Incident Duration Prediction Based on Latent Gaussian Naive Bayesian classifier

Dawei LI **Lin CHENG**[*]
School of Transportation, Southeast University
Nanjing, 210096, P. R. China;

Jiangshan MA
Department of Civil and Environmental Engineering, Tokyo Institute of Technology,
Tokyo, 152-8552, Japan

Received: 27-02-2011
Accepted: 20-04-2011

Abstract

The probability distribution of duration is a critical input for predicting the potential impact of traffic incidents. Most of the previous duration prediction models are discrete, which divide duration into several intervals. However, sometimes the continuous probability distribution is needed. Therefore a continuous model based on latent Gaussian naive Bayesian (LGNB) classifier is developed in this paper, assuming duration fits a lognormal distribution. The model is calibrated and tested by incident records from the Georgia Department of Transportation. The results show that LGNB can describe the continuous probability distribution of duration well. According to the evidence sensitivity analysis of LGNB, the four classes of incidents classified by LGNB can be interpreted by the level of severity and complexity.

Keywords: incident management; incident duration; Bayesian networks; LGNB classifier

1. Introduction

Highway incidents are a major cause of traffic congestion and delay. Some studies have estimated that around 60% of all traffic congestion on highways is caused by incidents [1]. There are several methods for estimating traffic delay caused by incidents in the literature. The deterministic queuing model [2,3] and traffic shock wave theory [4,5] are two major methods used in the delay estimation and prediction.

It should be noted that all of the aforementioned models need incident duration as a critical input. Previous study based on dynamic traffic assignment also shows that the network wide incident delay is sensitive to the incident duration [6].

Incident duration is the time period from the occurrence to the clearance of an incident, Based on the definition, it has three elements, i.e., detection, response, and clearance.

Recent researches have applied a variety of techniques to analyze the incident duration. Many studies have been carried out to estimate the statistical distribution of incident durations. Golob (1987) analyzed the duration of incidents involving large trucks, and demonstrated that the accident duration fits a lognormal distribution [7]. Giuliano (1989) organized the incidents to several categories and found that the distributions of incident duration of most incident categories are lognormal distributed. Some studies also developed models to predict incident duration [8]. Sullivan (1997) developed a model, IMPACT to predict the incident occurrence and the associated delays assuming that the incident duration fits a lognormal distribution [9]. Some studies applied several variations of linear regression models by

[*] Corresponding author: gist@seu.edu.cn

Published by Atlantis Press

treating characteristics such as incident type, weather condition, and number of vehicles and lanes involved as independent variables [10,11]. Ozbay (1999) constructed decision trees which do not require knowledge of all observable incident characteristics 12[12]. Nam (2000) used hazard-based models which provide information not only on the total incident duration, but also on the probability that an incident will be cleared in the next small time interval, known to have already existed for a certain period of time [13]. Smith (2002) also suggested using the nonparametric regression, and estimated incident duration based on similar incidents in the past [14]. It should be noted that almost all incident duration prediction models except decision trees, require complete knowledge of incident characteristics that chosen for duration prediction. However, this knowledge is often incomplete in practice. Techniques exist to solve this problem based on interpolated values of unknown independent variables [15], but this reduces accuracy and complicates the prediction process. Decision tree models do not suffer from this limitation, but these models are deterministic and do not give the reliability of prediction. So some researchers developed some probabilistic models for duration prediction based on Bayesian inference [15, 16]. However, most of the prediction models only output one probable value or one interval of duration, rather than the continuous probability distribution in the feasible region, which will be more useful in delay uncertainty analysis and travel time reliability analysis on the real time.

The primary objective of this study is to develop a continuous model based on latent Gaussian naive Bayesian (LGNB) classifier, using data from incident logs maintained by the Georgia Department of Transportation.

