Journal of Transportation Research Board, No.1897, TRB, National Research Council, Washington D.C,2004, pp. 88-95.
2. Abdel-Atry, M., and Pande, A., Classification of real-time traffic speed patterns to predict crashes on freeways. Presented at the 83rd annual Meeting of the Transportation Research Board(TRB), Washington D.C., 2004
3. Anderson, I. B., Bauer, K. M., Harwood, D. W., and Fitzpatrick, K. Relationship to Safety of Geometric Design Consistency Measures for Rural Two-lane Highways. Transportation Research Record: Journal of Transportation Research Board, No.1658, TRB, National Research Council, Washington D.C., 1999, pp 43-51
4. Council, F. M. Safety Benefits of Spiral Transitions on Horizontal Curves on Two-lane Rural Roads. Transportation Research Record: Journal of Transportation Research Board, No.1635, TRB, National Research Council, Washington D.C., 1998, pp 10-17
5. Golob, T., and W. Recker. Relationship Among Urban Freeway Accidents, Traffic Flow, Weather, and Lighting Conditions. Institute of Transportation Studies, University of California, Berkeley, 2001
6. Gwynn, D. W. Relationship of Accident Rate and Accident Involvements with Hourly Volumes. Traffic Q.,1967 , July pp. 407-418
7. Lee, C., F. Saccomanno, and B. Hellinga. Analysis of Crash Precursors on Instrumented Freeways. In Transportation Research Record: Journal of Transportation Research Board, No.1784, TRB, National Research Council, Washington D.C., 2002, pp. 1-8
8. Lee, C., B. Hellinga, and F. Saccomanno. Real-time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. In Transportation Research Record: Journal of Transportation Research Board, No.1840, TRB, National Research Council, Washington D.C., 2003, pp. 67-77
9. Oh, C., Chang, J., and Chang, M., Relationship between V/C and accident rate for freeway facility section. Journal of Transportation Research Society of Korea, Vol. 17, No. 2, 1999, pp 21-27
10. Oh, C., J.-S. Oh, S. G. Ritchie, and M.-S. Chang., Real-Time Estimation of Freeway Accident Likelihood.. Presented at 80th Annual Meeting of the Transportation Research Board, Washington, D.C., 2001
11. Park Hyo-Sin et al., Development of Accident Prediction Models for Freeway Interchange Ramps. Journal of Transportation Research Society of Korea, Vol. 25, No. 3, 2007, pp 123-135.
12. Vogot, A., and Bared, J. Accident Models for Two-lane Rural Segments and Intersections. Transportation Research Record: Journal of Transportation Research Board, No.1635, TRB, National Research Council, Washington D.C., 1998, pp 18-29
13. Zegeer, C.V., Reinfurt, D. W., Hummer, J., Herf, L., and Hunter, W. Safety Effects of Cross-section Design for Two-lane Roads. Transportation Research Record: Journal of Transportation Research Board, No.1195, TRB, National Research Council, Washington D.C., 1988, pp 20-32