The Application of Empirical Bayes Approach for Identifying and Ranking Hazardous Junctions Case Study: Singapore Signalized Junctions

Aine KUSUMAWATI
Assistant Professor
Faculty of Civil & Environmental Engineering
Institut Teknologi Bandung
Jl. Ganesa No. 10, Bandung 40132
Fax: +62-22-2512395
INDONESIA
E-mail: aine@trans.si.itb.ac.id

WONG Yiik Diew
Associate Professor
School of Civil & Environmental Engineering
Nanyang Technological University
Blk N1-01B-51, 50 Nanyang Avenue
SINGAPORE 639798
E-mail: cydwong@ntu.edu.sg

Abstract: The identification and ranking of hazardous road locations are important parts of road safety improvement programs. This paper describes the application of Empirical Bayes (EB) approach for identifying and ranking hazardous junctions. Accident, traffic, and junction geometric/environment data from 203 four-legged and 186 three-legged signalized junctions across western part of Singapore were collected. Accident prediction models were developed and safety of the junctions was estimated. After that, hazardous junctions were identified using probability of selecting the worst site concept and then ranked using *PSI* (potential for safety improvement) and *LH* (level of hazard) criteria. A total of 38 junctions were found as hazardous. The result shows that the use of *PSI* criterion is more favorable than *LH* criterion as it is better able to detect the top hazardous junctions with the largest number of accidents in the study period.

Key Words*: Empirical Bayes, hazardous junctions, level of hazard, potential for safety improvement*

1. INTRODUCTION

It is widely accepted that an effective way to improve road safety is through road safety improvement programs. Such programs involve identification of hazardous sites, prioritization/ranking of hazardous sites for treatment, diagnosis of safety problems, selection and prioritization of feasible treatments and studying the effect of applied treatments, with all aspects to be considered within the limitation of available budget.

In a road safety context, a hazardous site is commonly defined as any site that exhibits an accident potential that is significantly high when compared with some norm or average accident potential which is established from other sites with similar characteristics. The identification of hazardous sites is important so as to avoid wasting resources due to treating the sites that are wrongly identified as unsafe sites and leaving the truly hazardous sites untreated. The process of identification is usually followed by ranking the hazardous sites according to pre-specified criteria to measure the hazard level given limited budget. In this way, the authorities can determine which sites need to be prioritized for safety treatment.

Various methods are available to identify and rank hazardous sites. Commonly, the hazardous sites are selected simply by ranking all the sites according to the hazard level where high-risk sites are assigned smaller ranks and vice versa. The worst sites are then selected from the

outcome of the ranking; the sites with smaller ranking are regarded as the more hazardous sites. Alternatively, a certain limit for accident count or accident rate is set then hazardous sites were selected based on that limit. In this context, a site is defined as hazardous if its observed accident count or accident rate or both exceed the preset limit. However, it is undoubtedly that this method is very sensitive to random variation in accident counts and to the regression-to-mean problem (Hauer, 1986; Elvik, 1997). Basically, the method will (incorrectly) identify hazardous sites as those sites which have high accident counts or accident rates.

Another approach is to select hazardous sites using a critical threshold that is chosen in such a way that it is exceeded by only a small proportion of the sites. Thus, a site is said to be hazardous if its observed accident count or accident rate exceeds this critical threshold. An example of this kind of method is the rate-quality control method described in Stokes and Mutabazi (1996). McGuigan (1981) proposed ranking sites according to their potential for accident reduction (PAR), which is the difference between the reported (actual) number of accidents at a site and the expected number of accidents at sites with similar characteristics. Soon after, McGuigan (1982) postulated that using PAR in ranking hazardous sites is more likely to maximise the cost-effectiveness of a road safety improvement programme than using accident count or accident rate. On the other hand, Maher and Mountain (1988) concluded that using accident count as ranking criterion may perform as well as or better than using PAR due to inaccuracy in the estimation of expected number of accidents at a site which is required in the PAR method.

After safety estimation by Empirical Bayes (EB) approach has become popular, EB estimate has been used as a criterion for the purpose of identification and ranking hazardous sites. Hauer (1996) used the EB safety estimate directly to rank sites. Persaud *et al.* (1999) used it in a method called potential for safety improvement (PSI). This method is quite similar to the PAR method, except that the EB safety estimate is used instead of accident count. Here, PSI is estimated as the difference between the EB safety estimate and what is normal for similar sites. Saccomanno *et al.* (2001) defined hazardous site as a site where the observed number of accidents exceeds either the accident prediction (Poisson) model estimate or EB estimate by at least one standard deviation. They concluded that the EB estimate has yielded fewer hazardous sites than accident prediction (Poisson) model.

This paper describes the application of EB approach for identifying and ranking hazardous signalized junctions in Singapore. Accident prediction models for four-legged and three-legged signalized junctions were developed using Negative Binomial (NB) regression model. The models were thus used in estimating safety of the junctions using EB approach. The hazardous junctions were identified using probability of selecting the worst site concept and then ranked according to *PSI* (potential for safety improvement) and *LH* (level of hazard) criteria.