2. Methodology

Formally, a Bayesian network for a set of random variables ***U*= {**X_1,…, X_n**}**. is a pair $B=(G,\Theta)$. The first component, *G*, is a directed acyclic graph whose vertices correspond to the random variables X_1, …, X_n, and whose edges represent direct dependencies between the variables. The graph *G* encodes independence assumption: each variable X_i is independent of its nondescendants given its parents in *G*. The second component of the pair, namely, represents the set of parameters that quantifies the network. It contains a parameter $\theta_{X_i|\Pi_{x_i}} = P_B(x_i \mid \Pi_{X_i})$ for each possible value x_i of X_i, and Π_{x_i} of Π_{X_i} , where Π_{X_i} denotes the set of parents of X_i in *G*. A Bayesian network *B* defines a unique joint probability distribution over ***U*** given by

$$P_B(X_1,...,X_n) = \prod_{i=1}^{n} P_B(X_i \mid \Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{X_i}} \qquad (1)$$

When we use Bayesian networks as classifiers, and the class variable is *C*, the other attribute variables are A_1… A_n. The aim is to find the most likely value of *C*, given the information about attribute variables, that is

$$i^* = \arg\max P(C = c_i \mid A_1,...A_n) \qquad (2)$$

Bayesian networks have been widely used in different fields because of the ability of uncertain inference [17]. As a special instance of Bayesian networks, LGNB classifier for duration prediction can be shown in Fig. 1.

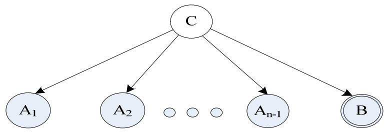


Fig. 1. The structure of LGNB classifier.

In this Bayesian network, *C* is a latent class variable. The possible values of *C* represent the classes that an incident belongs to. The count of classes can be learned from the data. It is assumed that incidents duration fits a lognormal distribution, so Ln(*Duration*) fits a Gaussian distribution, and it can be represented as a continuous node such as node B in Fig.1). There is not a conditional probability table as the discrete models, but conditional Gaussian distribution of B for each possible value of C, that is

$$L(B \mid C = c_i) = N(\mu_i, \sigma_i) \qquad (3)$$

With this form of Bayesian networks, we can get the possible distribution of each incident. Then we can also get the information about the probability of an incident, known to have already existed for a certain period of time [18]. If ln(*Duration*) is $N(\mu,\sigma)$ with density function noted as $f(x)$, the probability distribution function (PDF) of the incident duration when the incident already exists T_0 can be obtained by applying Bayesian theory:

$$f^{'}(x) = k \cdot L(x) \cdot f(x) \qquad (4)$$

Where $L(x) = \begin{cases} 0 & if \;\; x \le T_0 \\ 1 & if \;\; x > T_0 \end{cases}$

k is a constant defined as follows:

$$k = [1 - \Phi(\frac{\ln(T_0) - \mu}{\sigma})]^{-1} \quad (5)$$

There are many algorithms can be deployed to learn the parameters of discrete Bayesian networks, such as MLE algorithm and EM algorithm. However, it is not a easy thing to learn the parameters of a continuous Bayesian network. In this paper, an approximate method is proposed to learn the parameters of LGNB classifier: At first, the continuous variable is discreted with small interval, and the LGNB classifier is transformed to a NB classifier. Using MLE algorithm or EM algorithm, the parameters of this NB classifier can be determined. At last, the CPT of discreted continuous vartiable can be fitted into Gaussian distribution.

Table 1. Summary of incidents used to calibrate and validate models.

	Calibration Set	Validation Set
Number of incidents	1470	1503
Median duration (minutes)	52	47
Standard deviation of duration (minutes)	161	142
Incidents less than 30 minutes	0.32	0.30
Incidents at least 30 minutes, but less than 60 minutes	0.26	0.28
Incidents at least 60 minutes, but less than 90 minutes	0.13	0.17
Incidents at least 90 minutes, but less than 120 minutes	0.12	0.09
Incidents at least 120 minutes	0.17	0.16

3. Data Discription

The data we used in this study was also used by (Boyles et al. 2007). The incidents logs were maintained by the Georgia Department of Transportation. These logs contain a list of incidents occurring in the Atlanta metropolitan area, including the type, start and end times, the number of various types of vehicles involved, the affected lanes, the geographic location, and flags to indicate the presence of assorted types of damage.

Table 2. Description of variables.

