BAYESIAN HIERARCHICAL ANALYSIS

ON CRASH PREDICTION MODELS

HUANG HELAI

NATIONAL UNIVERSITY OF SINGAPORE

2007

BAYESIAN HIERARCHICAL ANALYSIS

ON CRASH PREDICTION MODELS

HUANG HELAI

B.E., M.E. **(***Tianjin University***)**

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF CIVIL ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2007

ACKNOWLEDGEMENTS

A journey is easier and more fruitful when people travel together since interdependence is certainly more valuable than independence. This thesis is the result of four years of research in National University of Singapore, whereby I have been accompanied and supported by many people. It is pleasant that I have now the opportunity to express my gratitude for all of them.

I wish to express my deepest gratitude to my supervisor, Associate Professor Chin Hoong Chor for his constructive advices, constant guidance, exceptional support and encouragement throughout the course of the study. During these years, I have known Prof Chin as a strict and principle-centered mentor with excellent and unique discernment about the reality as well as the future. He showed me different ways to approach a problem and the need to be persistent to accomplish any goal. He could not even realize how much I have learned from him. I am really feeling fortunate that I have come to get know Prof Chin in my life.

I would like to thank the members of my PhD committee who monitored my work and gave me invaluable suggestions on the research topic: Professor Quek Ser Tong and Associate Professor Phoon Kok Kwang. Special thanks also go to my module lecturers and some other professors in Department of Civil Engineering in NUS: Dr. Meng Qiang, Associate Professor Lee Der Horng, Associate Professor Cheu Ruey Long, Associate Professor Chua Kim Huat, David.

I am also greatly indebted to the technicians in the traffic laboratory Mr Foo Chee Kiong, Mdm. Chong Wei Leng and Mdm. Theresa for their immense support and accompany during my study period.

Heartfelt thanks and appreciation are also due to my colleagues and friends namely, Dr. Mohammed Abdul Quddus, Mr. Foong Kok Wai, Zhou Jun, Kamal, Shimul, Ashim for their nice company and encouragement during the study period.

I gratefully acknowledge the National University of Singapore for providing research scholarship covering the entire period of this study.

Last, but not least, I would like to take this opportunity to give special gratitude to my parents for giving me life in the first place, for educating me with aspects from both arts and sciences, for unconditional support and encouragement to pursue my interests.

Huang Helai

National University of Singapore
August 2007

TABLE OF CONTENTS

CHAPTER TWO

REVIEW OF CRASH PREDICTION MODELS

CHAPTER THREE

MODELING MULTILEVEL DATA AND EXCESS ZEROS IN CRASH FREQUENCY PREDICTION

CHAPTER FOUR

CRASH FREQUENCY PREDICTION MODEL ON SIGNALIZED INTERSECTIONS

CHAPTER FIVE

BAYESIAN HIERARCHICAL BINOMIAL LOGISTIC MODEL IN CRASH SEVERITY PREDICTION

CHAPTER SIX

SEVERITY OF DRIVER INJURY AND VEHICLE DAMAGE IN TRAFFIC CRASHES AT SIGNALIZED INTERSECTIONS

CHAPTER SEVEN

CONCLUSIONS AND RECOMMENDATIONS

SUMMARY

Crash prediction model is one of the most important techniques in investigating the relationship of road traffic crash occurrence and various risk factors. Traditional models using generalized linear regression are incapable of taking into account the within-cluster correlations, which extensively exist in crash data generating or collecting process.

To overcome the problem, this study develops a Bayesian hierarchical approach to analyze the traffic crash frequency and severity. Zero-inflated Poisson model with location-specific random effects is proposed to capture both the multilevel data structure and excess zeros in crash frequency prediction. And for crash severity prediction, a hierarchical binomial logistic model is developed to examine the individual severity in the presence of within-crash correlation. Bayesian inference using Markov Chain Monte Carlo algorithm is developed to calibrate the proposed models and a number of Bayesian measures such as the deviance information criterion, cross-validation predictive densities, and intra-class correlation coefficients are employed to establish the model suitability.

The proposed method is illustrated using the Singapore crash records. Comparing the predictive abilities of the proposed models against those of traditional methods, the study proved the importance of accounting for the within-cluster correlations and demonstrated the flexibilities and effectiveness of the Bayesian hierarchical method in modeling multilevel structure of traffic crash data.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
BCI	Bayesian Credible Interval
BI	Bayesian Inference
BIC	Bayesian Information Criterion
BUGS	Bayesian Inference Using Gibbs Sampling
CBD	Central Business District
CPM	Crash Prediction Model
CSPM	Crash Severity Prediction Model
CV	Cross Validation
DF	Degree of Feedom
DIC	Deviance Information Criterion
GEV	Generalized Extreme Value
GLM	Generalized Linear Regression Model
HBL	Hierarchical Binomial Logistic Model
ICC	Intra-class Correction Coefficient
IIA	Independence of Irrelevant Alternatives
IID	Independently and Identically Distributed
IRR	Incidence Rate Ratio
LTA	Land Transport Authority
MCMC	Markov Chain Monte Carlo algorithm
ML	Multiple Linear Regression Model
MLE	Maximum Likelihood Estimation

MPSE	Mean Predictive Square Error
NB	Negative Binomial Regression Model
OBL	Ordinary Binomial Logistic Model
REP	Random Effect Poisson Model
REZIP	Zero-inflated Poisson Model with Random Effects
S.D.	Standard Deviation
SPF	Safety Performance Function
TCS	Traffic Computer System
ZIP	Zero-inflated Poisson Model
ZIPS	Zero Inflated Power Series

LIST OF SYMBOLS

α_i	the random location-specific effects assumed to be independently and identically distributed at the location level
$\boldsymbol{\beta}$	A vector of estimable coefficients representing the effects of the covariates
β_{0j}	The intercept term of j^{th} crash in individual level model of CSPM
β_{pj}	The p^{th} regression coefficients of j^{th} crash in individual level model of CSPM
γ_{00}	The intercept term for regressing β_{0j} in crash level model of CSPM
γ_{p0}	The intercept term for regressing β_{pj} in crash level model of CSPM
γ_{0q}	The qth regression coefficient for regressing β_{0j} in crash level model of CSPM
γ_{pq}	The q^{th} regression coefficient for regressing β_{pj} in crash level model of CSPM
δ_{it}	A term representing the exponential value of ε_{it}
ε_{it}	Random effect error term in the NB model uncorrelated with $\boldsymbol{X}_{it}$
$\Phi(\cdot)$	The cumulative distribution function of the standard normal distribution
λ_{it}	The modified Poisson parameter for random effects
μ	The mean of a Poisson distribution
μ_{it}	The expected number of events of an observation unit *i* in a given time

	period *t* in the Poisson regression model
$\tilde{\mu}_{it}$	The expected number of events of an observation unit i in a given time period t in the NB regression model
π_i	The probability of $Y_i = 1$ in Binomial distribution
(θ, θ)	The parameter for gamma distribution of α_i
$\boldsymbol{\theta}$	A vector of estimable coefficients representing the effects of the covariates $\boldsymbol{A}_{it}$ in ZIP model
σ_i	$\ln(\alpha_i)$
τ	Sharp parameter in ZIP(τ) model
τ_0	The variance of the random effects u_{0j}
ψ_i	Location-specific random effect in Logit part of REZIP model
$\prod_{i=1}^{n}(\cdot)$	Product of given function from 1 to n observations
$\sum_{i=1}^{n}(\cdot)$	Summation of a given function from 1 to n observations
$\boldsymbol{A}_{it}$	Covariates vector in Logit part of ZIP model
d_{im}	A set of m dummy variables only one of which is equal to 1 for any observation
$D^{(\backslash s(it))}$	The remaining data set except $s(it)$
$E_{I(0.025,0.975)}$	95% Bayesian credible interval of predictive mean in MCMC simulation
$E(\cdot)$	Mean or expected value

$g(\delta)$	The distribution of δ
i	The index for observation site or individual
k	Overdispersion parameter in NB model
l_{it}	Indicator variable in ZIP model
Logit(π_i)	$\log(\pi_i/(1-\pi_i))$
m_i	An arbitrary variable used to calculate Vuong statistics
$\overline{m}$	Mean value of m_i
M	The prior knowledge in the model specification
$n(f)$	the number of actual observed frequency of " f " in CV
n_{it}	Observed number of events of an observation unit i in a given time period t
N	The total number of observation
p	Probability of success in Bernoulli trial
p_{it}	The probability of zero crash state in ZIP model
Probit(π_i)	The inverse of the standard cumulative normal distribution function (π_i)
$\Pr(n_{it} \mid \mu_{it})$	Probability density function of n_{it} given the value of μ_{it}
$\hat{\Pr}_1(n_{it} \mid \mu_{it})$	Predicted probability of observing n_{it} based on zero-inflated count data model
$\hat{\Pr}_2(n_{it} \mid \mu_{it})$	Predicted probability of observing n_{it} based on standard Poisson or NB regression model

q	Probability of failure in Bernoulli trial
$R_{it}(0)$	Poisson probability with zero crash
$s(it)$	A sub-group of the observed data set
S_m	Standard deviation of m_i
t	The index for observation period
T_i	The observation number for observation unit i
$u(f)$	Disaggregate predictive probability-based utility
u_{0j}	Within-crash random effects of β_{0j}
u_{pj}	Within-crash random effects of β_{pj}
u_{it}	Utility function in CV
$V(\cdot)$	Variance
V	Vuong statistics
(V, B)	Latent variable in data augmentation step in BI
X	Random variable in Bernoulli trial
$\boldsymbol{X}_{it}$	A vector of covariates for observation unit i in a given time period t
X_{pij}	The pth covariate for ith driver-vehicle unit in jth crash in individual level CSPM model
y_{ij}	Binary severity variable for the ith driver-vehicle unit in jth crash
y_{it}	The observed dependent variable in an observation unit i in a given time period t

$\hat{y}_{it}$	The predictive value of y_{it}
y^*	The latent dependent variable
Z	Number of successes out of N Bernoulli trials
Z_{qj}	The q^{th} covariate of the j^{th} crash in crash level model of CSPM

CHAPTER ONE

INTRODUCTION

1.1 THE PROBLEM

Road safety is a socio-economic concern. With the rapid development of motorization in the past 50 years, the increase of road traffic crashes has become one of the major global health problems. Worldwide, an estimated 1.2 million people are killed in road crashes each year and as many as 50 million are injured (Peden et al., 2004). International studies ranked road traffic crashes as the ninth most serious cause of death in the world in the year 1990. It was forecasted that without increased efforts and new initiatives, the total number of causalities on the roads will increase by some 60% in 2020 and as much as 80% in low income and middle-income countries, which will by then be the third most serious cause of death.

From the economic perspective, the magnitude of road traffic crashes places a huge economic burden on society. For example, in 2005, there were 172 fatal, 71 serious injuries, 6,463 slight injuries, and 81,580 Properties-Damage-Only (PDO) crashes in Singapore. A scientific estimate (Chin, 2007) showed that the total cost of road crashes occurring in 2005 is S$527.25 million, which is about 0.3% of the year's GDP in Singapore. The estimated cost per fatal crash is S$837,475.

Due to the tremendous life and property loss, more and more attention has been placed in various ways on improving the road safety situations. One important way is traffic

safety management. Based on the understanding of the traffic system properties, and integrated with other transport functions, traffic safety management is targeted to developing, implementing, and assessing road safety countermeasures. To ensure the cost-effectiveness of source location, traffic authorities always desire to identify where the most serious "problem" sites are, and to know whether the proposed countermeasures will work or are working effectively. However, it is sometimes very difficult to obtain a comprehensive understanding of traffic system safety because road traffic is such a complicated system, which may be affected by a diversity of risk factors including environmental situations (e.g. weather, street lighting), geometric features (e.g. the layout on the roadway and roadside, the grade), traffic conditions (e.g. traffic volume), regulatory measures (e.g. signals), and driver and vehicle characteristics (e.g. driver age, driver gender, vehicle type, in-vehicle safety protection measures). Moreover, the understanding of traffic system safety may be further obscured since crash occurrences are necessarily discrete, often sporadic and random events. Hence, obtaining unbiased estimation and prediction of traffic system safety has become the central concern for research as well as for practical purposes in road safety management. In practice, the need to obtain estimates of system safety specifically arises from:

1) Entity identification which deviates from a norm and requires rectification,

2) Assessment of the effects of safety countermeasures,

3) Evaluation of standards, programs, rule-making or policies either prospectively or retrospectively, and

4) Other unspecified occasions.

1.2 RESEARCH BACKGROUND

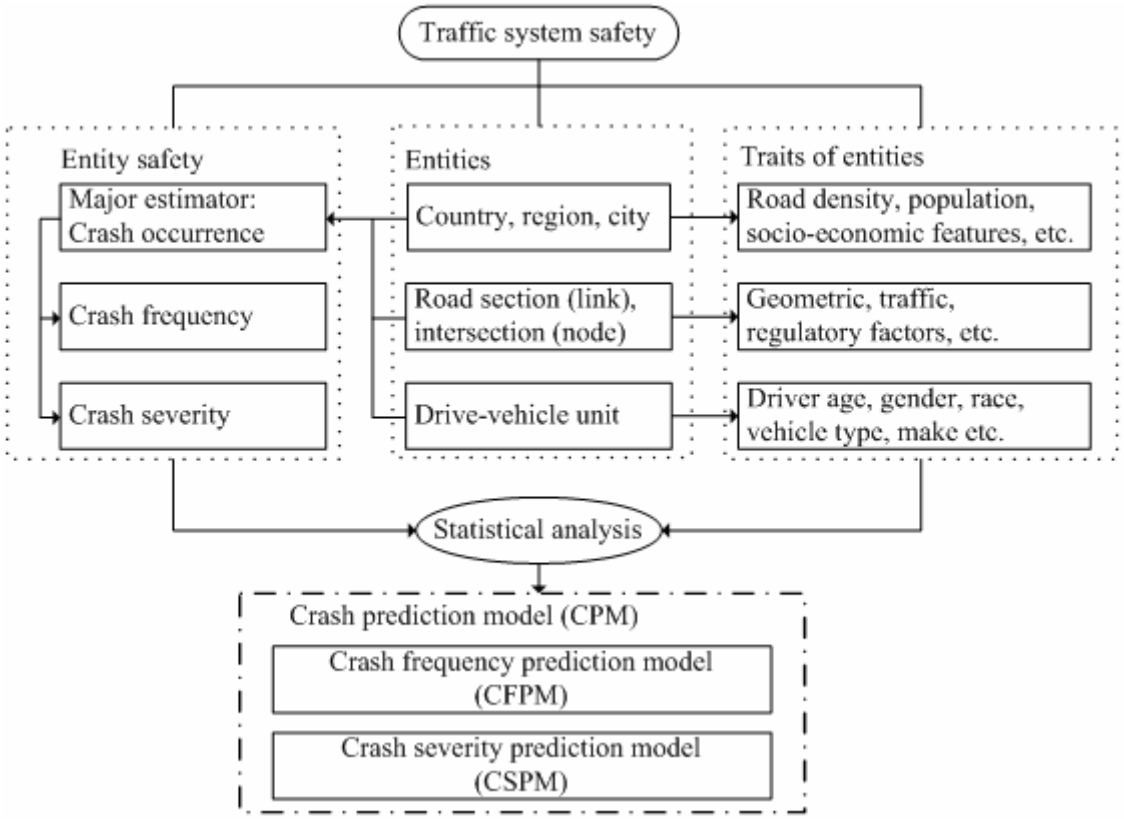

Figure 1.1 Mind Map of the Research Background

Traffic system consists of entities which are differentiated by a variety of traits. For example, as shown in the Figure 1.1, traffic facilities in a country, region, or city can be viewed as one such entity in some macroscopic analysis. The traits for this kind of entity can be such factors as road density, population, and some other social-economic features. Traffic entities can be, more intuitively, a road section or an intersection, with various geometric, traffic, and regulatory factors as traits. Furthermore, a driver-vehicle unit can also be treated as an entity, with traits of driver age, gender, annual distance traveled, vehicle type, make and so on. Most studies of traffic system safety tend to focus on one or several specific entities. While some researchers conduct the

regional evaluation on road safety, some others focus on the microscopic analysis of driving behaviors. Hence, traffic system safety analysis is more or less equivalent to understanding the safety of various particular traffic entities and their interactions.

Although the methods to estimate the system safety vary in a wide range, most studies on road safety have relied on traffic crash statistics to address a range of the above-mentioned safety-related concerns. Hauer (1992) defined system safety as the expected number of crashes in each severity class, which is a characteristic property of a certain system during a specific period of time. Since crash occurrence is likened to a symptom of some undesirable problems in the traffic system, it is reasonable to assume that the answers to such problems can be obtained by examining the symptoms, i.e. the frequency and severity of crash occurrence (Chin and Quek, 1997).

Since traffic entities can be characterized by their traits, either observable or unobservable, it is the usual practice in safety research to establish a statistical relationship between these traits in crash causation and the crash occurrence. This safety statistical model is called as crash prediction model (CPM), which is the major concern of this thesis. Some other researchers also define this kind of models as safety performance function (SPF). The term “crash prediction model” will be used consistently in the rest of this thesis.

Frequency and severity are two major concerns in understanding the relationship of crash occurrence and various risk factors (Hauer, 2006). CPMs are developed to estimate and predict the crash frequency as well as the crash severity. In this thesis, the prediction models for crash frequency and severity are termed “crash frequency

prediction model" (CFPM) and "crash severity prediction model" (CFSM), respectively. A significant number of studies have been conducted on investigating the suitability of various CPMs.

1.2.1 Crash Frequency Prediction Models (CFPM)

Researchers have been using various statistical techniques to model the crash frequency, ranging from the use of multiple linear regression models (ML) to methods involving exponential distribution families such as Poisson and negative binomial (NB) regression models. It has been observed that for random, discrete, nonnegative and sporadic crash data, ML models have several undesirable statistical limitations such as the assumption of normality (Jovanis and Chang, 1986; Joshua and Garber, 1990; Miaou and Lum, 1993). To overcome the problems associated with ML models, Jovanis and Chang (1986) proposed the Poisson regression model, which showed the advantages of Poisson model over linear regression technique in modeling the crash frequency.

Poisson distribution also suffers from an important limitation. Poisson regression model may be appropriate only when the mean and the variance of the crash frequencies are approximately equal, which is a basic property of Poisson process. But this latent assumption has been denied in many traffic studies (e.g. Miaou, 1994; Shankar et al., 1995; Vogt and Bared, 1998), in which the variance of the crash frequency is significantly greater than the mean. To overcome this over-dispersion problem, NB model has been found to be more suitable than Poisson model by introducing a stochastic component to relax the mean-variance equality constraint

(Lawless, 1987; Miaou, 1994; Shankar et al., 1995; Poch and Mannering, 1996; Barron, 1998).