2. METHODOLOGY

2.1 Accident Prediction Model

Accident count is discrete, and hence does not follow the normal distribution. Thus, it cannot be modelled using ordinary linear regression model. Currently, there are a number of modelling approaches to deal with discrete count data. The more familiar approach is Poisson

regression model. However, it is sometimes the case that the Poisson regression model is not appropriate to model accident count data whenever there is large variability within the data. In this case, a NB regression model may be used as they can cater to the over-dispersion in the data. NB model assumes that the distribution in the number of accidents at a site is Poisson and the distribution of expected accident counts in the population is Gamma.

If Y_i is an independent random variable that follows a NB distribution with expected value μ_i, then the probability function of Y_i is given by:

$$f(Y_i = y_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad ; \quad i = 1, \ldots, n \tag{1}$$

with α being the dispersion parameter, and $\alpha \geq 0$.

Here,

$$E(Y_i) = \mu_i = \exp\left(X_i^{'}\beta\right) = \exp\left(\sum_1^p x_{ij}\beta_j\right) \tag{2}$$

and

$$Var(Y_i) = \mu_i + \alpha\mu_i^2 \tag{3}$$

The Poisson regression model can be regarded as a limiting model of the NB regression model as α approaches zero. However, it is noted that the NB model does not fit well if the data are under-dispersed as it requires that the variance is greater than the mean. The appropriateness of using NB model over Poisson regression model can be determined by the statistical significance of an estimated coefficient α which is a parameter related to the degree of over-dispersion. When α is significantly larger than zero (as measured by the t-statistic), then the suitability of the NB regression model is confirmed. Otherwise, the Poisson model would prevail (Poch and Mannering, 1996).

Evaluation of the goodness of the fit between the observed values y_i and the fitted values $\hat{\mu}_i$ for Poisson and NB regression models can be assessed by a number of statistics. Two well known ones are the Pearson X^2 and Scaled Deviance G^2 statistics. Here the models are assumed to be nested, with the larger model having the greater number of parameters. These statistics follow χ^2 distribution with *(n-p)* degrees of freedom where *n* is the number of data points (observations) and *p* is the number of estimated parameters.

The Pearson X^2 statistic is defined as sum of squares of standardised observations, while the Scaled Deviance G^2 can be defined as twice the logarithm of the ratio of the likelihood of the data under the larger model, to that under the smaller model. The Pearson X^2 and Scaled Deviance G^2 statistics for the Poisson regression model are:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{4}$$

$$G^2 = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right] \tag{5}$$

For the NB regression model:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i + \alpha \hat{\mu}_i^{\,2}} \tag{6}$$

$$G^2 = 2\sum_{i=1}^{n} \left(\ln\left(\frac{y_i}{\hat{\mu}_i}\right)^{y_i} - (y_i + 1/\alpha)\ln\left(\frac{1+\alpha y_i}{1+\alpha \mu_i}\right) \right) \tag{7}$$

A model passes the goodness of fit criteria when the values of Pearson X^2 and Scaled Deviance G^2 statistics are greater than or equal to the value of χ^2 distribution with *(n-p)* degrees of freedom for a certain confidence level (in this case, 95%).

2.2 EB Approach for Safety Estimation

The safety of an entity is defined as 'the number of accidents by kind and severity, expected to involve or to occur at the entity, per unit of time, in a certain period' (Hauer, 1992 and 1997). Hauer (1997) postulated that the count of accidents (*K*) observed at an entity is a biased estimate of its expected number of accidents (*m*), and proposed the EB approach for estimating the safety of an entity. Safety estimation in the EB approach is based on an entity's traits (such as gender, age, traffic or geometry) and the entity's historical accident records. Thus, it requires information on mean and variance of the safety for a reference population of similar entities.

However, there were inherent difficulties in defining the reference population in EB method. There is a problem when a sizeable number of reference population does not exist. Even if the number of entities is large enough, it does not mean that the reference population will be easily defined. Elvik (1988) rationalized that a group of entities cannot form a population unless they are sufficiently similar among themselves that the accident counts defined for the group of entities are able to be fitted by a probability distribution. Hence, to overcome the difficulties in defining the reference population, Hauer (1997) suggested using multivariate modeling method to estimate the safety of the reference population and this approach was adopted in this research.

To derive the EB model, Hauer (1997) started with introducing two variables namely *m* (the expected number of accidents of an entity occurring within a group of similar entities during a specified time period) and *K* (historical accident counts of the entity in a specified period). Let the mean and variance of *m* in the reference population (group of entities) be $E(m)$ and $Var(m)$. Here, $E(m)$ is taken as the safety of the population. Where information on the historical accident counts of an entity is not available, $E(m)$ is also the best estimate of safety *m* of the entity. However, if such information is available then the best estimate of safety *m* of the entity that has *K* accidents is given by $E(m|K)$. In this regard, $E(m|K)$ and $Var(m|K)$ respectively denote the mean and the variance of *m* in the subpopulation that recorded *K* accidents.