ID	Variable Name	Description
1	WeekDay	1=Weekday; 2=Weekend
2	Day	1=Day; 2= Night
3	IncidentType	1=Accident; 2=Stall;3=Debris; 4=Other
4	DetectionType	1=Call Report; 2=Operator detected;3=Other; 4=Unknown
5	LocationType	1=Freeway; 2=Ramp;3=Intersection; 4=Arterial; 5=Other
6	AffectedLaneType	1=Lanes; 2=Off road;3=Shoulder; 4=Core area; 5= None
7	NumLanesAffected	1=None; 2=One lane;3=More than two lanes
8	FirePresence	1=Not; 2=Presence
9	HazMatPresence	1=Presence; 2=Not
10	InjuryCount	1=None; 2=One; 3=Two; 4=Three;5=More than four
11	FatalCount	1=None; 2=More than one
12	NeedPolice	1=No; 2=Yes
13	NeedHero	1=No; 2=Yes
14	NeedTruckWrecker	1=No; 2=Yes
15	SignalDamage	1=No; 2=Yes
16	GuardRailDamage	1=No; 2=Yes
17	AutomobileCount	1=None; 2=One; 3=Two; 4=Three; 5=More than four
18	MotorcycleCount	1=None;2=More than one
19	PickupVanCount	1=None;2=More than one
20	MotorhomeBusCount	1=None;2=More than one
21	LightTruckCount	1=None;2=More than one
22	TractorTrailerCount	1=None;2=More than one
24	Ln(Duration)	Continuous variable
25	C	Latent class variable in LGNB

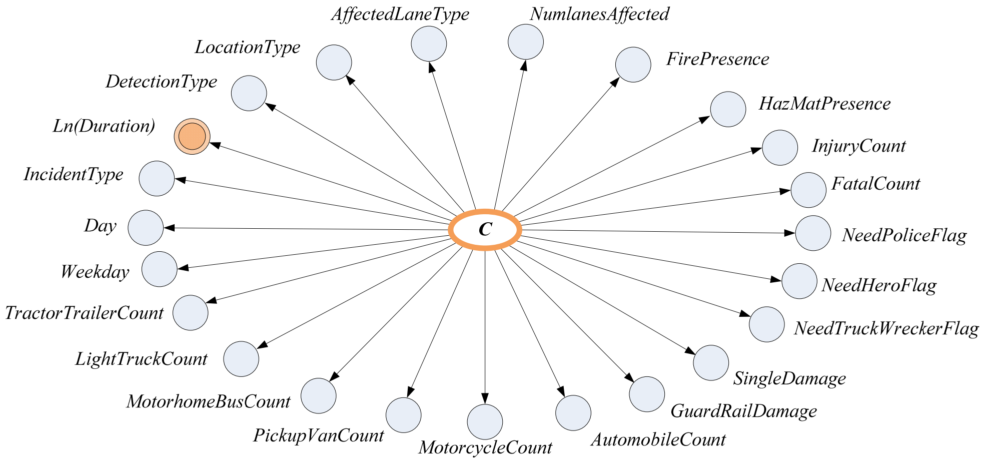


Fig. 2. The structure of LGNB classifier for duration prediction.

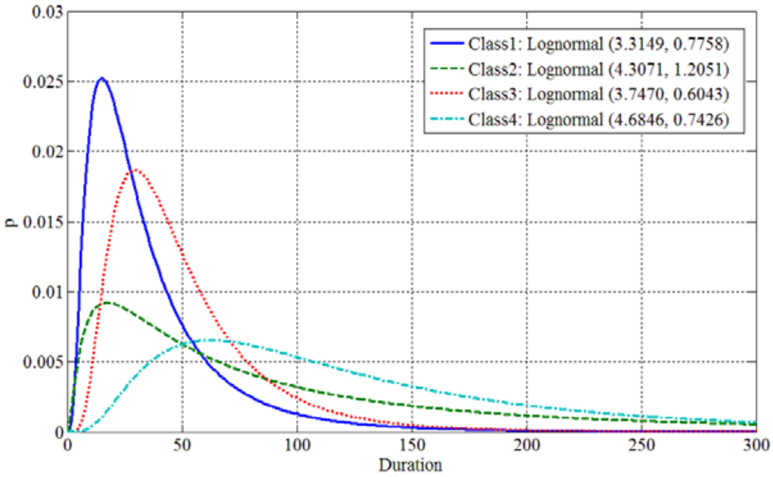


Fig. 3. The distribution of Duration for each class.