1.2.2 Crash Severity Prediction Model (CSPM)

To account for the nominal or ordinal features of crash severity data, categorical data analysis techniques for discrete dependent variables have generally been employed in most previous crash severity studies. While some researchers (Mannering and Grodsky, 1995; Shankar and Mannering, 1996; Mercier et al., 1997; Al-Ghamdi, 2002) used binomial/multinomial logit or probit models to explore the significance of risk factors by taking crash severity as a nominal, some others (O'Donnell and Connor, 1996; Quddus et al., 2002; Rifaat and Chin, 2005; Abdel-Aty and Keller, 2005) employed ordered logit or probit models to account for the ordered nature of severity levels.

1.3 RESEARCH PROBLEMS

1.3.1 Multilevel Data Structure

As shown above, generalized linear regression models (GLM) are traditionally used in both CFPM and CSPM. While those GLMs adapt appropriate dependent variables to the specific features of crash frequency or severity, they suffer from the underlying limitation that all samples in the dataset are assumed to be independent of one another. However, in crash data generating process or collecting process, there are often hierarchies between the different samples, which imply some unobserved heterogeneities due to multilevel data structure.

Specifically, in CFMP, Poisson and NB distributions are incapable of taking into account some unobserved heterogeneities due to spatial and temporal effects of crash data. In particular, in both Poisson and NB models, it is presupposed that the crash occurrence distributions for the sites with similar observed characteristics are the same. Furthermore, crash counts for a specific location in different time periods are assumed to be independent of one another. But indeed, some hidden features may necessarily exist between different traffic sites and crash occurrences for a specific site may often be correlated serially. Consequently, without appropriately accounting for the location-specific effects and potential serial correlations, the standard errors in the regression coefficients may be underestimated.

In CSPM, the techniques used in most past studies, assuming independence between samples (e.g., a crash or a driver), also suffer from limitations in some special data structure with present of clustering data. For example, it is reasonable to assume that the characteristics of the vehicles within which casualties are traveling will affect their probability of survival. If this is the case, then casualties within the same vehicle would tend to have more similar severity than casualties within different vehicles, and the assumption of residual independence will not be met. The same argument may be extended to encompass the effect of similarities between different crashes, road sections, or geographical regions. Hence, the models without considering the within-cluster correlations, especially when the correlations exist significantly, would result in inaccurate or biased estimates for factor effects.

1.3.2 Excess Zeros in Count Data

Another challenge with existing CFPM is the distribution of excess zero crash observations in some crash data. It is obvious that the distribution of annual crash frequencies with extra zeros may be qualitatively different from the simple Poisson and parent NB distribution (Shankar et al., 1997). If the Poisson or NB distributions are applied in this case, estimation may be mistakenly regarded as the presence of over-dispersion in the data whereas over-dispersion may merely be a natural result of an incorrectly specified model.

To better reflect this special situation, Lambert (1992), in his study on defects in manufacturing, introduced a technique called zero-inflated model by proposing a dual-state system. In recent years, this technique has been employed successfully in road crash frequency prediction (e.g. Miaou, 1994, Shankar et al., 1997, Chin and Quddus, 2003). However, the zero-inflated models are also incapable of accommodating the within-location correlation as well as between-location heterogeneities associated with multilevel data structure. Hence, it would also be interesting whether the accounting of multilevel structure into zero-inflated model will further improve the performance of CFPM.

1.4 RESEARCH OBJECTIVE, METHODOLOGY AND SCOPE

1.4.1 Research Objectives

Based on the identified research problems, two main objectives are formulated for this research, which are:

a) to examine and model the multilevel data structure in CPMs, i.e. CFPM and CSPM.
b) to explore a theoretical framework to determine the suitability of applying various safety statistical models in predicting crash frequency and severity.

1.4.2 Methodology

To achieve the above objectives, hierarchical models that allow multilevel data structure to be properly specified and estimated, are employed. Specifically, in CFPM, based on the investigation of traditional count models such as Poisson and NB models, innovative microscopic traffic crash prediction models are developed to capture both multilevel data structure and excess zero crash observations in the crash frequency data. This is done by developing the random effect Poisson model (REP), the zero-inflated Poisson model (ZIP), and zero-inflated Poisson model with random effects (REZIP). As for CFSM, a hierarchical binomial logistic model (HBL) is proposed to account for the within-cluster correlation of crash severity.

In model calibration, this study develops Bayesian inference (BI) with Markov Chain Monte Carlo (MCMC) algorithm to estimate the proposed models. In Bayesian models,

given model assumptions and parameters, the likelihood of the observed data is used to modify the prior beliefs of the unknowns, resulting in the updated knowledge summarized in posterior densities. BI has intrinsic advantages in explicitly accounting for hierarchical structure over likelihood-based estimation due to its potential to model all sources of sampling uncertainty in the hierarchical models (Congdon, 2003). Due to the absence of built computing programme, the Bayesian inferences for the proposed models are innovatively realized by programming using BUGS language (Bayesian Inference Using Gibbs Sampling).

A number of statistical measures in the Bayesian framework are proposed to assess the suitability of the proposed models, such as Deviance Information Criterion (DIC) and cross validation predictive densities (CV). Furthermore, an Intra-class Correlation Coefficient (ICC) is employed to estimate the proportions of variances associated with different levels and hence to examine the advantage of the hierarchical models over the traditional models. Moreover, the proposed methods are illustrated and validated using Singapore intersection data. After identifying the critical factors contributing to crashes at intersections, possible causes and potential countermeasures for each of the identified factors are discussed and suggested.

1.4.3 Scope of the Study

While the proposed method may apply to most traffic crash situations on various roadway types, the statistical models developed in this study are mainly illustrated on the prediction of traffic crash frequency and severity at urban signalized intersections. The models are based on police recorded crash data and field survey data for geometric, traffic and regulatory characteristics. In CFPM, a total of 52 signalized intersections are sampled which are supposed to be representative of all intersections in Singapore.

Although we proposed the full hierarchical models, only random intercept models are illustrated to avoid excess complexity as the large set of covariates are used. The random effects on covariate coefficient can be easily extended within the proposed methodological framework.

1.5 ORGANIZATION OF THE THESIS

This thesis is organized under seven chapters as structured in Figure 1.2.

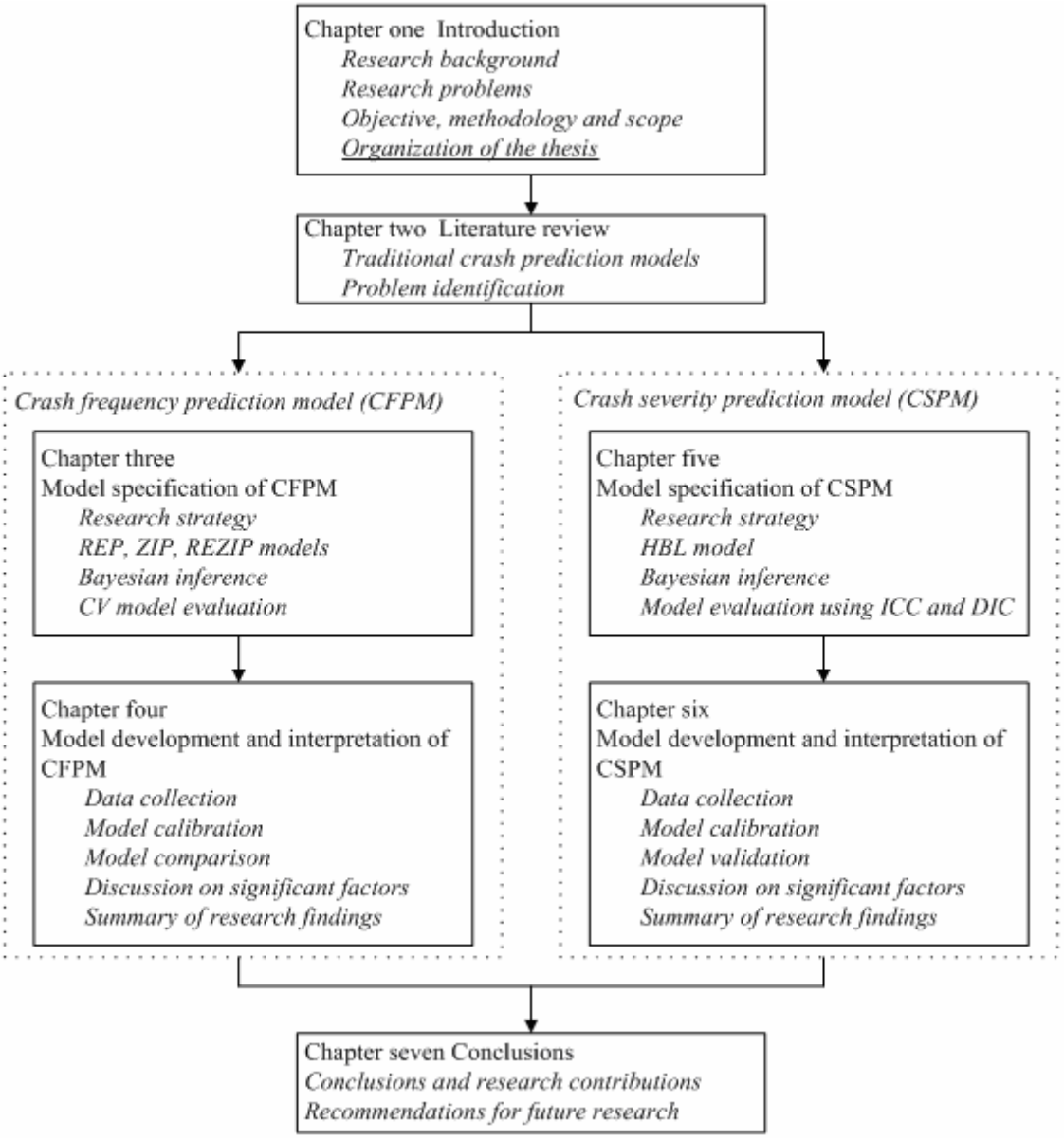

Figure 1.2 Structure of the Thesis

Chapter 1 is the introductory chapter which provides the research background, identifies the research problems, lays out the research objective, methodology and scope, and finally presents an outline of the thesis.

Chapter 2 provides a critical literature review for traditional CPMs. The research problems are specified in details and some existing solutions on the identified problems are also reviewed.

Chapter 3 and Chapter 4 are the crash frequency prediction model development. While Chapter 3 describes the methodology formulation of modeling multilevel data and excess zeros in CFPM, Chapter 4 summarizes an illustrative example for the proposed method using Singapore intersection data.

Chapter 5 and Chapter 6 are the crash severity prediction model development. Specifically, Chapter 5 proposes a Bayesian HBL model in modeling the multilevel data structure in crash severity. Chapter 6 uses the proposed method to examine the severity of driver injury and vehicle damage in traffic crashes at intersections using Singapore crash data.

Finally, conclusions derived from the analysis are summarized in Chapter 7, where research contributions and recommendations for further research are appended.

CHAPTER TWO

REVIEW OF CRASH PREDICTION MODELS

2.1 INTRODUCTION

Statistical modeling is a process of exploring and identifying the potential interrelationships of response variables and the explanatory variables in probabilistic forms. In road safety research, the wildly-used crash prediction model (CPM) is specifically targeted to examining the behavior of crash occurrence, including crash frequency and crash severity, for traffic entities. A variety of traits associated with the entities, as shown in Figure 1.1, are assumed to provide information on the behavior of the crash occurrence. Appropriate probabilistic forms and statistically significant traits are identified based on the examination of crash occurrence mechanism and model fitting performance on historical data.

In particular, crash frequency prediction model (CFPM) is developed when the crash frequency for the traffic entities is concerned, while crash severity prediction model (CSPM) is employed when the crash severity is focused. The fitted models of crash occurrence are useful in estimating the safety situation of traffic entities, in predicting the safety performance of existing or planning highway facilities, in providing information for safety countermeasure development and assessment and so on.

This chapter presents a critical review on traditional CFPM and CSPM. These include general description of crash occurrence mechanism, mathematical formulations, general forms, assumptions and potential weakness of conventional models, i.e.

Poisson and negative binomial (NB) regression models for CFPM and logit, probit and ordered models for CSPM.

2.2 CRASH FREQUENCY PREDICTION MODEL (CFPM)

2.2.1 Crash Occurrence Mechanism

A traffic crash is, in theory, the result of a Bernoulli trial. Each time a vehicle enters an intersection, a highway segment, or any other type of entity (a trial) on a given transportation network, it will either crash or non-crash. For purposes of consistency, a crash is termed a "success" while non-crash is a "failure". For the Bernoulli trial, a random variable, defined as X , can be generated with the following probability model: if the outcome is a "success" (e.g. a crash), then $X = 1$, whereas if the outcome is a "failure", then $X = 0$. Thus, the probability model becomes

Table 2.1 Crash Occurrence as a Bernoulli Trial

X	1	0
$\Pr(x = X)$	p	q

where p is the probability of success (a crash) and $q = (1 - p)$ is the probability of failure (non-crash).

In general, if there are N independent trails (vehicles passing through an intersection, road segment, etc.) that give rise to a Bernoulli distribution, then it is natural to consider the random variable Z that records the number of successes out of the N trials.

Under the assumption that all trials are characterized by the same failure process, the appropriate probability model that accounts for a series of Bernoulli trials is known as the binomial distribution, and is given as:

$$\Pr(Z = n) = \binom{N}{n} p^n (1-p)^{N-n} \tag{2.1}$$

where $n = 0, 1, 2, ..., N$. In Equation (2.1), n is defined as the number of crashes (successes). The mean and variance of the binomial distribution are $E(Z) = Np$ and $VAR(Z) = Np(1-p)$ respectively.

For typical motor vehicle crashes where the event has a very low probability of occurrence and a large number of trials exist (e.g. million entering vehicles, vehicle-miles-traveled, etc.), it can be shown that the binomial distribution is approximated by a Poisson distribution. Under the binomial distribution with parameters N and p, let $p = \mu / N$, so that a large sample size N will be offset by the diminution of p to produce a constant mean number of events μ for all values of p. Then as $N \to \infty$, it can be shown that

$$\Pr(Z = n) = \binom{N}{n} \left(\frac{\mu}{N}\right)^n \left(1 - \frac{\mu}{N}\right)^{N-p} \cong \frac{\mu^n}{n!} e^{-\mu} \tag{2.2}$$

where, μ is the mean of a Poisson distribution. This approximate lends a reasonable support to the use of Poisson regression model in estimating the crash frequency.

On the other hand, the Poisson approximation to the binomial distribution in crash occurrence may also be understood from the aspect of traffic entity. Traffic crash occurrence in a traffic entity, e.g. an intersection or a road segment, is random, discrete and sporadic events that may follow Poisson process. Specifically, dividing the year into 8760 one-hour periods, the chance that more than one crash will occur in any single hour is negligible and the occurrence of crashes is likely to be independent for the different hours. The hourly number of crashes would then be binomially distributed with Binomial $(8760, p)$ where p is the probability of a crash in any given hour. Since p is very low, this distribution is extremely close to the Poisson distribution with the mean of $(8760 \times p)$. Even when the crash probability is indeed variable from one hour to the next, the number of crashes will still have approximately a Poisson distribution.

Consequently, by assuming the crash occurrence as Poisson process, the Poisson distribution has been commonly employed to describe the crash frequency at various traffic entities. When considering the variations of the process associated with different traits of entities, Poisson regression model have been thus conventionally adapted in a number of CFPM studies (e.g. Maycock and Hall, 1984; Jovanis and Chang, 1986; Joshua and Garber, 1990; Jones, Janseen, and Mannering, 1991; Miaou and Lum, 1993). The assumptions and mathematical forms of Poisson regression model are briefly reviewed in the following.

2.2.2 Poisson Regression Model

As discussed in the crash occurrence mechanism, Poisson distribution may be a reasonable description for crash occurrence when crashes are considered to occur both randomly and independently in time. The Poisson distribution has only one adjustable parameter, namely the mean of the distribution μ, which must be positive. This requirement may be unsatisfactory in the case of an additive model, in which the μ does not necessarily have a lower bound. To ensure μ to be positive, a commonly used formulation is a log-linear relationship between the expected numbers of crashes in an observation unit i in a given time period t, i.e. μ_{it} and the covariates $\boldsymbol{X}$, which is

$$\mu_{it} = E(y_{it}) = \exp(\boldsymbol{X}_{it}\boldsymbol{\beta}) \tag{2.3}$$

where, $\boldsymbol{X}_{it}$ is a vector of covariates (traits) which describe the characteristics of a observation unit i (traffic entity, e.g. an intersection, a road segment) in a given time period t (e.g. annual) and $\boldsymbol{\beta}$ is a vector of estimable coefficients representing the effects of the covariates. Note that y_{it} is the number of observing crashes in an observation unit i in a given time period t. Therefore, the probability of observing y_{it}, when μ_{it} is given, can be expressed as

$$\Pr(y_{it} \mid \mu_{it}) = \frac{\exp(-\mu_{it})\mu_{it}^{\,y_{it}}}{y_{it}!} \tag{2.4}$$

where μ_{it} is a deterministic function of $\boldsymbol{X}_{it}$ and randomness in the model comes from the Poisson specification for y_{it}.

To estimate μ_{it}, i.e. $\boldsymbol{\beta}$, which is the effect of the covariates on the dependent variable, the method of maximum likelihood estimation (MLE) is commonly used (Green, 1997). In general, the likelihood function for independently Poisson-distributed random variables is

$$L(\mu_{it} \mid y_{it}) = \prod_{i=1}^{N} \prod_{t=1}^{T} \frac{\exp(-\mu_{it})\mu_{it}^{\ y_{it}}}{y_{it}!} \tag{2.5}$$

The basic idea of maximum likelihood is that given the data, an estimate of $\boldsymbol{\beta}$ can be determined by maximizing this function and hence the likelihood of having generated the data (King, 1989).

Correspondingly, the log-likelihood function is then

$$\begin{aligned} l(\boldsymbol{\beta}) &= \ln(L(\mu_{it} \mid y_{it})) \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} [y_{it} \ln(\mu_{it}) - \mu_{it} - \ln(y_{it}!)] \end{aligned} \tag{2.6}$$

Standard numerical maximization methods can easily be applied to this globally concave function by using one of many computer programs (e.g. Greene, 1995).

However, the Poisson regression model has some potential problems in describing the crash process. One important constraint is that the mean must be equal to the variance.

If this assumption is not valid, the standard errors will be biased and the test statistics derived from the model will be incorrect. Many researchers have modified the simple Poisson assumption by assuming that the parameter is distributed, usually in a Pearson type III distribution. A historical and bibliographical account of the problem associated with the use of the Poisson model has been well documented (Haight, 1967). In a number of recent studies (Miaou, 1994; Shankar et al., 1995; Vogt and Bared, 1998), the crash data were found to be significantly overdispersed, i.e. the variance is much greater than the mean. This will result in incorrect estimation of the likelihood of crash occurrence.