The probability distribution $f(m|K)$ was built by assuming that the probability distribution of m in the entities of the reference population is described by a Gamma probability density function, which is denoted by $g(m)$. That is, for $m \geq 0$,

$$g(m) = \frac{a^b m^{b-1} e^{-am}}{\Gamma(b)} \tag{8}$$

The parameters *a* and *b* are related to mean $E(m)$ and variance $VAR(m)$ as follows:

$$E(m) = \frac{b}{a} \quad , \quad Var(m) = \frac{b}{a^2} \tag{9}$$

$E(m)$ is estimated by using the result of accident prediction model where the expected accident count *K* is given by μ, which is similar to $E(m)$, with variance equal to $Var(\mu)$. Hauer (1992) showed that $Var(\mu) = E(m) + Var(m)$. For the case of accident prediction model that is built using NB regression model then $Var(m) = \alpha E(m)^2$ because the variance of the NB regression model is given by $Var(\mu) = \mu + \alpha \mu^2$ and also $\mu = E(m)$. Here, α is the dispersion parameter of the NB model and $b = 1/\alpha$. Using Bayes' theorem and assuming that the accident count *K* is Poisson-distributed, the probability distribution for $f(m|K)$ is thus:

$$f(m|K) = \frac{(1+a)^{K+b} m^{K+b-1} e^{-m(1+a)}}{\Gamma(K+b)} \tag{10}$$

with mean and variance

$$E(m|K) = \frac{K+b}{1+a} \quad , \quad Var(m|K) = \frac{K+b}{(1+a)^2} \tag{11}$$

It has to be noted that the EB method relies upon assumption that the count of accident at an entity *K* is Poisson-distributed while many researchers have indicated that accident counts need not always be Poisson-distributed (Nicholson, 1985; Ng et al, 1995). However, Kusumawati (2008) examined 1186 samples of historical accident count sequences and found that only 8.6% of the samples were rejected to be Poisson-distributed although the remaining samples that could not be rejected as Poisson-distributed need not always had index of dispersion equal to one (as required by Poisson distribution).

2.3 Identification and Ranking of Hazardous Junctions

The identification of hazardous junctions is based on the probability of selecting a site with accident potential exceeding what is normal for sites similar to the site being investigated. The accident potential of a site is represented by the EB safety estimate $E(m|K)$ while the normal accident potential is represented by the average accident potential of junctions similar to the junction being investigated, which is the result of safety estimation using accident prediction model $E(m)$.

Given a site with accident potential represented by $E(m|K)$ - mean of the posterior distribution, and a threshold value represented by $E(m)$ - mean of the prior distribution. Here, $E(m|K)$ or $\hat{m}$ is the EB estimate. The site is identified as hazardous if there is a significant

probability that its accident potential exceeds the value that is normal for similar sites in the reference population. So, mathematically, a site is identified as hazardous if:

$$\left[P\left(E\left(m|K\right) > E\left(m\right)\right) = 1 - \int_0^{E(m)} \frac{(1+a)^{K+b} m^{K+b-1} e^{-m(1+a)}}{\Gamma(K+b)} dm \right] \geq \delta \tag{12}$$

where δ represents the minimum accepted confidence level, which can be any value smaller than one. In this case, the value of δ is taken as 0.95.

After the hazardous junctions are identified, next step is to rank the junctions for priority treatment. In this case, two criteria that make use of the EB estimate were considered for ranking the hazardous sites for treatment priority. The first one is to rank the junctions according to the potential for safety improvement (PSI), which is defined as the positive difference between EB estimate $E\left(m|K\right)$ and the normal accident potential for similar sites in the reference population $E\left(m\right)$:

$$PSI = E\left(m|K\right) - E\left(m\right) \tag{13}$$

All hazardous sites will have a PSI value greater than zero though not all sites having PSI larger than zero are necessarily hazardous.

The level of hazard (LH) is another criterion that sounds logical to be used in ranking the hazardous sites. It is a ratio, instead of difference, of MEB estimate $E\left(m|K\right)$ to the normal accident potential for similar sites in the reference population $E\left(m\right)$. The level of hazard (LH) is defined as:

$$LH = \frac{E\left(m|K\right)}{E\left(m\right)} \tag{14}$$

As in the PSI method, all hazardous sites will have an LH value greater than one though not all sites having LH greater than one are necessarily hazardous.

3. MODEL DEVELOPMENT

3.1 Data

In building up the database, on-site visits were made to hundreds of signalized junctions in the western part of Singapore over the period of 2003 up to mid 2004. At each site, information was captured on the geometric lay-out as well as the operational features. These junctions were also checked against various records to ensure there were no major geometric/operational changes during the part of the study period (1999-2003). Altogether, a total of 203 four-legged signalized junctions and 186 three-legged signalized junctions were assembled in the final junction sample. Junction geometric and operational data were obtained from field data collection. At first, twenty seven types of geometric and operational variables

were collected, however some variables were omitted in modelling stage due to lack of data variation or multicollinearity between variables.

The road traffic accident data were extracted from a computerised accident database obtained from Singapore Traffic Police Department (TPD). It should be noted that the TPD accident records cover only reported fatal and injury accidents. Therefore, any minor accident not incurring any injury to person or involves only damage to property is not reflected in the accident database because such accidents may be settled between the accident parties, and are usually not reported to the TPD. For the purpose of this research, the accident data were extracted using SAS application software package in order to obtain accident counts pertaining to respective junctions and all the available information related to the accidents.