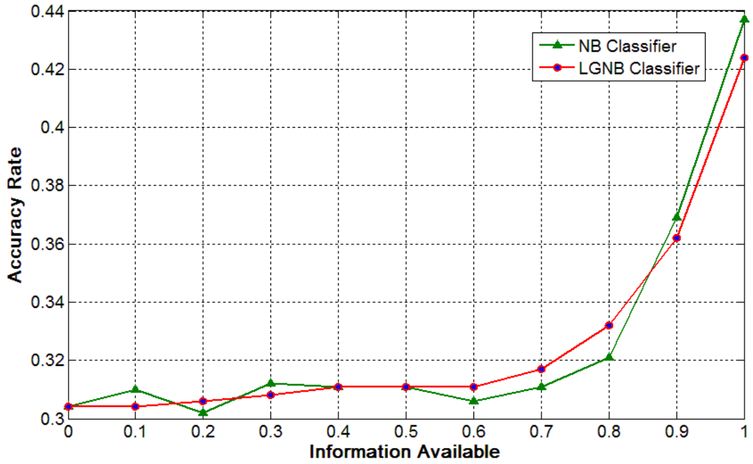


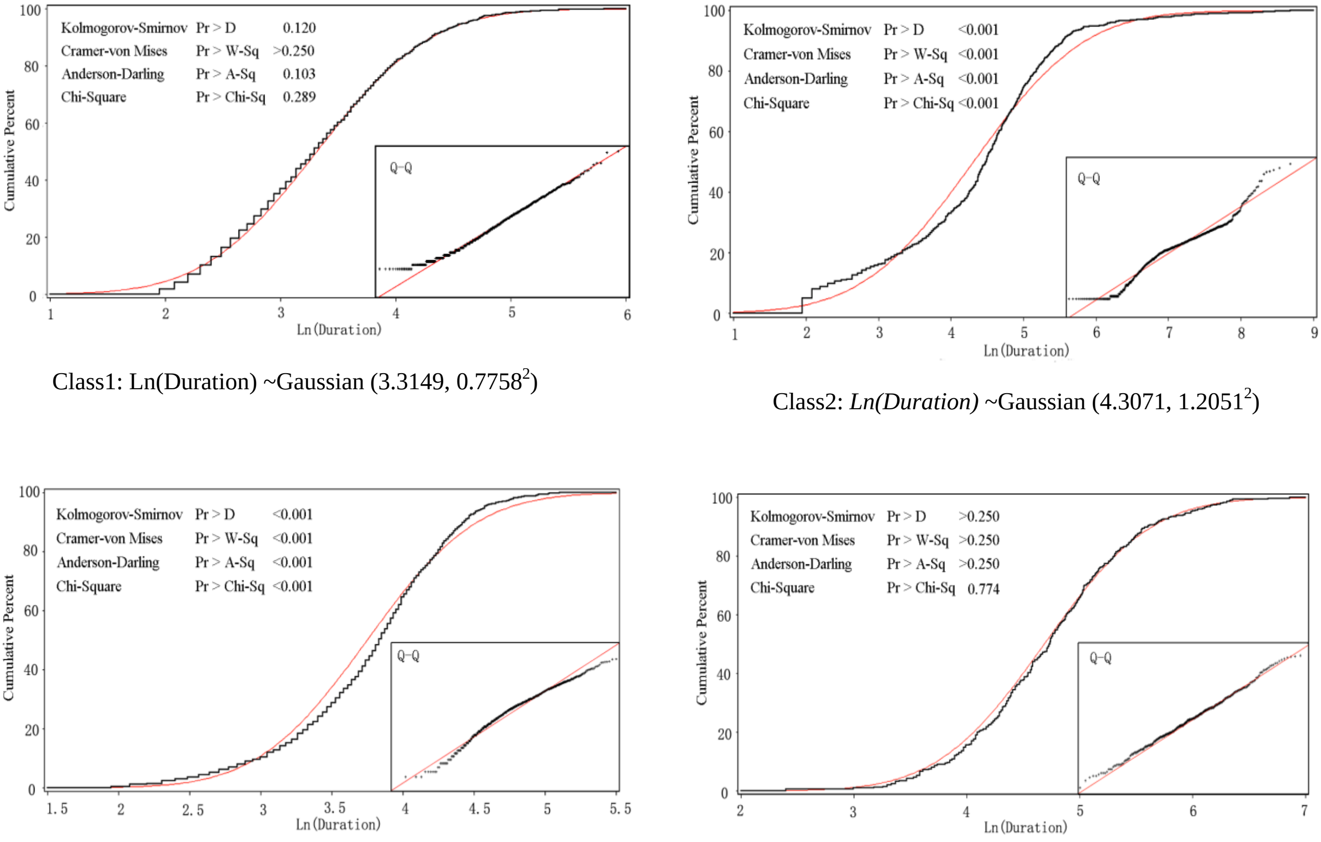
Fig. 4. Accuracy evaluation of LGNB classifier.