In overcoming the problem of over-dispersion, several researchers, like Miaou (1994), Kulmala (1995), Shankar et al. (1995), Poch and Mannering (1996), and Abdel-Aty and Radwan (2000) have employed the NB distribution instead of the Poisson. By relaxing the condition of mean equals to variance, NB regression model is more suitable in describing discrete and nonnegative events. The mathematical formulation of NB regression model is described in the following.

2.2.3 Negative Binomial Regression Model

To overcome the over-dispersion problem, the NB regression model relaxes the "equality" constraint between mean and variance by introducing a stochastic component into the Poisson model even though the source of over-dispersion in event count data cannot be distinguished (which will be discussed in detail in the later section of this chapter). Mathematically, the Equation (2.3) can be rewritten as

$$\widetilde{\mu}_{it} = \exp(\boldsymbol{X}_{it}\boldsymbol{\beta} + \varepsilon_{it}) \tag{2.7}$$

where ε is a random error that is assumed to be uncorrelated with $\boldsymbol{X}$. Hence, the relationship of $\widetilde{\mu}$ and original μ in Poisson model follows readily

$$\begin{aligned}\widetilde{\mu}_{it} &= \exp(\boldsymbol{X}_{it}\boldsymbol{\beta})\exp(\varepsilon_{it}) \\ &= \mu_{it}\exp(\varepsilon_{it}) \\ &= \mu_{\mathrm{it}}\delta_{it}\end{aligned} \tag{2.8}$$

where δ_{it} is defined to equal $\exp(\varepsilon_{it})$. An assumption needs to be made about the mean of the error term (δ_{it}) to identify NB regression model (Long 1997). The most convenient assumption is that

$$E(\delta_{it}) = 1 \tag{2.9}$$

which implies that the expected count after adding the new source of variation is the same as it was for the Poisson regression model, i.e.

$$\begin{aligned}E(\widetilde{\mu}_{it}) &= E(\mu_{it}\delta_{it}) \\ &= \mu_{\mathrm{it}}E(\delta_{it}) \\ &= \mu_{\mathrm{it}}\end{aligned} \tag{2.10}$$

The distribution of observations given X and δ is still Poisson, i.e.

$$\Pr(y_{it} \mid \boldsymbol{X}_{it}, \delta_{it}) = \frac{\exp(-\tilde{\mu}_{it})\tilde{\mu}_{it}^{\,y_{it}}}{y!} = \frac{\exp(-\mu_{it}\delta_{it})(\mu_{it}\delta_{it})^{y_{it}}}{y_{it}!} \tag{2.11}$$

However, since δ is unknown we cannot compute $\Pr(y \mid \boldsymbol{X}, \delta)$ and instead need to compute the distribution of y_{it} given only $\boldsymbol{X}$. To compute $\Pr(y \mid \boldsymbol{X})$ without conditioning on δ, we average $\Pr(y \mid \boldsymbol{X}, \delta)$ by probability of each of δ. If g is the probability density function (pdf) for δ, then

$$\Pr(y_{it} \mid \boldsymbol{X}_{it}) = \int_0^{\infty} [\Pr(y_{it} \mid \boldsymbol{X}_{it}, \delta_{it}) \times g(\delta_{it})] d\delta_{it} \tag{2.12}$$

The solution of this integral in Equation (2.12) depends on the form of $g(\delta_{it})$. Ideally, the choice of this function reflects some knowledge or theory about the process that generates the over-dispersion. However, such information is rarely, if ever, available. Furthermore, few functions will produce compound Poisson distributions that are computationally tractable. In practice, the gamma distribution is usually chosen. There are two main advantages to this choice. First, the solution to Equation (2.12) that follows from this choice can easily be used to obtain parameter estimates. Second, the gamma distribution is quite flexible. It can vary from highly skewed to symmetric shapes, depending on the values of the two parameters that characterize it.

Assuming that $g(\delta_{it})$ has a gamma distribution with mean 1 and variance k. The resulting probability distribution under the NB assumption is

$$\Pr(y_{it} \mid \mu_{it}, k) = \frac{\Gamma(y_{it} + 1/k)}{\Gamma(1/k)\, y_{it}!}\left(\frac{k\mu_{it}}{1+k\mu_{it}}\right)^{n}\left(\frac{1}{1+k\mu_{it}}\right)^{1/k} \tag{2.13}$$

in which $k(\geq 0)$ is often referred to as over-dispersion parameter. If k reduces to zero then the NB regression model reduces to the Poisson regression model. In this way, the Poisson regression model is nested within the NB regression model and a t-test for $k = 0$ can be used to evaluate the significant presence of over-dispersion in the data. In NB regression model, it is assumed that unconditional mean μ_{it} is independently distributed over time. For this specification, the mean and variance will be respectively

$$E(y_{it} \mid \mu_{it}, k) = \mu_{it} \tag{2.14}$$

$$Var(y_{it} \mid \mu_{it}, k) = \mu_{it}(1 + k\mu_{it}) \tag{2.15}$$

and the mean-variance relationship of the distribution is given by

$$Var(y_{it} \mid \mu_{it}, k) = E(y_{it})[1 + kE(y_{it})] \tag{2.16}$$

Estimation of μ_{it} can be obtained through standard maximum likelihood as mentioned in the previous section and is given by

$$L(\mu_{it}) = \prod_{i=1}^{N}\prod_{i=1}^{N} \frac{\Gamma(y_{it}+1/k)}{\Gamma(1/k)y_{it}!}\left(\frac{k\mu_{it}}{1+k\mu_{it}}\right)^{n}\left(\frac{1}{1+k\mu_{it}}\right)^{1/k} \tag{2.17}$$

This function is maximized to obtain coefficient estimates for $\boldsymbol{\beta}$ and k. Several researchers, like Miaou (1994), Kulmala (1995), Shankar et al. (1995), Poch and Mannering (1996) and Abdel-Aty and Radwan (2000) have employed this NB distribution and they have proved that NB regression model is better than Poisson model in fitting the overdispersed crash data.

However, the NB regression model is not without limitations. As mentioned above, although NB regression model provides an over-dispersion parameter (k) to relax the constraint between mean and the variance of crash data, the source of over-dispersion in event count data cannot be distinguished. Specifically, as with Poisson regression model, NB regression model also assume an "independent" relationship between different observations. When some special data structures are present, i.e. correlations exist in crash data, NB as well as Poisson regression models are obviously not adequate. The next section presents a discussion on the possible sources of over-dispersion in crash data and some existing solutions.

2.2.4 Potential Problems and Existing Solutions

In the Poisson regression model, variation in μ is introduced through observed heterogeneity. Different values of $\boldsymbol{X}$ result in different values of μ. In the NB regression model, variation in $\tilde{\mu}$ is due both to variation in $\boldsymbol{X}$ among individuals but also to the heterogeneity introduced by ε. For a given combination of the values for the independent variables, there is a distribution of $\tilde{\mu}$'s rather than a single μ. However, both Poisson and NB regression models assume the observations are independent with each other. Consequently, some unobserved heterogeneities due to spatial and temporal effects of crash data may not be taken into account appropriately. In particular, in both Poisson and NB regression model, it is presupposed that the crash occurrence distributions for the sites with similar observed characteristics are the same. Furthermore, crash counts for a specific location in different time periods are assumed to be independent with each other.

But indeed, some hidden features may necessarily exist between different traffic sites and crash occurrences for a specific site may often be correlated serially. Traffic crash is a complex event with a large number of factors involved. Ideally, all of the relevant factors should be included in the model. In practice, however, some of the factors may not be available or even collectable for study. A model may only consider the most important factors as independent variables and omit the others. It assumes that similar sites (site with same selected independent variables) have the same mean of crash occurrence. In the real world, however, similar site may be different in omitted factors and thus may have different means. This introduces additional variance to the data and causes the over-dispersion. Consequently, without appropriately accounting for the

location-specific effects and potential serially correlations, the estimates of the standard error in the regression coefficients may be underestimated.

One way to overcome these problems is to treat them in a time series cross-sectional panel with different locations and time periods, as suggested by Hausman et al. (1984) in their study of patent applications. Using the panel data, the hidden features can alternatively be captured by individual (location) heterogeneity. In employing the model in what may be its first application in traffic crash studies, Shankar et al. (1998) showed that the introducing of location-specific random effects and time indicators into the NB regression model can significantly improve the explanatory power of crash models. In recent years, the proposed hierarchical model (random effect model) is increasingly applied to develop the crash prediction models (e.g. Yang, 2003, Chin and Quddus, 2003a).

Another possible source of over-dispersion in existing CPMs is the distribution of excess zero crash observations in some crash data. This "excess zeros" occurs frequently as the outgrowth of three sources: a) crash severity: minor crash may not be reported; b) near crash: it may also indicate a potentially dangerous traffic location even though no crashes have been recorded (Shanker et al., 1997); c) specific types of crashes: some traffic location is possibly safe regarding to specific types of crashes (Chin and Quddus, 2003b).

It is obvious that the distribution of annual crash frequencies with extra zeros may be qualitatively different from the simple Poisson and parent NB distribution (Shankar et al., 1997). If simply applying the Poisson or NB distribution in this case, estimation

may be mistakenly regarded as the presence of over-dispersion in the data whereas over-dispersion may merely be a natural result of an incorrectly specified model. To better reflect the situation, a dual-state system may be assumed. In this, one state is the zero-crash state, in which the traffic location, e.g. an intersection or a roadway section, can be regarded as virtually safe, while the other state is the non-zero-crash states, in which the crash frequencies are assumed to follow some known distributions such as the Poisson and NB.

To handle this dual-state system, Lambert (1992), in his study on defects in manufacturing, proposed a technique called zero-inflated Poisson (ZIP) regression. This ZIP model provides a practical way to explicitly model existence of the two states as well as allow for both the probability of a perfect state (i.e. zero-defect state) and the mean of the imperfect state (i.e. non-zero-defect state) to depend on the covariates. This has been recently applied in a variety of fields to account for the excess zero count data, for example, in applications to sociology (Land et al., 1996), industry (e.g. Xie et al., 2001; Ghosh et al., 2006), management (Karen and Kelvin, 2005), and biomedicine (Hall, 2000).

In traffic analysis field, zero-inflated models are examined and increasingly employed to investigate the relationship between traffic crashes and the covariates. Miaou (1994) started the exploration of using the ZIP structure to analyze crash frequency. Shankar et al. (1997) conducted an empirical inquiry to explore the conditions under which the zero-inflated models are more appropriate than simple Poisson and NB regression models on crash analysis research. In a more recent study, Chin and Quddus (2003b) proposed an evaluation framework to determine the suitability of applying different

count models in crash studies, and they demonstrated that the zero-inflated probability process is an appropriate technique for modeling specific types of crashes in which the data contain many zero counts. Furthermore, the applications of ZIP model have been found in analysis of truck crashes (Miaou, 1994), motor vehicle crashes (Lee et al., 2002; Qin et al. 2005), run-off-road crashes (Lee and Mannering, 2002), pedestrians and motorized traffic crashes, (Shankar et al., 2003), occupational injuries (Wang et al., 2003), crash occurrence at signalized T-intersections (Kumara and Chin, 2003), and crash rate prediction for two-lane highway segments (Xiao et al., 2004).

However, Lord et al. (2005, 2007) have questioned the basic dual-state assumption of zero-inflated models. The essential objection is that no highway is "virtually safe" to allow a non-crash state. They have also provided several reasons for the presence of excess zeros other than the dual-state explanation. Nonetheless, despite the lack of intuitive appeal, zero-inflated models may still be used for three reasons. Firstly, zero crash observations exist everywhere in road network, and no alternative solution is currently available to systematically account for the excess zeros. Secondly, as long as the regression results are not extrapolated beyond the range of observation of the study period, the model may still be valid even though no highway is "virtually safe". Thirdly, in the absence of any better alternative, this model may yet be suitable for prediction rather than estimation purpose.

Although ZIP model is capable of handling the dual-state system in crash data with excess zero observations, it does not accommodate the within-location correlation as well as between-location heterogeneities, which is the basic motivation for the need of hierarchical models. Hall (2000), in his research of zero-runoff sub-irrigation system,

developed the ZIP models with random effects which have demonstrated to fit better than corresponding fixed effects models with zero-inflation and mixed effects models without zero-inflation for the repeated measures and split-plot data sets. Hence, it would be also interesting if a theoretical combination of zero-inflated model and random effect model can further improve the performance of CFPM.

2.3 CRASH SEVERITY PREDICTION MODEL (CSPM)

In addition to the crash frequency, crash severity is another important concern of road safety. Instead of the count data in crash frequency, description of crash severity level is generally associated with the nominal or ordered features. Some generalized linear models (GLM) have generally been employed to account for these features in most previous crash severity studies. In particular, while some researchers (Jones and Whitfield, 1988; Lui et al., 1988; Shibita and Fukuda, 1994; Mannering and Grodsky, 1995; Shankar and Mannering, 1996; Mercier et al., 1997; Simoncic, 2001; Al-Ghamdi, 2002) used binomial/multinomial logit/probit model to explore the significance of risk factors by taking crash severity as a nominal variable, some others (O'Donnell and Connor, 1996; Quddus et al., 2002; Rifaat and Chin, 2005; Abdel-Aty and Keller, 2005) employed ordered logit/probit models to account for the ordered nature of severity levels. This section presents a critical review on these models conventionally used in crash severity studies.

2.3.1 Logit and Probit Models

The logit and probit models are commonly used when the crash severity is classified as nominal categories. Binary logit/probit models are applied when accounting only two states of severity level, for example, injury or non-injury, fatal or non-fatal; while multinomial logit/probit models extend the analysis on two states to multiple states of severity levels.

In modeling the two-state severity levels using binary logit/probit models, the dependent variable Y for i^{th} observation unit (e.g. a crash, a driver) can only takes one of two values: $Y_i = 0 \; or \; 1$ representing the two states of severity levels respectively. The binary logit model denotes the probability of $Y_i = 1$ by $\pi_i = \Pr(Y_i = 1)$, which follows a binomial distribution. A logistic transformation can be interpreted as the logarithm of the odds of severity level 1 vs. severity level 2. The logistic transformation of the probability π_i is given by

$$\text{Logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \tag{2.18}$$

The binary logit model is obtained by treating the Equation (2.18) as a link function in the generalized linear model framework,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{X}_i\boldsymbol{\beta} \tag{2.19}$$

So probability π_i can be solved:

$$\pi_i = \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{1-\exp(\boldsymbol{X}_i\boldsymbol{\beta})} \tag{2.20}$$

where, $\boldsymbol{X}_i$ is a vector of explanatory variables such as geometric, traffic, and situational factors, as well as the driver-vehicle characteristics which are assumed to have effects on severity level. $\boldsymbol{\beta}$ is the effect coefficient vector of the explanatory variables. For all

possible values of $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$, the logistic transformation ensures that π remains in the [0, 1] interval. As π approaches 0, logit(π) tends toward $-\infty$; as π approaches 1, logit(π) tends toward $+\infty$.

The binary probit model provides an alternative to the logit model. Again, a nonlinear model in π is transformed so that a monotonic function of π is linear with respect to explanatory variables. The probability π_i is given by the standard cumulative normal distribution function:

$$\pi_i = \int_{-\infty}^{\boldsymbol{\beta}\boldsymbol{X}_i} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)dt = \Phi(\boldsymbol{X}_i\boldsymbol{\beta}) \tag{2.21}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The probit transformation is given by the inverse of the standard cumulative normal distribution function.

Solving Equation (2.21) for $\boldsymbol{X}_i\boldsymbol{\beta}$ yields

$$\text{Probit}(\pi_i) = \Phi^{-1}(\pi_i) = \boldsymbol{X}_i\boldsymbol{\beta} \tag{2.22}$$

Thus, the probit model can be written as

$$\pi_i = \Phi(\boldsymbol{X}_i\boldsymbol{\beta}) \tag{2.23}$$

For the normal and the logistic distribution have similar shapes, probit and logit models are very similar. In practice, the logistic distribution may be preferred due to the simplicity of probability distribution and density functions. In case of crash severity studies, the logit model is preferred because of its ease in interpretation in terms of log-odds ratio which probit model cannot do since probit model has no simple closed-form expression for the odds-ratio.

Though binary logit model is applied broadly in severity studies, it may not be adequate when more than two states of the injury severity are considered. The multinomial logit model extends the logit model to more than two states. For the nominal dependent variable, the multinomial logit model (McFadden, 1973) is the most widely-used discrete choice model due to its simple mathematical structure and ease of estimation. This discrete choice model is based on the principle that an individual chooses the outcome that maximizes the utility gained from that choice. Based on this principle and the assumption that the error term is generalized extreme value (GEV) distributed, McFadden (1981) derived the simple multinomial logit model. The final form of the model is as follows:

$$\pi_i(y_i = j) = \frac{\exp(\boldsymbol{X}_i \boldsymbol{\beta}_j)}{\sum_J \exp(\boldsymbol{X}_i \boldsymbol{\beta}_j)} \tag{2.24}$$

where $\pi_i(y_i = j)$ is the probability of individual i having alternative j in a set of possible choice categories J, $\boldsymbol{X}_i$ is a vector of measurable characteristics that determine alternative j; $\boldsymbol{\beta}_j$ is a vector of statistically estimable coefficients.

However, the multinomial logit model has the limitation of independence of irrelevant alternatives (IIA) (Ben-Akiva and Lerman, 1985), such that the odd of m versus n ($m, n \in 1...J$) is not affected by other alternatives, i.e.

$$\frac{\pi_i(y_i = m)}{\pi_i(y_i = n)} = \exp(\boldsymbol{X}_i[\boldsymbol{\beta}_m - \boldsymbol{\beta}_n]) \tag{2.25}$$

This expression is only a function of the respective utilities of alternatives m and n, and is not affected by the introduction/removal of other alternatives. This analytical feature implies that the relative shares of the two given alternatives are independent of composition of the set of alternatives.

The limitation of IIA in multinomial logit model was also identified by Shankar, Mannering and Barfield (1996), Chang and Mannering (1999), Lee and Mannering (2002) in their studies on crash severity. Shankar et al. (1996) classified severity of a crash to be one of four discrete categories: property damage, possible injury, evident injury and disabling injury or fatality. But according to them, property damage and possible injury crashes may share unobserved effects such as internal injury or effects associated with lower-severity crashes. However, the basic assumption in the derivation of the multinomial logit model is that error terms or disturbances are independent from one crash severity category to another. Shankar et al. (1996) suggested that if some severity categories share unobserved effects (i.e. have correlated disturbances), the model derivation assumptions are violated and serious specification errors will result.