The traffic flow data were obtained from the Green Link Determining System (GLIDE) records. The GLIDE is an intelligent traffic control system that controls traffic signal and manages traffic on Singapore's road network. Data obtained from GLIDE included junction layout, phase diagram that indicates the phase sequence, and traffic counts at 15-minute intervals for each approach lane of the junction. Traffic count data obtained from GLIDE, however, have some limitations. Firstly, data are recorded in the form of number of vehicles that passed a detector without classification by vehicle type. Secondly, the GLIDE detectors do not cover left-turn movements along the sliproads. Thirdly, a detector can at times be faulty due to various reasons and there would then be no traffic count data available for the affected lane(s).

Despite the limitations, this research still relied on the GLIDE to obtain information regarding traffic flows at the study junctions as the GLIDE provides the only practical source of traffic flow data on an area-wide basis. It would be overly resource-intensive to manually collect the data at the junctions otherwise. A study by Naing (2004) showed that the average deviation of the GLIDE count data as compared to manual count data was –7.3% (under-counting by GLIDE) when motorcycles were included into the analysis and +3.3% (over-counting by GLIDE) when motorcycles were excluded from the manual counts. To correct for unavailability of counts at the sliproads, several models were developed to estimate traffic flows at the sliproads which consist of models for estimating the left-turn sliproad flows at four-legged signalized junctions, at middle leg of three-legged signalized junctions, and at minor leg (stem) of three-legged signalized junctions, respectively. The models were developed through simple regression analysis using manual traffic count data from Singapore Land Transport Authority (LTA) and from field studies by Chia (2004) and Leong (2004). The resulting models can be seen in Kusumawati (2008).

Since the traffic flow data were collected only for 2004, estimations were done to obtain the data for other periods. Although it is desirable to find the specific growth rate applicable to each junction there was insufficient resources to calibrate the individual rates. It was assumed that all the junctions in study area had similar annual traffic growth rate corresponding to the annual growth rate of vehicle population in Singapore. A sensitivity analysis was then carried out to examine the impact of inaccuracy of the estimated traffic flow data on the developed models. It was found that varying traffic flow for four-legged signalized junctions by -30% to +30% of modelling value would result in expected accident number to change by between 0.734 to 1.255 of its initial value whereas varying traffic flow for three-legged signalized junctions for the same numbers as the four-legged signalized junctions would result in expected accident number μ to change by between 0.689 to 1.316 of its initial value. However, it was also found that the change of traffic flow value had some but lesser effect on

the outputs of EB methods. For example, varying the traffic flow in the four-legged signalized junctions model by -30% to +30% of the initial value of traffic would result an average change of 7% in the $E(m|K)$. Therefore, although the developed accident prediction models were sensitive to changes in traffic flow, their effects were moderated when their outputs were incorporated into the EB methods.

3.2 Model Estimation

The accident prediction models were developed based on time-aggregated accident counts data during 5-year (1999-2003) modelled period. The developed model takes the form:

$$\mu = k\left(Q_p Q_s\right)^{\gamma} e^{\Sigma(\beta_i G_i)} \tag{15}$$

where μ equals to expected accident number during 5-year period; Q_p and Q_s are total daily junction inflow in the primary and secondary direction, respectively; G_i equals to junction geometric/operational variables and k, γ, β_i, are the model parameters to be estimated. Table 1 presents the explanatory variables used for developing the accident prediction model.

Tabel 1 Explanatory variables of the accident prediction models

Explanatory variables	Codes	Types	Coding, as applicable
Daily junction inflow in the primary direction	Qp	Quantitative	
Daily junction inflow in the secondary direction	Qs	Quantitative	
Number of approach lanes	NAPLN	Quantitative	
Number of exit lanes	NEXLN	Quantitative	
Number of exclusive right-turn lanes	RTLANE	Quantitative	
Number of shared lanes	SHLANE	Quantitative	
Number of signal phases	NPHASE	Quantitative	
Classification of junction	CLASS	Qualitative	1 : if major junction 0 : otherwise
Sliproad	SLR	Qualitative	1 : if sliproad present on all junctions legs 0 : otherwise
	NOSLR	Qualitative	1 : if sliproad not present on all junctions legs 0 : otherwise
Median	MED	Qualitative	1 : if median present on all junctions legs 0 : otherwise
	NOMED	Qualitative	1 : if median not present on all junctions legs 0 : otherwise
Pedestrian crossing	PEDCROSS	Qualitative	1 : if pedestrian crossing present on all junctions legs 0 : otherwise
Downstream merger	MERG	Qualitative	1 : if downstream merger present on at least one junction leg 0 : otherwise
U-turn	UTURN	Qualitative	1 : if U-turn present on at least one junction leg 0 : otherwise
Red-light camera	RLCAM	Qualitative	1 : if red-light camera presents on at least one junction leg 0 : otherwise
Yellow box	YBOX	Qualitative	1 : if yellow box presents 0 : otherwise
Landuse	CITY	Qualitative	1: if junction is located in city area 0 : otherwise
	TOWN	Qualitative	1 : if junction is located in town area 0 : otherwise

The development of the accident prediction models was initiated by removing junctions with non-Poisson accident count sequence during the modelled period from the initial database for each kind of model. The examination of goodness-of-fit to Poisson distribution of the accident count sequence of every junctions were carried out using exact test (Nicholson and Wong, 1993), likelihood test (Rao and Chakravarti, 1956), Kolmogorov-Smirnov test (Campbell and Oprian, 1979) and multinomial test (Cressie and Read, 1984). In this case, there were 32 four-legged signalized junctions and 10 three-legged signalized junctions which had non-Poisson accident data sequences during the modelled period (1999-2003). These junctions were therefore removed from the database.