The incident database is quite extensive, but it has a limitation that there is no data-field that gives the exact occurrence time of the incident. We can only get the time-stamp that the operator first input an incident in the database. On the other hand, the average detection time as estimated by the Navigator system in the survey is 7 occurred 7 minutes before the logged time.

Incidents occurring in January and February 2004 were used in this research. In the original incident logs, any scheduled incidents such as construction-related closures were excluded from this analysis. The remaining incidents were randomly divided into two groups, the first group (Calibration Set) was used to train the GLNB classifier, and then tested on incidents in the second group(Validation Set). Table 1. contains selected descriptive statistics for the incidents found in these two sets.

4. Model Development

The structure of LGNB is given as Fig. 2. The variables contained in the classifiers are described in Table 2. The parameters can be learned from data using Kevin Murphy's Bayes Net Toolbox (BNT) for Matlab [19]. Because the class node is latent, we need to determine its size (the number of possible values it can obtain). More possible values, more complicated the classifier will be. So we set its size equal 1 first, and then increase



Class1: Ln(Duration) ~Gaussian (3.3149, 0.7758^2)

Class2: *Ln(Duration)* ~Gaussian (4.3071, 1.2051^2)

Class3: *Ln(Duration)* ~Gaussian (3.7470, 0.6043^2)

Class4: *Ln(Duration)* ~Gaussian (4.6846, 0.7426^2)

Fig. 5. Distributions fit test of LGNB.

it by 1, until the expected loglikelihood of the model doesn't increase significantly. At last we set the size equal 4. When the probability distribution of Ln(*Duration*) is known, we can get the distribution of *Duration* for each class as shown in Fig. 3.

5. Model Evaluation And Analysis

5.1. *Accuracy evaluation*

To evaluate the prediction accuracy of LGNB classifier, we compare it with NB classifier[15]. The models were run repeatedly on the incidents in the validation set, with varying amounts of data which was made available for the models. For example, in one run, only ten percent of the information of cases (randomly chosen) was used to classify the incidents; this corresponds to a real-world situation in which only a small amount of information is known about the incident. In order to compare accuracy with discrete models, the continuous variable Duration in LGNB classifier is divided to 5 discrete intervals as the NB classifier. The CPT of duration is computed according to the continuous probability distribution shown in Figure 3. The frequency of correct classification is noted in Figure 4. We can see that when no information is available, both classifiers predict duration using the prior margin distribution, and consider the duration of every incident is less than 30min, which is the largest portion of the training set, and the accuracy is 30%. When more information is available, more incidents are classified correctly. It can be found that LGNB classifier has the same performance as NB classifier, which is proposed and compared to a standard linear regression by Boyles (2007) [15]. However, the purpose of LGNB is to classify incidents according to the fitted continuous probability distribution but not the length of duration. Therefore, the distribution fit test is a more appropriate method to evaluate the performance of LGNB.

5.2. *Distribution fit test of LGNB*

In the following part of this paper, we will test the assumption that Ln(Duration) fits a Gaussian distribution and evaluate the classification performance of LGNB. With LGNB classifier we divided the Validation Set into 4 classes. Class 1 contains 482 cases, Class 2 contains 417 cases, Class 3 contains 411 cases, and Class 4 contains 193 cases. We do distribution fit test for each class, and Figure 5 shows the test result. It can be found that all of the four classes approximately fit the Gaussian distribution which calibrated in last section, though Class 2 and Class3 are rejected in the hypothesis tests when the confidence level is 0.05.

5.3. *Sensitivity analysis of LGNB*

The LGNB has divided all the incidents into four classes, in this part we will analyzed the characters of incidents in each class. By conducting an evidences sensitivity analysis, the effect of the information about other variables on the latent class variable can be measured.