On the other hand, according to Long (1997), a significant advantage of the multinomial probit model is that the errors can be correlated across choices, which eliminates the IIA restriction. However, computational difficulties make the multinomial probit model impractical.

2.3.2 Ordered Logit and Probit Models

When the dependent variable is ordinal in nature, it should not preferably be treated as nominal. Multinomial logit/probit model cannot handle ordinal dependent variable. Consequently, there will be loss of efficiency due to information being ignored. One way to deal with this problem is to use ordered logit/probit model. The ordered logit/probit models discern unequal differences between ordinal categories in the dependent variable (McKelvey and Zavoina, 1975; Greene; 2000).

In crash severity modeling, researchers (e.g., O'Donnell and Connor ,1996; Duncan et al.,1998; Khattak, 2001; Kockelman et al., 2002, Rensky et al., 1999; Quddus et al., 2002) have recognized that the discrete measure of severity is ordinal in nature and have applied the ordered logit/probit models to severity studies. The difference between the two models lies in the assumption of errors. O'Donnell and Connor (1996) and Rensky et al. (1999) have further indicated that the results from the ordered probit and ordered logit are similar. However, ordered probit model is preferable because the assumption that the distribution of errors is normally distributed is more likely to be valid. A mathematical description of ordered probit model is presented in the following.

The ordered probit model is usually motivated in a latent (i.e. unobserved) variables framework. In CSPM, the general form of the ordered model is

$$y_i^* = \boldsymbol{X_i}\boldsymbol{\beta} + \varepsilon_i \tag{2.27}$$

$$y_i = m \quad \text{if } \tau_{m\text{-}1} \leqslant y^* < \tau_m \qquad \text{for } m = 1 \text{ to } M \tag{2.28}$$

where, y represents the crash severity and can be ordered in M several levels (e.g., slight injury, serious injury and fatal) and y^* indicates the injury propensity. $\boldsymbol{X_i}$ is a vector of explanatory variables describing characteristics of the victim, vehicle, crash and the environmental , $\boldsymbol{\beta}$ is a vector of parameters to be estimated and ε_i is the error term.

In Equation (2.27) and Equation (2.28), the latent variable y_i^* ranging from $-\infty$ to $+\infty$ is mapped to an observed ordinal variable y. The threshold values $\tau's$ are unknown parameters to be estimated. The extreme categories, 1 and M, are defined by open-ended intervals with $\tau_0 = -\infty$ and $\tau_M = +\infty$. The mapping from the latent variable to the observed categories is illustrated in Figure 2.1 below:

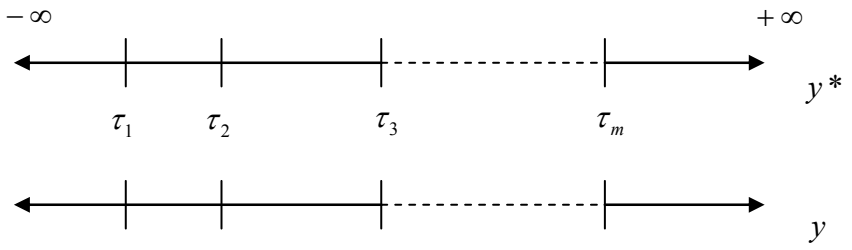

Figure 2.1 Mapping of Latent Variable to Observed Variable

To calibrate the model, distribution of error term (ε) need to be assumed to estimate $\boldsymbol{\beta}$. For the ordered probit model, ε is assumed distributed normal with mean 0 and variance 1. Hence, the probability of a particular value of y_i given $\boldsymbol{X_i}$ can be computed. According to following formulation, the predicted probability of any type of injury severity, m for given $\boldsymbol{X_i}$ is

$$\Pr(y_i = \mathrm{m} \mid \boldsymbol{X_i}) = \Phi(\tau_m - \boldsymbol{X_i\beta}) + \Phi(\tau_{m-1} - \boldsymbol{X_i\beta}) \tag{2.29}$$

The model is unidentified since a change in the β_0 (the first component in $\boldsymbol{\beta}$ representing the intercept) in the structural model can always be compensated for by a corresponding change in the thresholds (τ_1 and τ_2). As suggested by Long (1997), there is an infinite number of parameterizations that could be made to identify the model, only one of two are commonly used that is either β_0 or τ_1 is constrained to 0. The choice of parameters to be used is arbitrary and does not affect $\boldsymbol{\beta}$ or the associated significance tests, as well as the computed probabilities in Equation (2.29).

The contribution to the likelihood for the i th observation depends on which value of severity m is observed. For each of the ordered responses ($m = 1,...,M$), the product over all observations have been taken for which $y = m$ and the likelihood can be written as

$$L = \prod_{i=1}^{n} \prod_{m=1}^{M} \Pr(y_i = m \mid \boldsymbol{X}_i)^{d_{im}} \tag{2.30}$$

where d_{im}=1 if $y_i = m$, and 0 otherwise. Thus, d_{im} define a set of m dummy variables only one of which is equal to 1 for any observation.

Then the final form of the log-likelihood can be written as

$$l(\boldsymbol{\beta}) = \ln(L) = \sum_{i=1}^{n} \sum_{m=1}^{M} d_{im} \ln\left[\Phi(\tau_m - \boldsymbol{X}_i\boldsymbol{\beta}) - \Phi(\tau_{m-1} - \boldsymbol{X}_i\boldsymbol{\beta})\right] \tag{2.31}$$

2.3.3 Potential Problems

As reviewed above, the GLMs, i.e. logit/probit or ordered models, have been applied broadly to account for the nominal or ordered feature of crash severity levels in modeling the crash severity. These models have been proved to be useful in many studies. However, a potential problem arises when factors influencing the severity levels of any individual casualty are seen to be operating at a variety of scales, with these scales comprising successive levels of a hierarchy. These may be associated with the personal characteristics of the casualty at the lowest level of the hierarchy, the features of the vehicle within which they are located or the distinguishing events of the crash in which they are involved. At the highest levels of the hierarchy, they may be extended to the properties of the road section upon which the crash took place; or even the attributes of the geographical region or country where it occurred.

However, since the techniques used in most past studies assumed independence between different observations, these techniques may not be incapable of accounting for the possible within-cluster correlations. Actually, this within-cluster correlation has already been identified in some earlier studies; for example, Evans (1992, 1993) found that in a multiple vehicle crash, the risk of fatality was dependent on the characteristics of the other vehicles. Hence, the models without considering the covariance between individuals in the same cluster (e.g., a same crash), especially when the covariance is significant, would result in inaccurate or biased estimates of factor effects.

2.4 SUMMARY

This chapter presents a critical review on the traditional CPMs. In according to the features of response variable, various GLMs are broadly applied to build probabilistic formulations on the relationship of the crash frequency or severity with a variety of possible covariates, such as geometric, traffic, environmental factors as well as driver-vehicle characteristics.

However, potential problems are identified in both CFPM and CSPM with their applications in certain areas of road crash predictions. One of the most fundamental problems with the application of GLMs is that each observation (e.g. a crash or a vehicle) entered into the estimation procedure corresponds to an individual situation, either for crash frequency or severity. Hence, the residuals from the model exhibit independence. However, a consideration of the data structure suggests that in some cases the assumption of independence may often not hold true with the present of a multilevel structure of crash data.

This possible existence of multilevel structure within crash data is commonly ignored. However, disregarding hierarchies, where they are present, can lead to the production of models giving unreliable estimates of prevision, incorrect standard errors, confidence limits, and tests (Skinner et al. 1989). In the rest of this thesis, methodological formulations using Bayesian hierarchical modeling technique are proposed to take account of potential multilevel data structures in modeling crash frequency and severity. Specific CFPM (in Chapter 3 and Chapter 4) and CSPM (in Chapter 5 and 6) are separately developed, which are illustrated using Singapore intersection data.

CHAPTER THREE

MODELING MULTILEVEL DATA AND EXCESS ZEROS IN CRASH FREQUENCY PREDICTION

3.1 INTRODUCTION

In estimating the crash frequency, the Poisson and negative binomial (NB) models may be incapable of appropriately taking into account the unobserved heterogeneities when some special crash data structures are present. As reviewed in the Chapter 2 multilevel data structure (e.g. repeated observations at same sites for different time periods) and excess zero observations are two critical issues which may violate the latent assumptions in the Poisson process.

To better fit the crash data, some researchers (e.g. Shankar et al., 1998; Yang and MacNab, 2003, Chin and Quddus, 2003a) applied hierarchical data analysis techniques to deal with the multilevel data structures, while some others (e.g. Miaou, 1994; Shankar et al., 1997; Chin and Quddus, 2003b) employed the zero-inflated count model to account for the excess zero crash occurrences. Most of them have proven the effectiveness of techniques employed on improving the predictive performance of crash frequency prediction models (CFPM).

However, the model may still be inadequate if it involves both multilevel data structure and excess zeros in crash frequency prediction. Hence, it is interesting to examine

whether zero-inflated count model with random effects will further improve the existing CFPMs.

This study attempts to propose the use of zero-inflated Poisson model with location-specific random effects (abbreviated as REZIP). Furthermore, theoretical evaluating tools are also developed to determine the suitability of applying different count models (e.g. random effect Poisson model (REP), zero-inflated Poisson model (ZIP), and REZIP model) in road crash frequency prediction.

In model calibration, Bayesian analysis using Markov Chain Monte Carlo (MCMC) algorithm, instead of the classical maximum likelihood estimation (MLE) and likelihood ratio tests, is employed. Many advantages of Bayesian inference (BI) have been known in philosophical as well as practical aspects over the traditional MLE inference, which will be discussed in the section 3.4 of this chapter.

In selecting the appropriate model, the statistical tests of over-dispersion and zero-inflation are used to examine the crash data. Furthermore, model assessment measures based on cross validation predictive densities are proposed, which provide reliable and flexible tools to compare the model fitness between arbitrary non-nested models, e.g. REP, ZIP and REZIP models in this study.

Crash records and site characteristic data at signalized intersections in Singapore are used to illustrate the proposed methodology. The results demonstrate that the REZIP model can significantly improve the predictive performance of crash prediction models

for the subject dataset. The specific research strategy for CFPM development is presented in the following.

3.2 RESEARCH STRATEGY

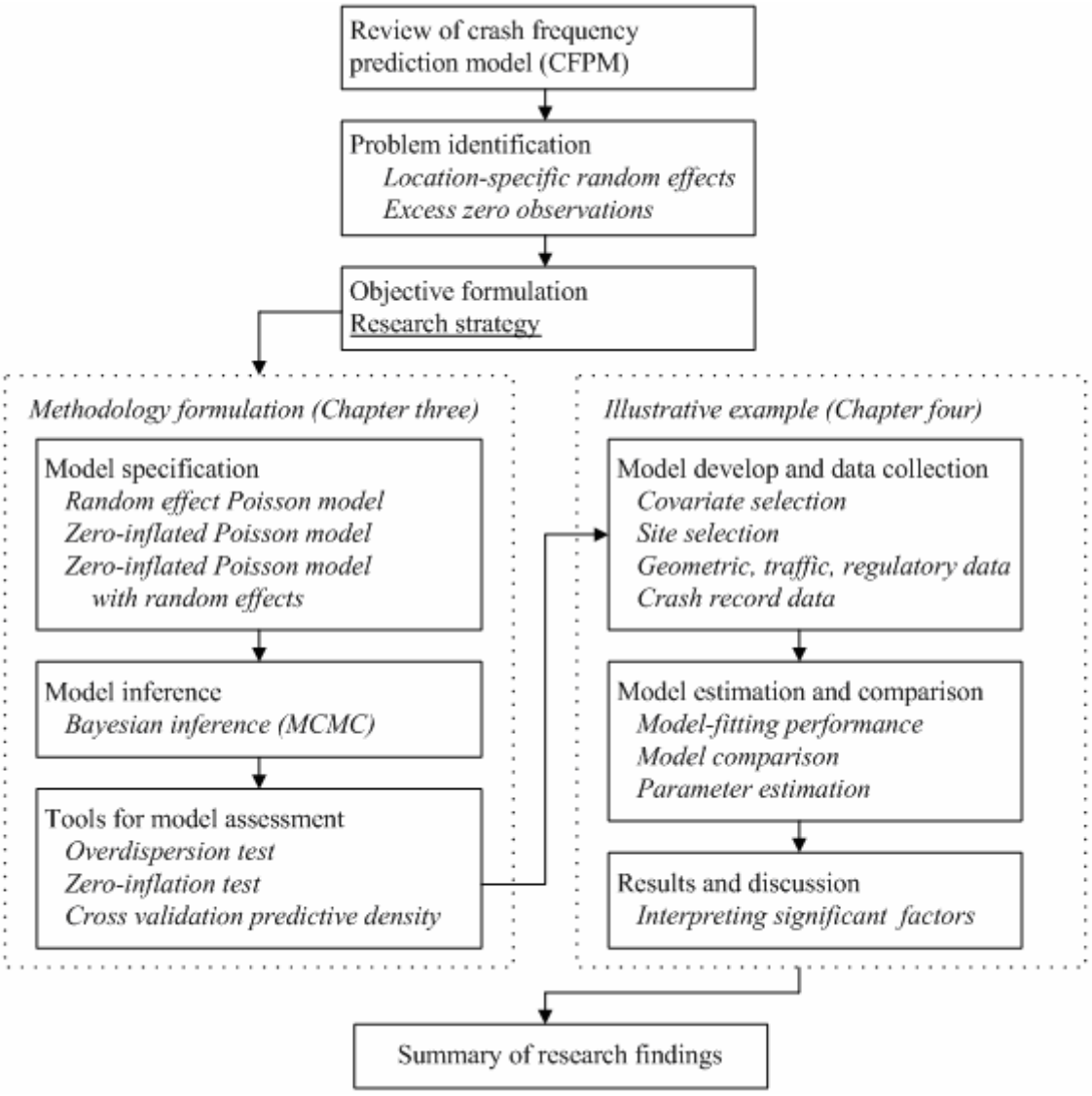

Figure 3.1 Research Strategy for CFPM Development

The research strategy for CFPM development is illustrated in Figure 3.1. After identifying the research problems (see section 2.2 for details) and specifying the objective and research strategy, the proposed methodology is formulated in this chapter, consisting of model specification, model inference, and tools for model assessment. Then, the model development using Singapore intersection data, the model estimation and comparison, and the result interpretation, as well as the summary of research findings will be presented in the Chapter 4.

3.3 MODEL SPECIFICATION

The section presents the model specifications for REP, ZIP, and REZIP models, associated with a discussion of the methodological evolution.

3.3.1 Random Effect Poisson Regression Model

The basic link function of REP model that modifies the Poisson regression model can be described as follows (Hausman et al., 1984):

$$\lambda_{it} = \mu_{it}\alpha_i = e^{\boldsymbol{X}_{it}\boldsymbol{\beta}+\sigma_i} \tag{3.1}$$

where, λ_{it} is the modified Poisson parameter for random effects, μ_{it} is the Poisson parameter representing the expected number of crashes at roadway location i in time period t ($i = 1...I$ and $t = 1...T_i$), α_i is the random location-specific effects assumed to be independently and identically distributed (IID) at the location level, and $\sigma_i = \ln(\alpha_i)$. $\boldsymbol{X}_{it}$ is the vector of covariates, whereas $\boldsymbol{\beta}$ is a vector of estimate coefficients. We

denote the total number of observations as N, which equals to $\sum_{i=1}^{I} T_i$. For simplicity, we use same length and same periods for observations on all sites, for instance annual crash frequency. And hence, all T_i are equal, simplified as T. The Poisson probability specification then becomes

$$\Pr(y_{it} \mid \boldsymbol{X}_{it}, \sigma_i) = \frac{\exp(-\mu_{it} e^{\sigma_i})(\mu_{it} e^{\sigma_i})^{y_{it}}}{y_{it}!} \tag{3.2}$$

where, y_{it} is observed number of crashes for roadway location i in time period t.

To ensure a positive value of λ_{it}, α_i ($=\exp(\sigma_i)$) is generally assumed a gamma distribution with parameters (θ, θ), so that $E(\alpha_i) = 1$, $Var(\alpha_i) = 1/\theta$. Hence, the joint density for this REP model can be derived as

$$\begin{aligned}&\Pr(y_{i1}, ..., y_{iT} \mid \boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}) \\ &= \frac{(\prod_t \mu_{it}^{y_{it}})\Gamma(\theta + \sum_t y_{it})}{(\prod_t y_{it}!)\Gamma(\theta)(\sum_t y_{it})!(\sum_t \mu_{it})^{\sum_t y_{it}}} (\theta/(\theta + \sum_t y_{it}))^{\theta} \left(\frac{\sum_t y_{it}}{\theta + \sum_t y_{it}} \right)^{\sum_t y_{it}}\end{aligned} \tag{3.3}$$

with $E(y_{it}) = \mu_{it}$ and $Var(y_{it}) = \mu_{it}\{1 + (1/\theta)\mu_{it}\}$.

Theoretically, while keeping the same mean as ordinary Poisson regression model, this REP model can explain the over-dispersion caused by within-location covariance. However, over-dispersion may also be the results of inappropriate model specification. Excess zeros is a common source of the potential misspecification in CFPM.

3.3.2 Zero-inflated Poisson Regression Model

In case of over-representation of zero crash observations, ZIP model (Lambert, 1992) may be employed to better fit the data. The basic assumption is that the population consists of two possible states: zero crash state with probability p_{it} and non-zero crash state with probability $(1-p_{it})$. The former consists of those traffic entities that always have zero crash while the latter may be assumed to follow some distribution such as Poisson. In this dual-state system, it is difficult to judge whether an entity with zero crash for a year is in the first or second state. Therefore, the overall probability of zero counts is a combination of the probabilities of zeros from each state, weighted by the probability of being in that state. Hence, the probabilities of zero (0) can be expressed as

$$\Pr[y_{it} = 0 \mid \boldsymbol{X}_{it}] = p_{it} + (1-p_{it})R_{it}(0) \tag{3.4}$$

where $R_{it}(0)$ is a Poisson probability with zero crash (i.e., $y_{it} = 0$) that occur by chance in the second state. On the other hand, the probability of positive counts is given by

$$\Pr[y_{it} > 0 \mid \boldsymbol{X}_{it}] = (1-p_{it})R_{it}(y_{it}) \tag{3.5}$$

where $R_{it}(y_{it})$ is the Poisson probability with positive counts ($y_{it} > 0$). Hence, the ZIP model can be expressed as

$$\Pr(y_{it} \mid \boldsymbol{X}_{it}) = \begin{cases} p_{it} + (1 - p_{it})\exp(-y_{it}), & \mathrm{y}_{it} = 0 \\ (1 - p_{it})\dfrac{\exp(-\mu_{it})\mu_{it}^{\,y_{it}}}{y_{it}!}, & \mathrm{y}_{it} > 0 \end{cases} \tag{3.6}$$

The Equation (3.6) can be further simplified as

$$\Pr(y_{it} \mid \boldsymbol{X}_{it}) = l_{it} p_{it} + (1 - p_{it})\left(\frac{\exp(\mu_{it})\mu_{it}^{\,y_{it}}}{y_{it}!}\right) \tag{3.7}$$

where, l_{it} is an indicator variable in which $l_{it} = 1$ when $y_{it} = 0$ and $l_{it} = 0$, otherwise.