Multivariate modelling was then carried out on the traffic flows and the geometric/ operational variables as presented in Table 1 using LIMDEP application software package. Variable selection was carried out using a backward elimination procedure whereby the least significant variable at 95% confidence level was progressively eliminated one by one from the model. This process was continued until all remaining variables were statistically significant to be retained in the model. It was found that *CLASS* and *CITY* were the only significant junction geometric/operational variables, and also NB was the more appropriate regression model over the Poisson regression model.

The resulting four-legged signalized junctions model is:

$$\mu = 1.137 \times 10^{-3} \left(Q_p Q_s\right)^{0.434} e^{(0.624 CLASS - 0.649 CITY)} \tag{16}$$

The resulting three-legged signalized junctions model is:

$$\mu = 1.706 \times 10^{-4} \left(Q_p Q_s\right)^{0.523} e^{(0.601 CLASS - 1.496 CITY)} \tag{17}$$

Tables 2 and 3 presents the complete statistical aspects of the four-legged and three-legged signalized junctions models, respectively. The fit of the variance functions of the four-legged and three-legged signalized junctions models to the averages of models squared residuals are presented in Figures 1 and 2, respectively. Those figures show that the averages of squared residuals were clustered around the variance function which indicates a good fitting model.

Table 2 Accident prediction model for four-legged signalized junctions

$\mu = e^{(-6.779+0.434\ln(Q_p Q_s)+0.624 CLASS-0.649 CITY)}$			
Degree of freedom	165		
Pearson X^2	148.424	$\chi^2_{0.05,165}$ = 196	
Scaled Deviance G^2	1 93.253	$\chi^2_{0.05,165}$ = 196	
Variable	Coefficient	t-value	P-value
Constant *c*	-6.779	-5.290	0.000
ln($Q_{pt}Q_{st}$)	0.434	6.346	0.000
CLASS	0.624	4.059	0.000
CITY	-0.649	-3.902	0.000
Dispersion parameter α	0.416	6.646	0.000

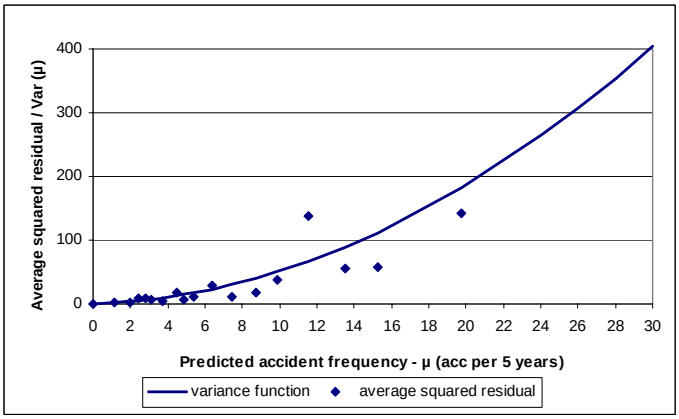

Figure 1 Variance function of four-legged signalized junctions model

Table 3 Accident prediction model for three-legged signalized junctions

$\mu = e^{(-8.676+0..523\ln(Q_p Q_s)+0.601CLASS-1.496CITY)}$			
Degree of freedom	167		
Pearson X^2	167.513	$\chi^2_{0.05,167}$ = 198.2	
Scaled Deviance G^2	178.399	$\chi^2_{0.05,167}$ = 198.2	
Variable	Coefficient	t-value	P-value
Constant c	-8.676	-5.311	0.000
$\ln(Q_p Q_s)$	0.523	5.869	0.000
$CLASS$	0.601	2.809	0.005
$CITY$	-1.496	-3.150	0.002
Dispersion parameter α	0.616	5.357	0.000

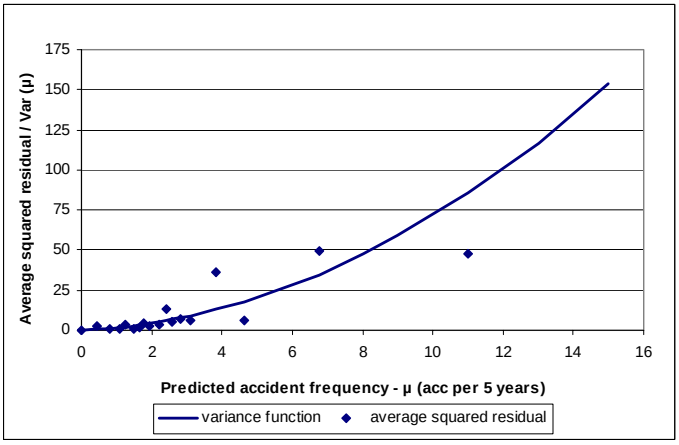

Figure 2 Variance function of three-legged signalized junctions model

4. IDENTIFICATION AND RANKING OF HAZARDOUS JUNCTIONS

The identification of hazardous junctions were based on a criterion that there is a significant probability that its accident potential exceeds the value that is normal for similar sites in the reference population. In this case, the accident potential of a junction is represented by $E(m|K)$, which was computed using the method presented in Section 2.2 based on the