Evidence sensitivity analysis may, for instance, give answers to questions like what are the minimum and maximum beliefs produced by observing a variable, which evidence acts in favor of or against a hypothesis, which evidence discriminates one hypothesis from an alternative hypothesis, and what if a certain observed variable had been observed to a value different from the actual value? Knowing the answers to these and similar questions may help to explain and understand the conclusions reached by the model as a result of probabilistic inference. It will also help to understand the impact of subsets of the evidence on a certain hypothesis and alternative hypotheses.

The sensitivity analysis report of LGNB classifier is shown in Table 3. This table shows the maximum and the minimum posterior probability of the latent class node due to certain evidences, which are entered in the network. For instance, the post probability table of *C* will be updated when each probable value of *IncidentType* is entered as an evidence. when the value "1" is entered as an evidence, the likelihood that this incident belongs to Class 3 will increase greatest with 29% (the meaning of "+29" in Table 3), which results in a posterior probability of 62%. When for the variable *IncidentType* the value "3" is entered as an evidence, the likelihood that this incident belongs to Class 3 will decrease greatest with 26% (the meaning of "-29" in Table 3). Table 3 only shows the variables affect the class variable significantly (increase or decrease the probability more than 1 percent), and the others are ommited.

We can find from the sensitivity analysis that only half of the variables affect the class variable significantly. We also can get some other findings from sensitivity analysis.

The type of incident is the major factor that affects the classifier of incidents. A stall incident has a large probability belongs to Class 1, and if the type of an incident is “others”, the probability that it belongs to Class 2 will increase. An accident most likely belongs to Class 3 or Class 4.

Location Type is also an important factor that affects the classifier of incidents. An incident occurring on the highway more likely belongs to Class 3 or Class 4. If an incident occurs on the arterial, the probability that it belongs to Class 1 or Class 2 will increase.

Table 3. Sensitivity Analysis of LGNB

	VarID	3	4	5	6	15	16	17	18	19	20	21
Class 1	Min(%)	-32	-24	-33	-33	-33	-33	-27	-20	-2	-1	-1
	value min	4	4	5	2	2	2	5	2	1	1	1
	Max(%)	+67	+17	+18	+46	+1	+1	+27	+1	+29	+1	+10
	value max	2	1	1	4	1	1	2	1	2	2	2
Class 2	Min(%)	-28	-27	-28	-28	-1	-6	-28	-28	-28	-28	-28
	value min	2	3	5	4	1	2	2	2	2	2	2
	Max(%)	+70	+17	+66	+47	+66	+1	+22	+1	+3	+1	+1
	value max	4	2	4	2	2	1	1	1	1	1	1
Class 3	Min(%)	**-26**	-12	-26	-25	-26	-1	-18	-1	-1	-3	-1
	value min	3	2	5	1	2	1	1	1	2	2	1
	Max(%)	**+29**	+21	+40	+2	+1	+10	+52	+10	+1	+1	+7
	value max	1	3	4	5	1	2	5	2	1	1	2
Class 4	Min(%)	-12	-10	-12	-12	-6	-1	-2	-1	-1	-1	-1
	value min	2	3	3	2	2	1	1	1	1	1	1
	Max(%)	+13	+24	+17	+1	+1	+30	+5	+25	+1	+31	+11
	value max	1	4	1	1	1	2	4	2	2	2	2

The incidents cause facility damage more likely belong to Class 2 or Class 4. We can see that, if an incident causes signal damage, the probability that it belongs to Class 2 will increase with 66%; oppositely, the probability it belongs to Class 1 will decrease with 33%, and the probability it belongs to Class 3 will decrease with 26%. If an incident causes guardrail damage, the probability that it belongs to Class 4 will increase with 30%, the probability that it belongs to Class1 will decrease with 33%.

The incidents which are serious accidents more likely belong to Class 3. We can see that if more than 4 automobiles involve in an incident, the probability that this incident belongs to Class 3 will increase with 52%.

According to the analysis above and the probability distributions, we assume that the four classes can be interpreted as follow: Class 1 is a set of incidents not serious or complicated; Class 2 is a set of incidents not serious but complicated; Class3 is a set of incidents serious but not complicated; Class4 is a set of incidents serious and complicated.