Lambert (1992) has proposed that p_{it} and the mean μ_{it} in the non-zero crash state be formulated as a logit and log-linear relationship respectively with their covariates,

$$\mathrm{logit}(p_{it}) = \ln\left(\frac{p_{it}}{1 - p_{it}}\right) = \mathbf{A}_{it}\boldsymbol{\theta} \tag{3.8}$$

$$\ln(\mu_{it}) = \boldsymbol{X}_{it}\boldsymbol{\beta} \tag{3.9}$$

where $\mathbf{A}_{it}$ and $\boldsymbol{X}_{it}$ are the covariate with $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as their coefficients vectors.

Depending on the specific analytical strategy, the covariates of $\boldsymbol{X}_{it}$ and $\mathbf{A}_{it}$ may or may not be same. In the case of similar covariates (i.e. $\boldsymbol{X}_{it} = \mathbf{A}_{it}$) affecting both p_{it} and μ_{it}, the number of parameters can be reduced by treating p_{it} as a function of μ_{it}.

Hence, a natural parameterization can be further proposed as follows,

$$\text{logit}(p_{it}) = -\tau \boldsymbol{X}_{it}\boldsymbol{\beta} \tag{3.10}$$

$$\ln(\mu_{it}) = \boldsymbol{X}_{it}\boldsymbol{\beta} \tag{3.11}$$

where τ is an unknown, real-value shape parameter. In this ZIP(τ) model, p_{it} of zero crash state is a simple multiplicative function of variables that explain the non-zero crash counts.

The mean and variance of the ZIP model are

$$E(y_{it} \mid \boldsymbol{X}_{it}, \mathbf{A}_{it}) = \mu_{it}(1 - p_{it}) \tag{3.12}$$

$$Var(y_{it} \mid \boldsymbol{X}_{it}, \mathbf{A}_{it}) = \mu_{it}(1 - p_{it})(1 + \mu_{it} p_{it}) \tag{3.13}$$

Obviously, if $p_{it} = 0$, the ZIP specification results in the standard Poisson but otherwise, the variance exceeds the mean, which is a possible source of over-dispersion.

To justify the appropriateness of zero-inflated count model over standard count model, Vuong statistics proposed by Vuong (1989) can be used. In this test, two models are considered: first $\hat{\Pr}_1(y_{it} \mid \mu_{it})$ is the predicted probability of observing n_{it} based on the zero-inflated count data model; second $\hat{\Pr}_2(y_{it} \mid \mu_{it})$ is the predicted probability for the

standard Poisson regression model. By examining the mean ($\overline{m}$) and standard deviation (S_m) of statistic

$$m_i = \ln\left[\frac{\hat{\Pr}_1(y_{it} \mid \mu_{it})}{\hat{\Pr}_2(y_{it} \mid \mu_{it})}\right] \tag{3.14}$$

Voung statistic is defined as

$$V = \frac{\overline{m}\sqrt{N}}{S_m} \tag{3.15}$$

which asymptotically follows a standard normal distribution. If $V > 1.96$ it favors the zero-inflated count model while $V < -1.96$ it favors the parent Poisson regression model but otherwise neither model is preferred.

3.3.3 Zero-inflated Poisson Model with Location-Specific Random Effects

Although ZIP model is capable of accounting for excess zeros by specifying the dual-state system, it is based on the assumption of independence among the observed samples. This assumption is possibly violated in repeated measures design such as crash count at some specific sites. While the crash occurrences may be independent between traffic sites, there almost certainly is correlation among repeated observations at the same sites. Hence, it is interesting and maybe sometimes necessary to consider the location-specific random effects into ZIP model. If these random effects truly exist, Equation (3.8) and Equation (3.9) may lead to erroneous estimations of factor effects.

In particular, location-specific random effects can be considered into ZIP model for both probability of being zero-crash state and count likelihood in non-zero-crash state. Hence, these two equations may be rewritten as follow.

$$\text{logit}(p_{it}) = \ln\left(\frac{p_{it}}{1-p_{it}}\right) = \mathbf{A}_{it}\boldsymbol{\theta} + \psi_i \tag{3.16}$$

$$\ln(\lambda_{it}) = \ln(\mu_{it}\alpha_i) = \boldsymbol{X}_{it}\boldsymbol{\beta} + \sigma_i \tag{3.17}$$

where ψ_i and σ_i are the location-specific random effects for the two states with independent normal distributed, i.e. $\psi_i \sim N(0, \varphi_\psi^2)$, and $\sigma_i \sim N(0, \varphi_\sigma^2)$. Due to some unobserved crash-inducing factors, it is reasonable to assume a correlation between different observations within specific site.

The modified Poisson parameter λ_{it} in Equation (3.17) is also a random variable rather than a deterministic function of $\boldsymbol{X}_{it}$ like μ_{it} in Equation (3.9). Correlation between λ_{it} and $\lambda_{it'}$ $(t \neq t')$ arising for different time period t in a particular location i will be accounted for by α_i while λ_{it} and $\lambda_{i't}$ $(i \neq i')$ for different locations will be assumed to be independent as α_i is assumed independent.

In case of similar covariates, the corresponding REZIP(τ) will then be

$$\text{logit}(p_{it}) = -\tau\boldsymbol{X}_{it}\boldsymbol{\beta} + \psi_i \tag{3.18}$$

$$\ln(\lambda_{it}) = \boldsymbol{X}_{it}\boldsymbol{\beta} + \sigma_i \tag{3.19}$$

3.4 BAYESIAN INFERENCE

3.4.1 Choice of Model Inference Algorithm

Algorithms of MLE inference for generalized linear models with random effects have been successfully built up for many years (e.g. Hinde, 1982). Currently, such hierarchical models can be fitted from a frequentist perspective with specialized computer software such as "MLwinN (Rasbash et al., 2000) and "HLM" (Raudenbush et al., 2001). The model calibration programs for REP and ZIP models are also available in some prevailed statistics software, such as STATA (STATA, 2005). Moreover, Hall (2000) proposed an EM algorithm to maximize the likelihood function for ZIP model with random effects, in which both the state of the process (zero state versus Poisson state) and the random effects were regarded as missing data.

On the other hand, with the recent development of computing capacity and Bayesian analysis techniques, some researchers have been working on calculating the models in a Bayesian framework (Gelman et al., 2003; Congdon, 2003). Bayesian inference (BI) is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations. Instead of giving "maximum likelihood" estimates for the studied unknowns totally based on the sample data in MLE inference, the essential characteristic of Bayesian methods is its explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis. Specifically, the ultimate aim of Bayesian data analysis is to obtain the marginal posterior distribution of all unknowns, and then integrate this distribution over the

unknowns that are not of immediate interest to obtain the desired marginal distribution. Or equivalently, using simulation, we draw samples from the joint posterior distribution and then look at the parameters of interest and ignore the values of the other unknowns.

The general procedure of Bayesian inference is summarized below:

1) Set up the likelihood part of the model, $p(y \mid \mu)$, where μ is the model parameters and y is the observable response data. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

2) Write the prior beliefs about the truth of the parameter value, $p(\mu)$, which could be based on various sources of information. If prior information is not so well formulated, temporarily, we set $p(\mu) \propto$ constant , with the understanding that the prior density can be altered to include additional information or structure.

3) Setting up a full probability model: $p(y, \mu) = p(y \mid \mu) \times p(\mu)$, which is a joint probability distribution for all observable and unobservable quantities in the problem.

4) Conditioning on observed data: calculating and interpreting the appropriate posterior distribution $p(\mu \mid y) \propto p(y \mid \mu) p(\mu)$, i.e. the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data. This is the major difficulty in the development and application of Bayesian techniques. Recently, data simulation for Bayesian posterior inference has made

great progress. The modern approach to Bayesian estimation has become closely linked to sampling-based estimation methods. For this problem, we draw simulations $\mu^1,...,\mu^L$, from the posterior distribution. Use the sample draws to compute the posterior density of any functions of μ that may be of interest.

5) Evaluating the fit of the model and the implications of the resulting posterior distribution.

In this study, BI is employed to calibrate the proposed models. This choice of BI over MLE in crash analysis is important for several reasons when multilevel data structure and extra zero observations are present.

Firstly, while in MLE, coefficients of factor effects are taken as fixed, BI appropriately represents the hierarchical data generating processes of crash occurrence by taking the parameters as unknowns with certain distributions (Gelman et al., 2003).

Secondly, BI can accumulate evidence from any information sources regarding crash prediction. In Bayesian models, any engineering experiences or justified previous findings may be considered into the posterior estimate of parameters by specifying the informative prior on those unknowns with preliminary information (Yang, 2003).

Thirdly, in modeling zero-inflated count data, Bayesian estimates perform better over MLE with respect to interval width and coverage probability when the probability of zero-crash state is chosen closer to unity (Ghosh, et al., 2006).

Moreover, since zero-inflated model could have multiple modes, the MLE are not always suitable for making inferences of parameters in this case while the Bayesian expected mean would be a better summary of the posterior than its modes (Angers and Biswas, 2003).

3.4.2 Bayesian Inference Using Gibbs Sampler

The Bayesian analysis of ZIP models have been proposed by several statistical researchers, e.g. Angers and Biswas (2003), Ghosh et al. (2006). Inspired by these works, this study attempts to conduct the BI further for REZIP model. Gibbs sampling method (Gelfand and Smith, 1990), as a Markov chain Monte Carlo (MCMC) algorithm, is employed to generate samples from non-standard joint posterior distribution of parameters. The basic idea behind the Gibbs sampling algorithm is to successively sample from the conditional distribution of each node given all the others in the graph (these are known as full conditional distributions): the Metropolis-within-Gibbs algorithm is appropriate for difficult full conditional distributions and does not necessarily generate a new value at each iteration. Under broad conditions, this process eventually provides samples from the joint posterior distribution of the unknown quantities. Empirical summary statistics, e.g. mean, median or quantiles, can be formed from these samples and used to draw inferences about their true values.

In the regression problem, the dual-state dependent variable y_{it} is represented by latent variables (V, B) using the data augmentation step (Tanner and Wong, 1987) as follows,

$$y_{it} = V_{it}(1 - B_{it}) \tag{3.20}$$

in which $V_{it} \sim Poisson(\lambda_{it})$ and $B_{it} \sim Bernoulli(p_{it})$.

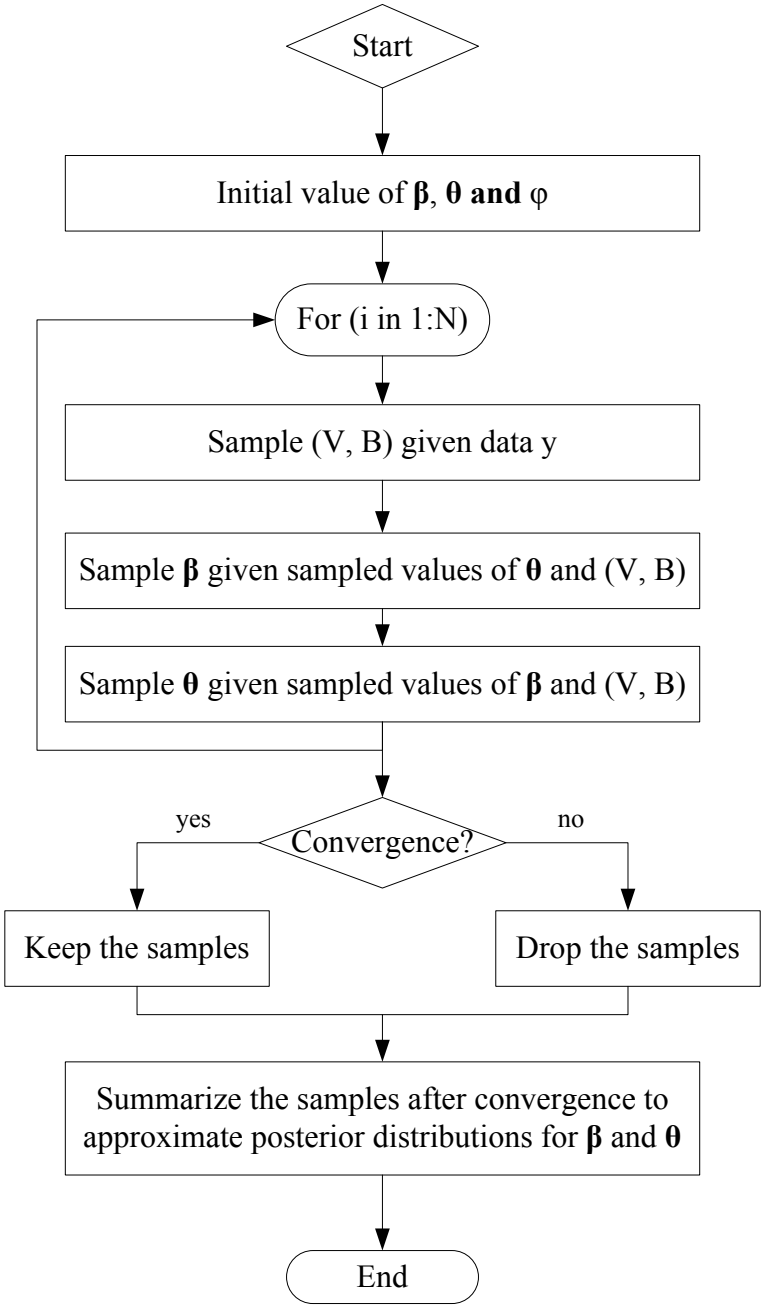

Figure 3.2 Bayesian Inference for ZIP Model Using Gibbs Sampler

Instead of sampling directly from the posterior of (p,λ), samples from the posterior of (p,λ,V,B), are obtained given the dependent variable y. The detailed Gibbs sampling algorithm for zero inflated power series (ZIPS) models is illustrated by Ghosh et al. (2006). The procedure using Gibbs sampler to calibrate the models can be summarized as in Figure 3.2.

In the algorithm, by specifying initial values of model parameters, data augmentation step is implemented to sample (V,B) given current values of (p,λ,n); then using Gibbs sampling method, $\boldsymbol{\beta}$ or $\boldsymbol{\theta}$ are sampled iteratively given the previously sampled values of (V,B) and $\boldsymbol{\beta}$ or $\boldsymbol{\theta}$; (V,B) are also updated with the current sampled values of $(\boldsymbol{\beta},\boldsymbol{\theta})$, this circle continues until convergence. The distributions of estimates are obtained by summarizing the results of a presupposed number of iterations (N N) after model convergence. The magnitude of N depends on the model convergence speed and the complexity of model structure.

When informative prior distributions are available, prior variance-covariance matrix of $\boldsymbol{\beta}$ and/or $\boldsymbol{\theta}$ may be used with some suitably structured matrix instead of identity matrix. In the absence of strong prior knowledge, uninformative priors can be assumed for model parameters such as $\boldsymbol{\beta},\boldsymbol{\theta} \sim N(\mathbf{0},1000\mathbf{I})$ and $\varphi_{\psi}^{2},\varphi_{\sigma}^{2} \sim Unif(0,100)$. On the other hand, reasonable initial values of parameters for the MCMC simulation chains can be obtained by fitting standard logistic and Poisson regression models.

The above MCMC sampling procedure was implemented using BUGS language (Bayesian Inference Using Gibbs Sampling) in WinBUGS (Spiegelhalter et al., 2003).

In this study, specific programmes for REP, ZIP and REZIP models are innovatively designed. Furthermore, Bayesian Output Analysis (BOA) programme (Smith, 2001) is adapted to obtain the distributional summary for MCMC simulation results and also to provide a variety of convergence diagnostic algorithms.

3.5 CROSS VALIDATION MODEL COMPARISON

The choice of the candidate models depends on the complexity and fitness of the subject crash data. In order to compare the specific predicting abilities of REP, ZIP, and REZIP models, cross-validation assessment (CV) in Bayesian framework is proposed. Compared with the other parametric criterion such as AIC and BIC, CV provides fairly flexible and reliable measures to examine the suitability for different output categories, which is extraordinarily useful for assessing the predicting ability on "zero" in this study. Moreover, Instead of just making a point estimate traditionally, cross validation predictive densities (Vehtari and Lampinen, 2002) in Bayesian framework describe the uncertainty in the estimates by obtaining distribution of the expected utility estimate. This is essential in model assessment by computing the probability of one model having a better expected utility than some other model.

In this study, a k-fold CV, instead of the leave-one-out CV, is employed to save the computational cost to overcome the difficulty of slow sampling speed in MCMC algorithm. In particular, the data set are divided into k roughly equal-sized groups and $s(it)$ denotes the set of data points in the group where the observation at site i in time period t belongs. In the k-fold CV, we evaluate the predictive ability of candidate models using the following steps:

1) Remove one group, i.e. $s(it)$, from the data set;
2) Fit the model with the remaining $(k-1)$ groups;
3) Use the fitted model to predict the removed group;
4) Summarize the prediction error by comparing the actual left-out data;
5) Repeat the entire procedure k times with different groups of data left out in turn;
6) Estimate various expected prescribed utilities to compare predictive ability for candidate models, as further discussed below.

In the Bayesian framework, the k-fold CV predictive densities for observation at site i in time period t are computed by the equation

$$p(\hat{y}_{it} \mid x_{it}, D^{(\backslash s(it))}, M) = \int p(\hat{y}_{it} \mid x_{it}, \theta, D^{(\backslash s(it))}, M) \times p(\theta \mid D^{(\backslash s(it))}, M) d\theta \qquad (3.21)$$

where $\hat{y}_{it}$ is the predictive crash number for observation (x_{it}, y_{it}); $D^{(\backslash s(it))}$ are the data in the remaining $(k-1)$ groups except $s(it)$; θ denotes all the model parameters and hyper-parameters of the prior structures and M is all the prior knowledge in the model specification, including all implicit and explicit prior specifications. This means that we have to fit the full model using data of $(k-1)$ groups for k times to yield the n predictive densities. In MCMC algorithm, we sample from $p(\theta \mid D^{(\backslash s(it))}, M)$ for each group, and this would normally take k times to sample from the full posterior. Thus, as k is greatly less than the total number of observations, the computational savings are considerable.