accident prediction model of the junction as given in Equations 16 or 17. The accident potential that is normal for similar junctions in the reference population is represented by $E(m)$, as obtained from the accident prediction model. Equation 12 was then used to identify the hazardous junctions in the study area based on 1999-2003 accident data. The methods identified a total of 38 junctions as being hazardous based on the criterion that the probability of $E(m|K)$ exceeding $E(m)$ is greater than 0.95, which consisted of 23 four-legged signalized junctions and 15 three-legged signalized junctions. The hazardous junctions are listed in Table 4.

Table 4 List of hazardous junctions

No	ID	$K_{1999–2003}$	$E(m)$	$E(m\|K)$	$P(E(m\|K)>E(m))$	$K_{2004–2006}$
1	X69	35	9.775	30.021	1.000	18
2	X89	15	4.204	11.072	1.000	14
3	X113	47	11.564	40.901	1.000	19
4	X153	43	18.862	40.271	1.000	9
5	T4	23	5.329	18.873	1.000	11
6	T9	19	3.670	14.297	1.000	4
7	T61	11	2.366	7.486	1.000	8
8	X117	27	13.082	24.839	0.999	12
9	X158	32	15.983	29.906	0.999	21
10	X188	18	7.001	15.189	0.999	12
11	X145	10	2.341	6.119	0.998	5
12	X146	13	4.268	9.854	0.997	4
13	X177	13	4.421	9.979	0.997	3
14	T157	10	2.826	7.382	0.997	2
15	X107	16	6.574	13.476	0.996	6
16	T101	11	3.558	8.668	0.996	0
17	T67	5	0.358	1.195	0.993	0
18	X9	25	14.373	23.477	0.990	9
19	X27	7	1.383	3.435	0.989	9
20	T91	8	2.366	5.707	0.988	2
21	T107	21	11.819	19.891	0.986	8
22	X94	9	2.929	6.263	0.985	2
23	X76	21	11.904	19.472	0.983	3
24	T60	8	2.674	5.988	0.982	0
25	X24	23	13.937	21.667	0.978	5
26	T12	10	4.186	8.375	0.978	6
27	X121	21	12.408	19.605	0.976	15
28	X181	11	4.834	8.952	0.975	7
29	X115	17	9.383	15.446	0.974	7
30	T130	17	9.649	15.941	0.973	10
31	X39	12	5.676	10.118	0.972	5
32	X11	26	16.979	24.881	0.968	14
33	T32	7	2.491	5.220	0.965	2
34	T7	6	1.860	4.070	0.963	3
35	T92	5	1.204	2.820	0.963	2
36	T20	8	3.303	6.452	0.960	3
37	X178	11	5.490	9.322	0.954	4
38	X88	9	4.001	7.124	0.953	9

The hazardous junctions were then ranked according to their *PSI* and *LH* values as explained in Section 2.3 and the result is presented in Table 5.

Table 5 Rank of hazardous junctions

No	ID	$K_{1999–2003}$	$K_{2004–2006}$	*PSI*	*LH*	Rank by *PSI*	Rank by *LH*
1	X69	35	18	20.246	3.071	3	6
2	X89	15	14	6.868	2.634	16	7
3	X113	47	19	29.337	3.537	1	3
4	X153	43	9	21.410	2.135	2	20
5	T4	23	11	13.544	3.541	5	2
6	T9	19	4	10.627	3.896	7	1
7	T61	11	8	5.120	3.164	21	5
8	X117	27	12	11.758	1.899	6	25
9	X158	32	21	13.923	1.871	4	26
10	X188	18	12	8.187	2.169	9	18
11	X145	10	5	3.779	2.614	23	8
12	X146	13	4	5.586	2.309	19	14
13	X177	13	3	5.557	2.257	20	15
14	T157	10	2	4.556	2.612	25	9
15	X107	16	6	6.902	2.050	15	22
16	T101	11	0	5.110	2.436	22	11
17	T67	5	0	0.838	3.343	37	4
18	X9	25	9	9.104	1.633	8	35
19	X27	7	9	2.051	2.483	34	10
20	T91	8	2	3.341	2.412	29	12
21	T107	21	8	8.072	1.683	10	31
22	X94	9	2	3.334	2.138	30	19
23	X76	21	3	7.568	1.636	13	34
24	T60	8	0	3.314	2.239	31	16
25	X24	23	5	7.730	1.555	12	37
26	T12	10	6	4.189	2.001	28	23
27	X121	21	15	7.198	1.580	14	36
28	X181	11	7	4.118	1.852	26	27
29	X115	17	7	6.064	1.646	17	32
30	T130	17	10	6.292	1.652	18	33
31	X39	12	5	4.443	1.783	24	28
32	X11	26	14	7.902	1.465	11	38
33	T32	7	2	2.730	2.096	36	21
34	T7	6	3	2.210	2.188	35	17
35	T92	5	2	1.616	2.342	38	13
36	T20	8	3	3.149	1.953	32	24
37	X178	11	4	3.832	1.698	27	30
38	X88	9	9	3.123	1.780	33	29