6. Conclusion

A continuous model for incident duration prediction is developed based on LGNB classifier, assuming that the duration of incidents with similar attributions fits a lognormal distribution. This assumption is tested using statistic diagrams and hypothesis tests, and it is found that this assumption is appropriate. According to the evidence sensitivity analysis of LGNB, the four classes of incidents classified by LGNB can be interpreted with different levels of severity and complexity.

LGNB classifier, as the simplest hybrid Bayesian networks, is used to predict the continuous probability distribution of duration, and the accuracy is not satisfied because of the unrealistic assumption. In the future, we will try more realistic models, and apply them in incident impact analysis and severity estimation.

Acknowledgements

This research is supported by the National High-tech R&D Program of China (2009AA11Z219). The authors would like to express their appreciation towards Dr. Stephen Boyles (University of Wyoming) in particular, for valuable assistance in obtaining and interpreting the data used for these models.

References

1. J. Lindley, Urban freeway congestion: quantification of the problem and effectiveness of potential solutions, *ITE journal*, 57 (1987) 23-51.
2. W. Chow, A study of traffic performance models under an incident condition, *Highway Research Record*, 567(1974) 31-36.
3. T. Olmstead. Pitfall to avoid when estimating incident-induced delay by using deterministic queuing models, *Transportation Research Record: Journal of the Transportation Research Board*, 1683 (1999) 38-46.
4. C. Wirasinghe. Determination of traffic delays from shock-wave analysis, *Transportation Research*, 12 (5) (1978) 343-348.
5. H. Mongeot H and J. Lesort. Analytical Expressions of Incident-Induced Flow Dynamics Perturbations: Using Macroscopic Theory and Extension of Lighthill-Whitham Theory, *Transportation Research Record: Journal of the Transportation Research Board*, 1710 (2000) 58-68.
6. V. Sisiopiku, X. Li and etc. Dynamic Traffic Assignment Modeling for Incident Management, *Transportation Research Record: Journal of the Transportation Research Board*, 1994 (2007) 110-116.
7. T. F. Golob, W. W. Recker and J. D. Leonard. An analysis of the severity and incident duration of truck-involved freeway accidents, *Accident Analysis & Prevention*, 19(5) (1987) 375-395.
8. G. Giuliano. Incident characteristics, frequency, and duration on a high volume urban freeway. *Transportation Research Part A: General*, 23 (5) (1989) 387-396.
9. E. C. Sullivan. New Model for Predicting Freeway Incidents and Incident Delays, *Journal of Transportation Engineering*, 123 (4) (1997) 267-275.
10. A. Khattak, J. Schofer and M. H. Wang. A Simple Time Sequential Procedure For Predicting Freeway Incident Duration, *Journal of Intelligent Transportation Systems*, 2 (1995) 113-138.
11. A. Garib, A. E. Radwan and H. Al-Deek. Estimating Magnitude and Duration of Incident Delays, *Journal of Transportation Engineering*, 123 (6) (1997) 459-466.
12. K. Ozbay, P. Kachroo. Incident Management in Intelligent Transportation Systems (Boston: Artech House, 1999).
13. D. Nam, F. Mannering. An exploratory hazard-based analysis of highway incident duration, *Transportation Research Part A: Policy and Practice*, 34 (2) (2000) 85-102.
14. B. L. Smith, B. M. Williams and O. R. Keith. Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting, *Transportation Research Part C*, 10(4)(2002) 303-321.
15. S. Boyles, D. Fajardo and S. T. Waller. A naive Bayesian classifier for incident duration prediction, in *Proc. 86th Meeting of Transportation Research Board* (2007).
16. K. Ozbay and N. Noyan. Estimation of incident clearance times using Bayesian Networks approach, *Accident Analysis & Prevention*, 38(3) (2006) 542-555.
17. A. K. A. Castro, P. R. Pinheiro, M. C. D. Pinheiro et al. Towards the Applied Hybrid Model in Decision Making:A Neuropsychological Diagnosis of Alzheimer's Disease Study Case, *International Journal of Computational Intelligence Systems (IJCIS)*, 4(1) (2011):89-9.
18. L. Fu and L. Rilett. Real-time estimation of incident delay in dynamic and stochastic networks, *Transportation Research Record: Journal of the Transportation Research Board*, 1603(1997) 99-105.
19. K. P. Murphy. The bayes net toolbox for matlab, *Computing science and statistics*, 33(2000) 1024-1034.