After obtaining the predictive densities, we would like to estimate how good the candidate models are by estimating how good those predictions (i.e. $\hat{y}_{it}$) are. The goodness of the predictive distribution $p(\hat{y}_{it} \mid x_{it}, D^{(\backslash s(it))}, M)$ can be measured by comparing it to the actual observation y_{it} with the utility,

$$u_{it} = u(y_{it}, x_{it}, D^{(\backslash s(it))}, M) \tag{3.22}$$

The goodness of the whole model can then be summarized by computing some summary quantity of the distribution of u_{it} over all data, for example, the mean

$$\overline{u}_{k-fold-CV} = E_{it}[u(y_{it}, x_{it}, D^{(\backslash s(it))}, M)] \tag{3.23}$$

In this study, two estimate utilities are employed. The first utility is the mean predictive square error (MPSE) given by

$$u_{it} = (E[\hat{y}_{it} \mid x_{it}, D^{(\backslash s(it))}, M] - y_{it})^2 \tag{3.24}$$

Hence, the corresponding model comparison criteria is

$$\overline{u} = \frac{1}{n}\sum_{\forall i,t}(E[\hat{y}_{it} \mid x_{it}, D^{(\backslash s(it))}, M] - y_{it})^2 \tag{3.25}$$

$$\overline{u}_{\alpha=0.95} = \frac{1}{n}\sum_{\forall i,t} Max\{(E_{I(0.025,0.975)}[\hat{y}_{it} \mid x_{it}, D^{(\backslash s(it))}, M] - y_{it})^2\} \tag{3.26}$$

where, $\overline{u}_{\alpha=0.95}$ denotes the mean maximum predictive square error with 95% probability confidence, and $E_{I(0.025,0.975)}$ is the 95% Bayesian credible interval of predictive mean in MCMC simulation. This provides a confidence level for the predictions.

To compare the predictive abilities for specific frequency in observations ($f = 0,1,2,...$), a second measure can be defined as disaggregate predictive probability-based utilities $u(f)$ by estimating the cumulative probability of $\hat{y} \in (f-0.5,\ f+0.5)$,

$$u(f) = \frac{1}{n(f)} \sum_{(\forall y_{it}=f)} \hat{P}(\hat{y}_{it} \in (f-0.5,\ f+0.5) \mid x_{it}, D^{(\backslash s(it))}, M) \tag{3.27}$$

where $n(f)$ is the number of actual observed frequency of "f". In the case of large sample sizes, $u(f)\ \forall f$ provides an overall percentage of model fitness.

3.6 SUMMARY

This chapter is the methodology formulation part of the study on CFPM. Following the development of research strategy, the methodology formulation is presented. To account for the multilevel structure and excess zeros in crash frequency data, REP, ZIP and REZIP models are developed associated with the evaluating tools for over-dispersion and excess zeros. Bayesian inference is chosen and developed for model calibration with a number of philosophical and practical advantages over traditional MLE algorithm for the proposed models. To compare the model fitting and predicting performance, a cross-validation algorithm in Bayesian framework, i.e. cross validation predictive density, is proposed innovatively. Several utility criteria are also developed. Using Singapore data, an illustrative study using the proposed methodology is conducted to develop the CFPM for intersection crashes, which is described in the next chapter.

CHAPTER FOUR

CRASH FREQUENCY PREDICTION MODEL ON SIGNALIZED INTERSECTIONS

4.1 INTRODUCTION

Signalized intersection is a hazardous location type on the road, which accounts for a substantial portion of traffic crashes, and the situation appears to be worsening. For example, in United States, 20% of all crashes and 7% of fatal crashes occur at signalized intersections (Porter and England, 2000). Furthermore, a 19% increase of the fatal crash frequency at traffic signals between 1992 and 1996 was reported while the number of all other fatal crashes only increased by 6% (Retting et al., 1999). In Singapore, as shown in Table 4.1, a total of 18008 intersection crashes were reported during 1998-2005, which represent about 33.7% of all traffic crashes.

To investigate the characteristics of intersection crashes, crash frequencies at signalized intersections in Singapore are examined in relation to various site characteristics. The models proposed in Chapter 3 (i.e. REP, ZIP, and REZIP models) are illustrated and examined. The result shows that REZIP model can significantly improve the predictive performance of CFPM for the subject dataset.

According to the research strategy presented in Figure 3.1, this chapter summarizes the main steps in model development, including data collection, model estimation and model comparison. Based on the model results, significant factors are then identified

and interpreted to understand the crash occurrence and to provide recommendations for countermeasure development at intersections.

Table 4.1 Road Crash Statistics in Singapore (1998-2005)

Year	Total crash	Intersection crash	Percentage
1998	5636	1963	34.83%
1999	6548	2336	35.68%
2000	7228	2595	35.90%
2001	7090	2533	35.73%
2002	6879	2500	36.34%
2003	6446	2087	32.38%
2004	6845	2227	32.53%
2005	6706	1767	26.35%
Total	53378	18008	33.74%

4.2 DATA COLLECTION

4.2.1 Site Selection

In order to develop a mathematical model that correlates crash occurrence at intersection to the intersection characteristics, one need to select intersections that have a wide variety of geometric, traffic and control characteristics. A total of 52 four-legged signalized intersections from the southwestern part of Singapore are selected to illustrate the process of establishing a suitable statistical model. Among which a number of intersections are in residential area and are characterized by low traffic volumes and few, if any, road crashes. Several intersections are in the vicinity of Central Business District (CBD) and are characterized by high traffic volumes and

crashes. There is diversity in geometric, traffic and control characteristics among the chosen intersections that lead to proper approach modeling crash occurrence at intersection. The list of selected intersections is given in Appendix A (Table A.1).

4.2.2 Traffic Crash Data

Traffic crash data from the year 1998 to 2005 is collected from Singapore Traffic Police Department. Each of the crash records in the database contains about over 50 fields that exhibit the driver, pedestrian, vehicle and roadway particulars related to the crash. A sample structure of the crash database is enclosed in Appendix A (Table A.2). In the crash records, crash location in road network is depicted by the grid code that may help to display the spatial crash distribution. Any crash within 100m from the center of the intersection is considered as intersection crash.

One can easily identify the total number of crashes at an intersection from the fields of their connecting street code (i.e. STREETCD1, STREETCD2) and crash IP number from the data file (Table A.2). Each intersection is divided into two separate roads, i.e. major and minor road, which are defined based on approach traffic volume. Crash counts are taken at each road in one-year interval. Thus, an intersection provides two observations per year and 16 observations to the study period. Consequently, a total of 832 observations are provided by the 52 intersections. The number of crashes calculated for each road per year is used as a dependent variable in developing the model. In total, 2644 crashes are identified for the selected sites in which 3.2% were fatal, 8.9% resulted in serious injury and the rest in slight or no injury. The distribution of crash counts is shown in the Figure 4.1.

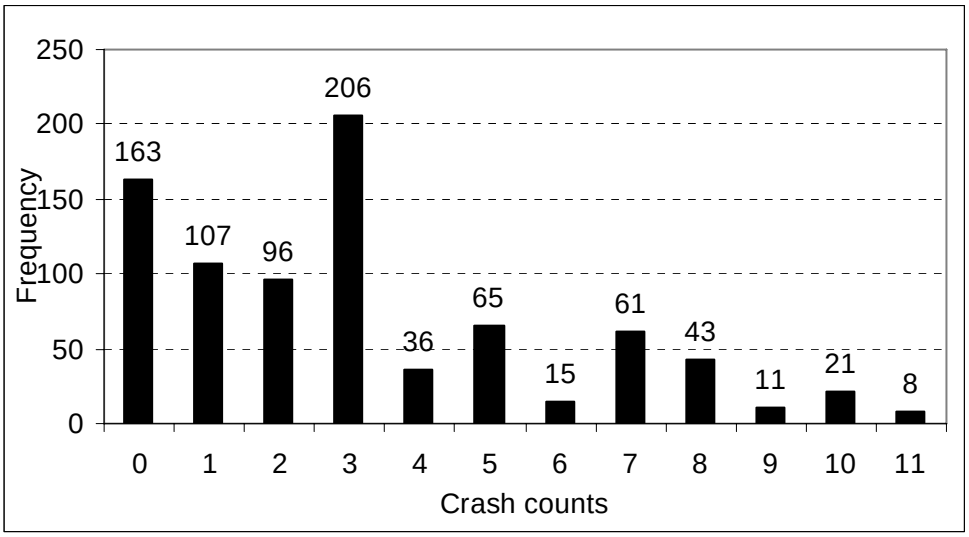

Figure 4.1 Distribution of Crash Counts in Observations

4.2.3 Site Characteristics

In order to examine the relation between crash occurrence and the possible factors, 23 covariates in the subject approach are collected representing traffic conditions, geometric features, and regulatory controls. In addition, 9 covariates are selected from the conflicting approach which may have interactive effects on the crash frequency in subject approach. The descriptive statistics of those variables are presented in Table 4.2.

Table 4.2 Covariates Used in the CFPM

Covariates of subject approach	Mean	S.D.	Min	Max
Number of lanes	5.79	1.90	2	10
Approach width (m)	20.84	6.84	7.20	36
Sight distance (m)	304.50	127.46	51.50	400
Curvature on approach road	0.39	0.49	0	1
Distance between cross walk and the curb(m)	0.48	1.06	0	5.6
Uncontrolled left-turn lane	0.78	0.42	0	1
Exclusively right turn lane	0.60	0.49	0	1
Presence of RLC	0.22	0.42	0	1
Presence of median	0.79	0.41	0	1
Median width greater than 2 m	0.25	0.43	0	1
Number of bus bays	1.71	1.26	0	4
Number of bus stops	2.62	1.31	0	4
Average distance of upstream and downstream bus stops from intersection	280.79	250.70	68.50	1000
Presence of pedestrians refuge	0.13	0.33	0	1
Total approach volume (ADT)	21.21	12.66	1.42	53.85
Approach right-turn volume (ADT)	7.68	5.07	0.51	28.78
Cycle duration	117.69	12.81	100	150
Number of phases per cycle	3.50	0.67	2	5
Percent of green time	0.39	0.13	0.2	0.8
Red length in pedestrian crossing (sec)	75.50	15.38	40	118.50
Speed limit	52.02	6.56	40	80
Signal control type	0.54	0.50	0	1
Covariates of conflicting approach				
Approach width (m)	20.77	6.80	7.20	36
Sight distance (m)	304.50	127.46	51.50	400
Curvature on approach road	0.39	0.49	0	1
Uncontrolled left-turn lane	0.78	0.42	0	1
Exclusively right turn lane	0.59	0.49	0	1
Presence of RLC	0.21	0.41	0	1
Presence of median	0.79	0.41	0	1
Total approach volume (ADT)	21.13	12.61	1.42	53.85
Approach right-turn volume (ADT)	7.61	5.06	0.51	28.78

In the data collection, some geometric elements are measured from the design layout of the intersection, including *Number of lanes, Presence of uncontrolled left-turn lane, Exclusive right-turn lane, Presence of median* at approach road near intersection, and. Most of the geometric data needed are collected from the site survey. These variables include *Approach width, Sight distance* to intersection, the existence of *Curvature on approach road, Distance between cross walk and the curb, Presence of red-light camera (RLC), Median width greater than 2m, Number of bus bays, Number of bus stops, Average distance of upstream and downstream bus stops from intersection, Presence of pedestrians refuge*.

Traffic characteristics at intersection include traffic demand pattern and traffic regulatory control variables. Exposure to crashes at intersection is likely to be dependent on traffic demand pattern, i.e. traffic volumes. Two types of traffic volumes considered in this study, i.e. *Total approach volume* and *Approach right-turn volume* at major or minor road at intersection. Average daily traffic (ADT) of total volumes and right turn volumes are collected from the loop detectors at the sites maintained by the Land Transport Authority (LTA) for 52 intersections.

Since traffic regulation has significant effects on traffic volumes at intersection, traffic regulation may affect crash occurrence at intersections significantly. Traffic regulatory control data, such as *Cycle duration, Number of phases per cycle, Percent of green time, Red duration in pedestrian crossing, Road speed limit* are included in this study. Furthermore, two types of signal control (*Signal control type*) are considered, e.g., adaptive signal control and pre-timed signal control. In the antecedent, all signalized intersections in Singapore were operated under the pre-timed signal control. In recent

years, most of signal controls are converted to adaptive type. The date of conversion is recorded in Traffic Computer System (TCS) maintained by LTA. The signal-timing plan and the number of phases per cycle are also collected from TCS record. The approach speed limit is collected from the accident data file.

In addition to the site characteristics for the subject approach, it is also reasonable to expect significant effects of the conflicting approach characteristics on the crash occurrence. In this study, a total of 9 covariates from the conflicting approach are included in the models, which are also shown in the Table 4.2.

4.3 MODEL CALIBRATION AND COMPARISON

Figure 4.1 shows the crash count distribution of the 832 observations, in which 163 involved no crashes and with a mode of 3 crashes per year per site. Hypothesis test on the "equality" constraint of the mean and variance imposed by the Poisson distribution against the alternative that the variance exceeding the mean, indicates that the over-dispersion parameter is significantly greater than zero ($t = 9.13,\ p = 0.001$). Besides the potential within-site correlation caused by repeated data collection measures, over-dispersion may also be led to by zero-inflation. The test with Vuong statistics ($V = 3.20,\ p < 0.001$) clearly shows that zero-inflated count model is favored over the parent Poisson model.

Since we use the same covariate set for both judgment of zero-crash state (p_{it}) and the parameter for non-zero-crash state (μ_{it}), the natural parameterization on variables in the two states, i.e. ZIP(τ) and REZIP(τ), are utilized as justified in section 3.3. To

evaluate quantitatively the predictive abilities of the candidate models, a 4-fold CV is implemented. In particular, the dataset is evenly divided into four groups with 208 samples in each group. Iteratively, the parameter estimates using data in any three groups are employed to estimate predictive distributions for the observations in the remaining group. Note that the prediction process is done at the same time with the model calibration by treating the data in the test group as missing with the three candidate models, so that a total of 12 models are investigated.

For each model calibration and prediction, three chains of 100,000 iterations are set up in WinBUGS based on the convergence speed and the magnitude of the dataset. After ensuring the convergence, first 20,000 samples are discarded as adaptation and burn-in, and only every tenth samples of the rest are retained for estimation to reduce autocorrelation, leaving a total of 12000 posterior samples.

Table 4.3 Cross-Validation Model Comparison

Utility	REP	ZIP	REZIP
$\bar{u}$	2.31	1.97	1.89
$\bar{u}_{\alpha=0.95}$	6.82	4.54	3.06
$u(f)\ \forall f$	0.29	0.38	0.44

The model comparison results of criteria defined in section 3.5 are shown in Table 4.3. Judged by the criteria of MPSE ($\bar{u}$), models accounting for excess zeros have been demonstrated to have a significant improvement in predictive abilities ($\bar{u}_{ZIP}$=1.97, $\bar{u}_{REZIP}$=1.89). The consideration of location-specific random effects in the ZIP model, resulting in REZIP model, yields the smallest predictive square errors. And REZIP

model also has the smallest credible interval width around the observations ($\bar{u}_{REZIP(\alpha=0.95)}$=3.06). Furthermore, we calculate the probability-based predictive utility $u(f)$ for the whole dataset. The result implies that, compared to REP model, REZIP model can increase the predictive accuracy as the overall percentage of model fitness by about 15%, i.e. from 29% to 44%. This explanatory and predictive power is considered acceptable for our purpose.

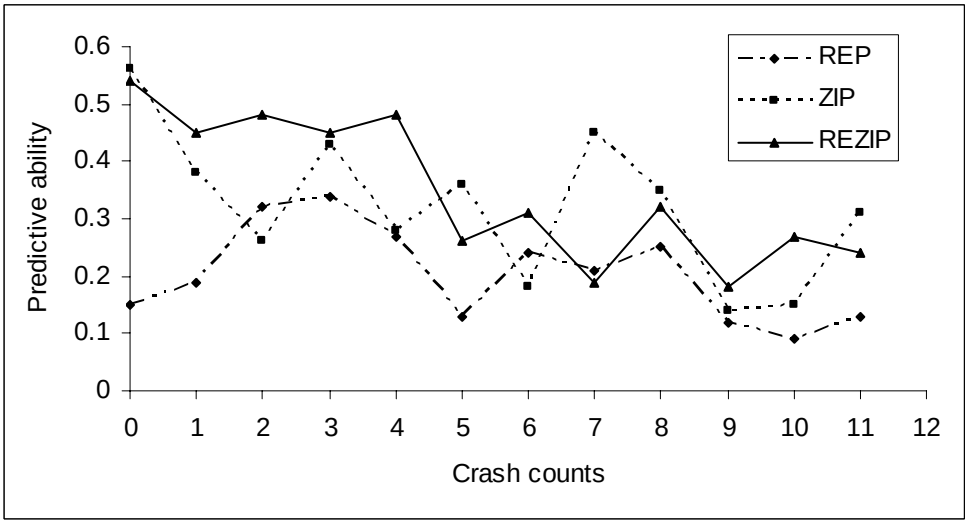

Figure 4.2 Model Comparison of Predictive Abilities using Cross-Validation

In particular, as shown in Figure 4.2, although the predictive abilities vary from the low to high crash frequencies, accounting for excess zeros in ZIP as well as REZIP models as a whole performs better in terms of predictive abilities for future observations. And this is the most apparent for prediction on "zero", where ZIP and REZIP show the significant improvement to predict the zero crash occurrence ($u(0)_{REP}$ = 0.15, $u(0)_{ZIP}$ = 0.56, $u(0)_{REZIP}$ =0.54). This is not surprising as the zero crash state is specifically modeled. One the other hand, as for the differences between ZIP

and REZIP, the ZIP predicts zero a little bit better possibly because REZIP provides more flexible model structure for the whole data rather than only the zero observations. This may also explain the outperforming of REZIP over ZIP in the low frequency predictions. When considering the frequency of crash counts in observations where most are below five, we can conclude that the REZIP will significant improve the predictive abilities compared to REP and ordinary ZIP.

4.4 PARAMETER ESTIMATES AND SIGNIFICANT VARIABLES

To understand how the different models affect the assessment of the various risk factors, we run each fitted model to obtain the parameter estimates using all 832 samples. As expected, some variations occur in the means of estimates as well as in the credible intervals. In Table 4.4, we list the estimates for those variables which are significant statistically in at least one of three candidate models, i.e. REP, ZIP, and REZIP. Same signs for most effect estimates are found in all three models, which imply that we can be relatively confident about the qualificative effect of risk factors (i.e. negative or positive) on crash frequencies. However, with respect to quantitative effects, it is not surprising to find a fairly large difference between REP and the other two zero-inflated models because of the entirely different model specifications. Moreover, regarding ZIP and REZIP models, two major differences are identified. Firstly, the credible intervals differ to some extent although parameter estimates are approximately similar. In particular, three non significant factors in the ZIP model appear to be significant in the REZIP model (i.e. *Time trend, Number of lanes, Distance of bus stops from intersection*) while another three significant factors in the ZIP model proved to be insignificant in the REZIP model (i.e. *Present of RLC, Cycle*

duration, *Uncontrolled left-turn lane in conflicting approach*). Secondly, there is a difference in the shape parameters (τ) ($\tau_{ZIP} = 1.50, \tau_{REZIP} = 2.08$). The larger shape parameter in the REZIP model indicates a steeper average trend towards zero-crash state with unit change in the risk factors.