5. ANALYSIS

The $E(m|K)$ in Table 4 represents the expected number of accidents that will occur per 5 years for a certain junction assuming constant accident rate and no treatment applied. The $E(m)$

represents the normal expected number of accidents that will occur per 5 years for a group of similar junctions in the reference population. A 3-year accident data of year 2004-2006 ($K_{2004-2006}$**)** was then used to validate whether it is true that the junctions were truly hazardous by comparing whether the $K_{2004-2006}$ for every junctions were indeed larger than their respective $E(m)$ that were already converted into the expected number of accidents per 3 years. It turns out that there were 6 junctions (X24, X76, X153, T60, T67, T101) out of the 38 hazardous junctions listed in Table 4 which had $K_{2004-2006}$ smaller than their respective $E(m)$. Thus, since only 3 years data were available for validation, it cannot be concluded yet that the 6 junctions were mistakenly identified as hazardous. It is possible that in the next two years (after 2006) the junctions would have total number of accidents exceeded their $E(m)$. However, when this research was completed, the accident data for 2007 and 2008 were not yet available.

After the hazardous junctions were identified, the junctions were ranked for remedial treatment due to constraint in budget for treatment. Nevertheless, one should not expect that a treatment can bring a junction into a zero accident state. The reason is that road traffic accidents are random; they are influenced by pure random variation as well as systematic random variation. Systematic variation, which is attributable by various causal factors can be controlled, however the pure random variation cannot be controlled; as long as there are exposures (vehicles), accidents shall occur. Thus, the most sensible approach is to bring the number of accidents at a junction back to normal condition which is representative for all similar junctions in the reference population.

The PSI and LH criteria used in this research is based on prioritizing junctions according to safety benefit, in terms of reduction in the number of accidents expected to occur in future period, that can be obtained by treating a certain junction. In the PSI criterion, the junctions are ranked according to the difference in the number of accidents expected to occur at a certain junction with the 'normal' number of accidents expected to occur in similar junctions in the reference population. However, by using the PSI criterion, a junction with a relatively low number of expected accidents can never be prioritized for treatment though the normal expected accidents for the similar junctions in the population is much lower. On the contrary, the LH criterion is able to detect such junction as it compares ratio of the number of accidents expected to occur at a certain junction to the 'normal' number of accidents expected to occur in similar junctions in the reference population. This situation can be seen, as an example, for T-61 junction. It is shown in Table 5 that the junction was ranked as number 21 by using the PSI criterion but it was ranked as number 5 by using the LH criterion. Nevertheless, it seems that treating a junction with higher expected number of accidents is more realistic than treating a junction with lower expected number of accidents.

Comparing the ranking results of top 5 and top 10 hazardous junctions using both criteria, the PSI criterion was able to detect more number of accidents that occurred in 2004-2006. The total number of $K_{2004-2006}$ for the top 5 and top 10 hazardous junctions using the PSI criterion were 78 and 137, respectively while the same number using the LH criterion were only 42 and 90. Thus, use of the PSI criterion was more favorable than the LH criterion for ranking of hazardous junctions.

Another criterion that can be used for ranking is accident cost. In this case, the accident cost for fatal accidents is larger than for serious injury accidents and the accident cost for serious

injury accidents is larger than for slight injury accidents. The junctions that shall be prioritized for treatment are the junctions with the largest accident cost among others. Nevertheless, the junctions with smaller accident cost may have higher level of hazard. This kind of method is not favored because it seems unfair to leave such junctions untreated thus exposing the road users to such high level of hazard for the simplistic reason that the road is less cost-effective to treat.

The ranking of junctions presented in Table 5 should not be taken as a final one. Road authority may consider other factors before deciding the final ranking order. Those factors as mentioned in UK DOT (1986) include treatment difficulty, extent of the budget, availability of staff resources, construction program constraints, pressures by elected representatives, pressure by the public and media, emotional reaction by community, and geographical spread of remedial work load. After the road authority decides on the junctions to be treated, the next stage in road safety improvement programs is to do in-depth analysis to identify possible remedial measures for each problem encountered by the junctions.

6. CONCLUSIONS

The identification and ranking of hazardous junctions is very important towards being able to select the junctions for priority treatment, given limited budget. The identification method was developed by applying the EB approach to estimate safety. The ranking criteria were also set based on the EB estimates. Two criteria were proposed to rank the hazardous junctions, which were potential for safety improvement (PSI) and level of hazard (LH).

Using the identification method, 38 hazardous signalized junctions were identified as hazardous. A 3-year accident data for year 2004-2006 ($K_{2004-2006}$**)** was then used to validate whether it is true that the junctions were truly hazardous. It turns out that there were 6 junctions out of the 38 hazardous junctions which had $K_{2004-2006}$ smaller than their respective $E(m)$. However, since only 3 years data were available for validation, it cannot be concluded yet that the 6 junctions were mistakenly identified as hazardous.

The hazardous junctions were afterwards ranked according to their PSI and LH values. The results showed that there was clear discrepancy in the ranking by PSI and LH criteria. The LH criterion seems not to be able to detect the most hazardous junctions as well as the PSI criterion. Therefore, ranking of hazardous junctions based on PSI criterion is preferred. However, it is noted that the ranking presented here shall only be taken as preliminary ranking since there are other factors which may affect the final ranking order.

ACKNOWLEDGEMENTS

The research study reported in this paper was performed in the School of Civil & Environmental Engineering, Nanyang Technological University, Singapore. The authors thank the Land Transport Authority and the Traffic Police Department for rendering assistance to the study. The authors also thank the reviewers for their constructive comments. The results and interpretation of analyses are, however, completely the views of the authors.

REFERENCES

Campbell, D.B. and Oprian, C.A. (1979) On the Kolmogorov-Smirnov test for the Poisson distribution with unknown mean, **Biometrical Journal, Vol. 21, No. 1**, 17-24.

Chia, P.L. 2004 **Traffic flows at signalised junctions**, Final Year Project, School of Civil and Environmental Engineering, Nanyang Technological University, Singapore.

Cressie, N. and Read, T.R.C. (1984) Multinomial goodness-of-fit tests, **Journal of the Royal Statistical Society, Vol. 46, No. 3**, 440-464.

Elvik, R. (1988) Some difficulties in defining populations of "entities" for estimating the expected number of accidents, **Accident Analysis & Prevention, Vol.20, No.4**, 261-275.

Elvik, R. (1997) Evaluations of road accident blackspot treatment: A case of the iron law of evaluation studies?, **Accident Analysis & Prevention, Vol. 29, No. 2**, 191-199.

Hauer, E. (1986) On the estimation of the expected number of accidents" **Accident Analysis & Prevention, Vol. 18, No. 1**, 1-12.

Hauer, E. (1992) Empirical Bayes approach to the estimation of "unsafety": The multivariate regression method, **Accident Analysis & Prevention, Vol. 24, No. 5**, 457-477.

Hauer, E. (1996) Identification of "Sites with Promise, **Transportation Research Record, Vol. 1542**, 54-60.

Hauer, E. (1997) **Observational before-after studies in road safety: Estimating the effect of highway and traffic engineering measures on road safety**, Pergamon.

Kusumawati (2008) **Traffic safety at road junctions**, PhD Thesis, Nanyang Technological University, Singapore.

Leong, M.Y. (2004) **Traffic flows at signalised junctions**, Final Year Project, School of Civil and Environmental Engineering, Nanyang Technological University, Singapore.

Maher, M.J. and Mountain, L.J. (1988) The identification of accident blackspots: A comparison of current methods, **Accident Analysis & Prevention, Vol. 20, No. 2**, 143-151.

McGuigan, D.R.D. (1981) The use of relationship between road accident and traffic flow in black-spot identification, **Traffic Engineering and Control, Vol. 22, No. 8**, 448-453.

McGuigan, D.R.D. (1982) Non-junction accident rates and their use in 'black-spot' identification, **Traffic Engineering and Control, Vol. 23, No. 2**, 60-65.

Naing, M.M. (2004) **Evaluation of GLIDE inductance loop detectors for volume counts**, MSc Dissertation, School of Civil & Environmental Engineering, Nanyang Technological University, Singapore.

Ng, C.H., Wong, Y.D., and Lum, K.M. (1995) The applicability of the Poisson distribution in traffic accident count analysis - The Singapore Experience, **Journal of The Institution of Engineers Singapore, Vol.35, No.2**, 49-53.

Nicholson, A.J. (1985) The variability of accident counts, **Accident Analysis & Prevention, Vol.17, No.1**, 47-56.

Nicholson, A. and Wong, Y.D. (1993) Are accidents Poisson distributed? A statistical test", **Accident Analysis & Prevention, Vol. 25, No. 1**, 91-97.

Persaud, B., Lyon, C., and Nguyen, T. (1999) Empirical Bayes procedure for ranking sites for safety investigation by Potential for Safety Improvement, **Transportation Research Record, Vol. 1665**, 7-12.

Poch, M. and Mannering, F.L. (1996) Negative Binomial analysis of intersection-accident frequencies, **ASCE Journal of Transportation Engineering, Vol. 122, No. 2**, 105-113.

Rao, C.R. and Chakravarti, I.M. (1956) Some small sample tests of significance for a Poisson distribution, **Biometrics, Vol. 12, No. 3**, 264-282.

Saccomanno, F.F., Grossi, R., Greco, D., and Mehmood, A. (2001) Identifying black spots along highway SS107 in Southern Italy using two models, **Journal of Transportation Engineering, Vol. 6**, 515-521.

Stokes, R.W. and Mutabazi, M.I. (1996) Rate-quality control method of identifying hazardous road locations, **Transportation Research Record, Vol. 1542**, 44-48.

UK DOT (1986) **Accident Investigation Manual**, UK Department of Transport, London.