Since the crash data are derived from historical records, and not from designed experiments, it seems difficult to assess the resulting differences by the differences themselves. Hence, the model assessment measures on predictive ability as illustrated in section 3.5 are especially useful for this kind of 'Happenstance Data' (Box et al., 1978). Moreover, the validity and practicality of the individual factors can also be examined based on engineering and intuitive judgment. In Table 4.4, Incidence Rate Ratios (IRR), i.e. $\exp(\beta)$ are calculated for the REZIP model results to facilitate interpretation of the variables. Apparently, if the IRR of a given variable is much less than 1.0, then an increase in value of the variable is associated with a significant improvement in safety and vice versa. Since the REZIP model have proved a relatively better fit for the data and in the predictive abilities, several interpretations of the parameter estimates may be made, as follows.

Table 4.4 Posterior Summary of Parameter Estimates

	REP	ZIP	REZIP			
Covariates	mean	mean	mean	95% BCI		IRR
Time trend	-0.01	-	-0.02	-0.03	0.00	0.99
Number of lanes	-	-	0.08	0.03	0.16	1.08
Sight distance	0.19	0.08	0.11	0.01	0.22	1.12
Presence of RLC	-	0.07	-	-	-	-
Presence of median	-	0.08	0.05	0.03	0.11	1.06
Distance of bus stops from intersection	0.17	-	-0.09	-0.16	-0.03	0.91
Total approach volume	0.11	0.11	0.07	0.01	0.11	1.07
Cycle duration	-0.18	-0.11	-	-	-	-
Number of phases per cycle	-	0.06	0.05	0.00	0.12	1.05
Red length in pedestrian crossing (sec)	0.21	0.06	0.10	0.02	0.18	1.10
Uncontrolled left-turn lane in conflicting approach	-	0.09	-	-	-	-
Conflicting approach total volume	0.25	0.19	0.15	0.03	0.28	1.17
Tau (shape parameter)		1.50	2.08	1.24	3.40	

As a whole, a small but significant decreasing time trend of crash occurrence is identified in the model (IRR 0.985, 95%BIC (-0.033, -0.003)). In traffic variables, both *Total approach volume* (IRR 1.068, 95%BIC (0.005, 0.112)) and *Conflicting approach volume* (IRR 1.165, 95%BIC (0.027, 0.278)) are found to be significantly associated with crash frequency. It is not surprising since exposure to crash is likely to depend on traffic volume. Among the geometric factors, both *Number of lanes* (IRR 1.078, 95%BIC (0.028, 0.159)), and *Presence of median* (IRR 1.055, 95%BIC (0.026, 0.113)) have negative effects on the intersection safety. This may be explained that a higher crash frequency is associated with larger intersections, with wider medians and more

traffic lanes. In such instances, not only are there more conflict points, the less defined space for vehicle turning and maneuver would have also contributed to more crashes. *Sight distance* (IRR 1.116, 95%BIC (0.005, 0.217)) is surprisingly identified as another negative geometric factor. The greater freedom of maneuver and potential higher speed with long sight distance may be the causes resulting in greater crash frequencies. This may be especially true when considering the complex risk factors regarding to the regulatory controls at signalized intersection. Kulmala, R. (1995) and Chin and Quddus, (2003a) also found the similar results in the studies of four-leg intersections.

Moreover, result also shows that longer distance of the bus stop from the intersection give rise to fewer crash occurrences (IRR 0.912, 95%BIC (-0.162, -0.025)). This is reasonable since the presence of a standing bus close to the intersection will influence the traffic maneuver near the intersection. Finally, two signal control variables are identified: *Number of phases per cycle* (IRR 1.025, 95%BIC (0.002, 0.121)), and *Red duration in pedestrian crossing* (IRR 1.104, 95% BIC (0.018, 0.183)). It is reasonable to expect higher crash risks during phase change periods and increase of the number of phases in fixed time means more potential conflicts. While a long red duration in pedestrian crossing per cycle does not give rise to higher crash risk, longer designed duration implies more pedestrian traffic crashes and is therefore a surrogate measure of pedestrian exposure.

4.5 Summary

The study of CFPM, as presented in Chapter 3 and Chapter 4, showed that when analyzing the crash frequency data with multilevel structure and excess zeros, REZIP model could be used as an alternative to the ordinary REP model or ZIP model. A methodological framework using Bayesian analysis was proposed for CFPM. This framework was shown to provide a reliable measure to fit various flexible models. A cross validation comparison method was used to evaluate the suitability of the models. The assessment measures proved to be useful and reliable to examine the predictive performance of the whole model as well as the realization of individual observations in the data, for instance, “zero” occurrence in crash data.

Using intersection data in Singapore, the illustrative results indicated that REZIP can significantly perform better in terms of predictive abilities over the other candidate models. The differences in parameter estimates in the three models (REP, ZIP, REZIP models) may not be sufficient to justify the suitability of any model. However, engineering and intuitive judgment based on the results estimated lends support to the selection of appropriate models. Furthermore, the differences between model results also imply that careful model development and assessment should be conducted since different specifications could result in quite different effect estimates as well as in their credible intervals.

It should be noted that all these considerations and treatments are aimed at accounting for the possible sources of over-dispersion in crash data. In particular, while random effect models take the physical data collection scheme into consideration, zero-inflated

models assume a dual-state data-generating process to explain the excess zeros. Hence, although the model selection depends on specific data, model specification considering both zero-inflated and random effects proposed in this study can be recognized as a theoretical improvement to better account for the possible sources of over-dispersion.

CHAPTER FIVE

BAYESIAN HIERARCHICAL BINOMIAL LOGISTIC MODEL IN CRASH SEVERITY PREDICTION

5.1 INTRODUCTION

Crash frequency and severity are two major concerns in understanding the relationship of crash occurrence and various risk factors. In Chapter 3 and Chapter 4, we developed statistical techniques to model the multilevel data and excess zeros in crash frequency prediction (CFPM). In addition to the crash frequency, crash severity is another important major symptom of traffic system safety. Before developing and implementing the traffic safety treatments, it would be very useful if a comprehensive understanding of the effects of risk factors on crash severity is available.

As reviewed in Chapter 2, generalized linear regression models (GLM) for discrete response variable, e.g. logit/probit model and ordered model, are commonly used in crash severity prediction models (CSPM). However, most crash severity studies ignored severity correlations between individuals involved in the same cluster, for example, occupants in the same vehicle, drivers in the same crash etc. Models without accounting for these within-cluster correlations will result in biased estimates in the factor effects.

This study proposes a Bayesian hierarchical analysis to examine the crash severity which is capable of appropriately modeling the multilevel data structure. To formulate the methodology, we take the driver-vehicle units involved in same crashes as the subject of study. The research justification and strategy for CSPM development is presented in the following.

5.2 RESEARCH JUSTIFICATION AND STRATEGY

Analysis of crash severity can be conducted in different ways for various purposes. Some studies focused on the crash frequencies at specific traffic sites associated with different severity levels (e.g. fatal, serious, slight) to investigate how geometric, traffic, and environmental factors affect the crash severity. While this kind of studies normally take each crash as the subject unit, analysis can also be undertaken based on the driver-vehicle units involved in crashes to examine individual severity. Compared to the crash-based severity studies, individual severity analysis is promising and may yield a disaggregate understanding about severity levels of different driver-vehicle groups. This is especially useful when the severity levels of driver-vehicle units with different characteristics are desired (Hauer, 2006).

Since the techniques used in most past severity studies assumed independence between different observations, these techniques may not be adequate in modeling multilevel individual severity of driver injury or vehicle damage in the presence of potential correlations between those involved in the same multi-vehicle crashes. Actually, this correlation between samples has already been identified in some earlier studies; for example, Evans (1992, 1993) found that in a multiple vehicle crash, the risk of fatality

was dependent on the characteristics of the other vehicles. Hence, the models without considering the covariance between individuals in the same crashes, especially when the covariance is significant, will result in inaccurate or biased estimates of factor effects.

As discussed previously, hierarchical modeling is a statistical technique that allows multilevel data structures to be easily specified and estimated (see Snijders and Bosker, 2000; Goldstein, 2003). Although the basic theories of hierarchical models have been developed and discussed for many years, it is only recently that many practical limitations on the use of hierarchical analysis have been overcome. A good number of applications of this modeling technique have been found in sociological research disciplines. In traffic safety research, Jones and Jorgenson (2003) presented a good exploration and discussion on the potential applications of the hierarchical models. Since then, the hierarchical modeling technique has been gaining an increasing amount of attention in accounting for the hierarchical data structure in road crash frequency and severity studies. For example, Jones and Jorgensen (2003) and Lenguerrand and Laumon (2006) developed hierarchical models to identify factors affecting crash severity, while Kim et al. (2007) employed the hierarchical crash prediction models for different crash types at rural intersections.

In the investigation of individual severity in crashes at signalized intersections in Singapore, a within-crash correlation was preliminarily identified, which will be shown in detail in Chapter 6. Motivated by this correlation and inspired by the existing studies with hierarchical models, we propose the use of a hierarchical binomial logistic (HBL) model to examine the significant risk factors related to severity of driver injury

and vehicle damage in traffic crashes. In particular, crash is considered as cluster and there are a number of sub-clusters per cluster, i.e. driver-vehicle units involved in a crash. A full Bayesian method using Markov chain Monte Carlo (MCMC) algorithm is employed for model calibration to explicitly model the two-level data structure, i.e. crash-level and individual-level. Using the Intra-class Correlation Coefficient (ICC) and Deviance Information Criterion (DIC) in model assessment and comparison, the use of random effects on crash level in the model is further validated to be effective in this study in accounting for the within-crash correlations.

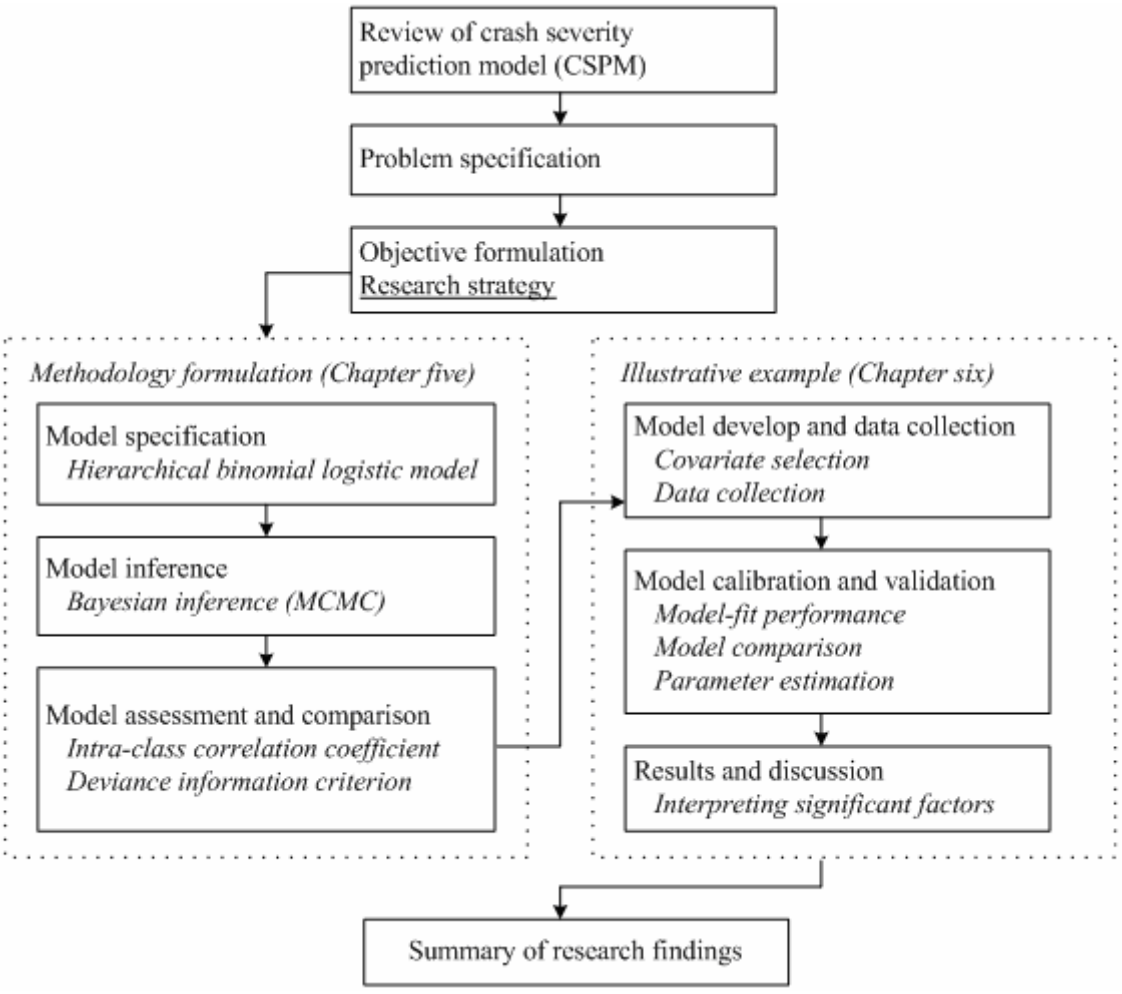

Figure 5.1 Research Strategy for CSPM Development

The specific research strategy for CFPM development is illustrated in Figure 5.1. In the rest of this chapter, the methodology is formulated, consisting of model development, inference, assessment and comparison. In Chapter 6, the illustrative study using Singapore crash data is presented. Specifically, data collection and model calibration are summarized to illustrate the proposed methodology and to understand the significant risk factors on individual severity. Summary of this study are presented finally.

5.3 HIERARCHICAL BINOMIAL LOGISTIC MODEL

In the presence of within-crash correlation of individual severity, models without appropriately considering the hierarchical data structure might yield inaccurate or biased parameter estimations. To account for this within-crash correlation, a HBL model with two-level specification is developed to estimate the effects of the selected covariates on severity level. Specifically, in the individual-level model (level 1), the response variable Y for the ith driver-vehicle unit in jth crash only takes one of two values: Y_{ij} = 1 in case of high severity, e.g. fatal or severe injury, while Y_{ij} = 0 in case of low severity, e.g. slight or no injury. The probability of Y_{ij} = 1 is denoted by $\pi_{ij} = \Pr(Y_{ij} = 1)$, which follows a binomial distribution; hence

$$\text{Logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} X_{pij} \qquad (5.1)$$

where X_{pij} is the pth covariate in the individual-level for ith driver-vehicle unit in jth crash, while β_{0j} and β_{pj} are the intercept and the regression coefficients. In the context

of the hierarchical model, the within-crash correlation is specified in the crash-level model (level 2) as:

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} Z_{qj} + u_{0j} \tag{5.2}$$

$$\beta_{pj} = \gamma_{p0} + \sum_{q=1}^{Q} \gamma_{pq} Z_{qj} + u_{pj} \tag{5.3}$$

In Equation (5.2) and Equation (5.3), both intercept β_{0j} and regression coefficients β_{pj} in Equation (5.1) vary with the different crashes. Specifically, two components are combined to decide the coefficient values. First, linear relationships are assumed for them with the crash-level covariates Z_{qj}, which is reasonable since the various crash features (e.g. street lighting, road surface condition) may result in different severity results. Second, besides the fixed parts which depend on the crash-level covariates Z_{qj}, random effects are also included to permit the potential random variations across the crashes (u_{0j} and u_{pj}). These between-crash random effects vary across the different crashes only but are constant for all the driver-vehicle units within a same crash. This specification enables the model to account for the within-crash correlations (Jones and Jorgensen 2003, Kim et al. 2007). Practically, the random effects are used to represent some unobservable variations between different crashes, which is the major difference between ordinary binomial logistic model (OBL) and HBL.

The full model with Equation (5.1), Equation (5.2) and Equation (5.3) is academically named as random slope model (Snijders and Bosker, 2000). When the random effects are assumed only on the intercept, a simplified form can be obtained by dropping the

crash-level covariate component $\sum_{q=1}^{Q}\gamma_{pq}Z_{qj}$ and the random part u_{pj}, which is referred to as random intercept model. The Equation (5.3) is thus modified to be:

$$\beta_{pj} = \gamma_{p0} \tag{5.4}$$

In this study, to avoid excess complexity as the large set of covariates used, only the random intercept model is investigated. Hence, the combined model is yielded by substituting Equation (5.2) and Equation (5.4) with Equation (5.1) and is represented as follows:

$$\text{Logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_{00} + \sum_{p=1}^{P}\gamma_{p0}X_{pij} + \sum_{q=1}^{Q}\gamma_{0q}Z_{qj} + u_{0j} \tag{5.5}$$

The random effects u_{0j} are generally assumed as a normal distribution with mean zero and variance τ_0^2, as suggested by Snijders and Bosker (2000). The variance of outcome (Y_{ij}) therefore consists of two components: the variance of u_{0j} (τ_0^2) which captures the between-crash variability (level 2), and the variance associated with logistic distribution which captures the within-crash variability (level 1).

In interpreting the fixed effect part of coefficient estimation, a similar way can be followed as with the OBL, the exponential of effect coefficients, i.e. $\exp(\gamma)$, can be calculated to obtain Odds Ratio (O.R.) estimates in HBL model. This provides a basic interpretation for the magnitude of γ: if O.R. is less than 1.0, a unit increase in the

variable X_{pij} or Z_{qj} will reduce the odds of being severe by a multiplicative effect of $\exp(\gamma)$ and vice versa. For the categorical covariates in the model where dummy variables are applied, $\exp(\gamma_a - \gamma_b)$ represents the odds ratios between these two categorical variables, *a* and *b*. In this case, the parameter or its estimate makes sense only by comparing one category with another.

5.4 BAYESIAN INFERENCE

There are several methods available for model calibration in hierarchical binomial logistic model (see Goldstein, 2003). As discussed in section 3.4, Bayesian analysis has a number of intrinsic advantages for calibrating hierarchical models over classical likelihood-based estimation methods. Several studies have also demonstrated the potentials of Bayesian inference (BI) in philosophical aspect as well in practical aspect in transportation applications (e.g. Washington et al., 2005; Mitra and Washington, 2007). Therefore, this study of CSPM also employs BI to calibrate the proposed two-level model (Gelman et al., 2003). A summarized description as well as the general procedure of BI can be found in section 3.4.

Specifically, in the absence of strong prior information for the model unknowns of the proposed HBL model, uninformative priors are assumed for all regression coefficients (γ_{00}, γ_{p0} and γ_{0q}) with normal distributions (0, 1000), and the variance τ_0^2 of the normal distributed random effects μ_{0j} with inverse gamma distribution (0.001, 0.001). The model was also programmed via the Gibbs sampler (Gilks et al., 1995) using BUGS language, which is implemented using WinBUGS (Spiegelhalter et al., 2003a). The 95% Bayesian Credible Interval (95% BCI) is used to examine the significance of

covariates, which provides probability interpretations with normality assumption on unknowns and confidence interval estimations (Gelman et al., 2003). Specifically, those coefficient estimations are identified as significant, whose 95% BCIs do not cover "0", i.e. the 95% BCIs of O.R. do not cover "1". Besides, engineering and intuitive judgment should be able to confirm the validity and practicality of the sign of each covariate and the rough magnitude of each estimated coefficient.

5.5 MODEL ASSESSMENT USING INTRA-CLASS CORRELATION COEFFICIENT (ICC)

An Intra-class Correlation Coefficient ρ (ICC) is normally defined to examine the proportion of specific crash-level variance (level 2) in overall residual variance (Jones and Jorgensen 2003; Kim et al. 2007). Since the logistic distribution for the individual-level (level 1) residual implies a variance of $\pi^2/3 = 3.29$, this implies that for a two-level logistic random intercept model with an intercept variance of τ_0^2, the ICC for between-crash residual is

$$\rho = \frac{\tau_0^2}{\tau_0^2 + \pi^2/3} \tag{5.6}$$

The ICC is an indicator of the magnitude of the within-crash correlation. A value of ρ close to zero means that there is a very small variation between the different crashes, indicating that OBL model may be adequate for the data. On the other hand, a relative large value of ρ implies a favor for hierarchical model, e.g. HBL model in this study.

5.6 MODEL COMPARISON USING DEVIANCE INFORMATION CRITERION (DIC)

To further ensure the advantage of employing HBL over OBL, an OBL model with the same covariates and dataset can also be estimated to compare with the calibrated HBL model. The OBL model may be given by dropping random effect part u_{0j}, which means ignoring the severity correlations between driver-vehicle units within the same crashes. So the Equation (5.5) changes to:

$$\text{Logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_{00} + \sum_{p=1}^{P}\gamma_{p0}X_{pij} + \sum_{q=1}^{Q}\gamma_{0q}Z_{qj} \tag{5.7}$$

For model comparison, a recently-developed criterion, Deviance Information Criterion (DIC), proposed by Spiegelhalter et al. (2003b), is employed. To introduce the DIC, an evolutional review of the traditional model comparison criteria is necessary.

Within the classical modelling framework, model comparison generally takes place by defining a measure of fit, typically a deviance statistic, and complexity, the number of free parameters (degree of freedom, DF) in the model. The deviance statistic (G^2) is defined as:

$$G^2 = -2(\log L_c - \log L_f) \tag{5.8}$$

in which L_c denote the likelihood of current model, and L_f denote the likelihood

estimated from the full (or saturated) model, or in other word, the maximum attainable likelihood for the data.

Since increasing complexity is accompanied by a better fit, models are compared by trading off these two quantities and, following early work of Akaike (1973), proposals are often formally based on minimizing a measure of expected loss (Akaike Information Criterion, AIC) on a future replicate data set as follows:

$$AIC(b) = -2(\log L_c) + 2b \qquad (5.9)$$

in which b is the number of variables in the model. Using this criterion, the model yielding the minimum AIC may be selected as the best model (Joshua and Garber, 1990).

In the case of large samples, the use of the G^2 statistic as a goodness-of-fit measure may not be a satisfactory procedure for rejecting one model in favor of another (Raftery 1986, 1995). The essence of the argument is that, when the sample size is large, it is much easier to accept (or at least harder to reject) more complex models because the likelihood-ratio test (G^2) is designed to detect any departure between a model and observed data. Adding more terms to a model will always improve the fit, but with large samples it becomes harder to distinguish a “real” improvement in fit from a trivial one.

One solution to this problem is to use the Bayesian information criterion (BIC) statistic in searching for parsimonious models that provide an “adequate” fit to the data. The

BIC index provides an approximation to a $-2\times\log$ transformed *Bayes factor*, which may be viewed as the ratio in likelihood between one model (M_0) and another model (M_1). The basic idea is to compare the relative plausibility of two models rather than to find the absolute deviation of observed data from a particular model. However the statistical methods for calculating the Bayes factor are complicated. Many applied researchers have found the BIC statistic popularized by (Raftery 1986, 1995) to be useful. It is defined as:

$$BIC = G^2 - DF\log n \tag{5.10}$$

This expression shows that BIC penalizes G^2 more, per degree of freedom, for a larger sample than for a smaller sample, at the rate of $\log n$.

A model comparison using the AIC or BIC both requires the specification of the number of parameter in each model, but in complex hierarchical models parameters may outnumber observations and these methods clearly cannot be directly applied (Gelfand and Dey, 1994). The most ambitious attempts to tackle this problem appear in the smoothing and neural network literature (Wahba, 1990). Spiegelhalter et al. (2003b) suggest Bayesian measures of complexity and fit that can be combined to compare models of arbitrary structure. It aims to identify models that best explain the observed data but with the expectation that they are likely to minimize uncertainty about observations generated in the same way. It is defined as:

$$DIC = D(\bar{\theta}) + 2p_D = \overline{D(\theta)} + p_D \tag{5.11}$$

in which $D(\theta)$ is termed as 'Bayesian deviance' in general:

$$D(\theta) = -2\log\{p(y \mid \theta)\} + 2\log\{f(y)\} \tag{5.12}$$

and, more specifically, for members of the exponential family with $E(Y) = \mu(\theta)$ we shall use the saturated deviance $D(\theta)$ obtained by setting $f(y) = p\{y \mid \mu(\theta) = y\}$.

p_D is motivated as a complexity measure for the effective number of parameters in a model, as the difference between the posterior mean of the deviance and the deviance at the posterior estimates of the parameters of interest. It is given as:

$$p_D = \overline{D(\theta)} - D(\bar{\theta}) \tag{5.13}$$

This is the so called "mean deviance minus the deviance of the means'. $D(\bar{\theta})$ is regarded as classical estimate of fit given by the MCMC simulation. And the posterior mean deviance $\overline{D(\theta)}$ can be taken as a Bayesian measure of fit or "adequacy". The DIC is formed by the sum of the classical estimate of fit and twice the effective number of parameters (p_D). Also we can consider DIC as a Bayesian measure of fit or adequacy, penalized by an additional complexity term p_D. Obviously, DIC is intended as a generalization of AIC.

5.7 SUMMARY

This chapter is the part of methodology formulation in the CSPM study. Based on the research justification, we take the individual severity of driver-vehicle units involved in same crashes as the subject of study. Following the development of research strategy, the methodology formulation is presented. A Bayesian HBL model is proposed to account for the within-crash severity correlations of individuals involved in same crashes. ICC is adopted to evaluate the magnitude of random effects. DIC is further introduced to compare the suitability of hierarchical logistic model to the ordinary logistic model. The proposed methodology is illustrated and validated using Singapore crash data, which is presented in Chapter 6.

CHAPTER SIX

SEVERITY OF DRIVER INJURY AND VEHICLE DAMAGE IN TRAFFIC CRASHES AT SIGNALIZED INTERSECTIONS

6.1 INTRODUCTION

To model the within-crash correlation, a Bayesian HBL model is developed in Chapter 5 to investigate the significant factors on the individual severity. According to the research strategy proposed in section 5.2, this chapter presents a study on individual severity of driver injury and vehicle damage at signalized intersection using Singapore crash data to illustrate and validate the proposed methodology. Following a description of data set for analysis, model calibration and validation results are summarized. Based on the parameter estimation, significant factors are identified and discussed. The summary of CSPM study is given finally.

6.2 DATA SET FOR ANALYSIS

For this study, crash data in Singapore from 2003 to 2005 are used. Of the total of 19832 reported crashes in this period, 4095 cases occurring at signalized intersections are extracted and used in the model. In these, 7840 driver-vehicle units are involved, resulting in an average involvement rate of 1.91 individuals per crash.

In the dataset, each observation is associated with a driver-vehicle unit involved in the crashes at intersections. Two categorical severity indicators are of interest, which are

driver injury severity: a) fatal or serious injury, DI(A), b) slight or no injury, DI(B); and vehicle damage severity: a) extensive damage, VD(A), b) slight or no damage, VD(B). To yield a net effect estimate of each potential factor on individual severity, a binary dependent variable is defined by combining the two severity indicators: a) DI(A) or/and VD(A), denoted as IS(A), representing high individual severity b) otherwise is low individual severity denoted as IS(B). A summary of severity statistics is given for years in Table 6.1.

TABLE 6.1 Summary of Crash Severity at Signalized Intersection by Years

Year	DI(A)	DI(B)	% of DI(A)	VD(A)	VD(B)	% of VD(A)	IS(A)	IS(B)	% of IS(A)
2003	39	2622	1.49	491	2170	22.63	508	2153	23.59
2004	37	2885	1.28	398	2524	15.77	412	2510	16.41
2005	36	2221	1.62	173	2084	8.30	192	2065	9.30
Total	112	7728	1.45	1062	6778	15.67	1112	6728	16.53

Note: DI(A): driver with fatal/serious injury; DI(B): driver with slight or no injury
VD(A): vehicle with extensive damage; VD(B): vehicle with slight or no injury
IS(A): DI(A) or/and VD(A); IS(B): otherwise

In addition to severity levels, a record of crash IP number, geometric features, traffic conditions, driver and vehicle characteristics is also reported. There are a total of 25 variables coded for each intersection crash in the dataset. A number of variables like location code, vehicle registration number, nature of vehicle registration etc. are excluded as they were irrelevant to the analytical purpose. A correlation matrix for those remaining variables, which are hypothesized to relate to the severity levels, is checked to avoid multi-collinearity as well as wrong signs or implausible magnitudes in the estimated coefficients. For the highly correlated variables, only the most significant variable is retained in the analysis; for example, weather condition is

excluded because of its high correlation with road surface. Finally, a total of ten covariates in the crash-level are used, i.e. *Day of week, Time of day, Intersection type, Nature of lane, Road surface, Street lighting, Road speed limit, Vehicle movement, Presence of red light camera (RLC),* and *Pedestrian involved*. In addition, to explore how differently the various driver-vehicle characteristics affected the severity levels, five covariates in the individual-level, i.e. driver-vehicle level, were selected, i.e. *Vehicle type, Driver age, Driver gender, Involvement of offending party, Passenger involved*. Unfortunately, several vehicle safety features such as airbags, and anti-lock brakes, are not included in the crash dataset. But although those variables may be important to affect the individual severity, they are not so useful in Singapore since most vehicles are less than 6 years old and are hence equipped with the latest protective features in modern cars. Moreover, the stringent compulsory annual inspection on all vehicles to ensure they are road worthy means that these features are in serviceable conditions.

The definitions of the selected covariates, together with their mean and standard deviation (S.D.), are presented in Table 6.2. For convenience of analysis, all these variables are split as groups of dummy variables based on the engineering experiences or existing findings in previous studies. For example, *Vehicle type* is categorized as three groups of two-wheel vehicle, light vehicle and heavy vehicle, since the vehicle weight had been identified relevant to injury severity (Evans and Frick, 1994).

TABLE 6.2 Covariates used in the CSPM

Covariates	*Description of the variables*	*Mean*	*S.D.*
Day of Week	If crash at weekend =1, otherwise=0	0.164	0.370
Time of Day			
Day time	If crash in 10am – 5pm =1, otherwise = 0	0.289	0.453
Night time	If crash at 8pm – 7am =1, otherwise = 0	0.434	0.496
Peak time	If crash at 7am – 10am or 5pm – 8pm =1, otherwise=0	0.278	0.448
Intersection Type			
X intersection	If crash at X type intersection =1, otherwise =0	0.014	0.115
T/Y intersection	If crash at T/Y type intersection =1, otherwise =0	0.232	0.422
Other types	If crash at other type intersection =1, otherwise =0	0.755	0.430
Nature of Lane			
Single lane	If crash on single lane =1, otherwise =0	0.025	0.155
Left-most lane	If crash on Left-most lane =1, otherwise =0	0.163	0.369
Right-most lane	If crash on right-most lane =1, otherwise =0	0.256	0.437
Centre lane	If crash on centre lane =1, otherwise =0	0.556	0.497
Road Surface	If road surface is dry = 0, otherwise =1	0.129	0.335
Weather Condition	If weather condition is fine = 0, otherwise =1	0.098	0.297
Street Lighting	If street lighting is fine = 0, otherwise =1	0.338	0.473
Road Speed Limit			
40 km/h	If road speed limit is 40km/h =1, otherwise =0	0.005	0.068
50 km/h	If road speed limit is 50km/h =1, otherwise =0	0.891	0.311
60 km/h	If road speed limit is 60km/h =1, otherwise =0	0.072	0.258
70 km/h	If road speed limit is 70km/h =1, otherwise =0	0.032	0.176
Vehicle Movement			
Single vehicle self-skidded	If Single vehicle self-skidded =1, otherwise = 0	0.031	0.172
Single vehicle against stationary object or pedestrian	If Single vehicle against stationary object or pedestrian =1, otherwise = 0	0.029	0.169
Between moving vehicle and stationary vehicle	If between moving vehicle and stationary vehicle =1, otherwise =0	0.882	0.323
Between moving vehicles	If between moving vehicles =1, otherwise = 0	0.053	0.223
Other movements	If other movements =1, otherwise = 0	0.006	0.076
Presence of Red Light Camera	If a red light camera is present =1, otherwise = 0	0.072	0.258
Pedestrian Involved	If passengers involved =1, otherwise = 0	0.051	0.220
Vehicle Type			
Two-wheel vehicle	If vehicle type is motor scooter or motorcycle =1, otherwise = 0	0.304	0.460
Light vehicle	If vehicle type is motorcar, station wagon, goods can, pick-up or minibus =1, otherwise =0	0.572	0.495
Heavy vehicle	If vehicle type is Bus, bendy, lorry, tip truck, trailer, crane or other heavy vehicles =1, otherwise =0	0.124	0.329
Driver Age			
<= 25	If driver age <= 25 = 1, otherwise =0	0.162	0.368
26 – 45	If driver age within 26-45 =1, otherwise =0	0.480	0.500
46 – 65	If driver age within 46-65 =1, otherwise =0	0.326	0.469
> 65	If driver age > 65 =1, otherwise =0	0.033	0.178
Driver Gender	If driver is female =1, otherwise =0	0.104	0.305
Involvement of Offending Party	If driver is likely at=fault =1, otherwise =0	0.627	0.484
Passenger Involved	If with passengers on board =1, otherwise =0	0.170	0.376

6.3 MODEL CALIBRATION AND VALIDATION

A preliminary examination of potential within-crash covariance in the collected data set identified a significant correlation between individuals involved in same multi-vehicle crashes, which represent 83.5% of all crashes at signalized intersections in Singapore. In particular, in a multi-vehicle crash, if the severity of driver-vehicle unit was IS(A), then the others had a probability of 31% also to be in IS(A). On the other hand, if a driver-vehicle unit was in IS(B), then the others had only 12% chance to be in IS(A). This significantly lower ratio clearly implies that the correlation among the individual severities in a multi-vehicle crash may exist. Hence, the proposed HBL model may be more appropriate in modeling the data than OBL model. The results for model calibration as well as quantitative assessment are presented in this section.

In the model calibration, beginning with the 15 covariates in the data set, each variable was tested for the statistical significance and the insignificant ones were eliminated. In the final model, three chains of 20,000 iterations each produced trace plots with a good degree of mixing, and Brooks, Gelman and Rubin convergence diagnostics (Brooks and Gelman, 1998) using Bayesian Output Analysis (BOA) program (Smith, 2001) indicated convergence. Particularly, after discarding 10000 burn-in samples and thinning to retain every fifth sample to reduce autocorrelation (leaving a total of 6000 posterior samples), the 0.975 quantiles of the corrected scale reduction factor (CSRF) for the parameters were each 1.2 or less. Posterior distributions were all uni-modal. The means, standard deviations and associated 95% BCI of estimated random effects and regression coefficients were monitored and listed in the Table 6.3.

TABLE 6.3 Posterior Summaries of Parameter Estimates

Parameters	Effect Estimate		Odds Ratio	95% BCI of Odds Ratio	
	Mean	S.D.		2.5%	97.5%
Fixed effects					
Time of Day					
Day time*	0	0	1.00	1.00	1.00
Night time	0.17	0.09	1.19	1.04	1.39
Peak time	-0.89	0.36	0.41	0.12	0.85
Intersection Type					
X intersection	-0.72	1.27	0.49	0.07	5.38
T/Y intersection	0.18	0.06	1.20	1.02	1.36
Other types*	0	0	1.00	1.00	1.00
Nature of Lane					
Single lane	-1.05	0.98	0.35	0.07	2.27
Left-most lane	-0.37	0.42	0.69	0.33	1.50
Right-most lane	0.23	0.08	1.26	1.07	1.83
Centre lane*	0	0	1.00	1.00	1.00
Street Lighting	-1.17	0.34	0.31	0.14	0.59
Presence of Red Light Camera	0.73	0.12	2.08	1.68	2.53
Pedestrian Involved	-0.96	0.46	0.38	0.14	0.92
Vehicle type					
Two-wheel vehicle	1.29	0.21	3.63	2.53	5.75
Light vehicle*	0	0	1.00	1.00	1.00
Heavy vehicle	-2.07	0.36	0.13	0.11	0.23
Driver Age					
<= 25	0.15	0.13	1.16	1.02	1.43
26 – 45*	0	0	1.00	1.00	1.00
46 – 65	-0.16	0.19	0.85	0.61	1.19
> 65	0.53	0.28	1.70	1.03	3.74
Involvement of Offending Party	0.49	0.13	1.63	1.21	2.14
Random Effects					
between-crash variance (τ_0^2)	1.34	0.87		0.56	2.29
within-crash variance	3.29				
ICC	0.289				

* represents the reference category used in the model for the multinomial variable

To check the model adequacy, underlying assumptions for the HBL model in Equation (5.5) were assessed. Posterior samples of the crash-level random effects (u_{0j}) can be thought of as residuals, and thus can be examined with usual model diagnostics. In the MCMC simulation, 200 random effects u_{0j} were randomly sampled, and the fact that they averaged very close to zero was reassured. Normal probability plots, revealing no strong abnormalities, also validate the normality and exchangeability assumptions.

As shown in Table 6.3, the variance of u_{0j} (τ_0^2), indicating the magnitude of the between-crash variance, is 1.34. Hence, the ICC is calculated by:

$$\rho = \frac{1.34}{1.34 + \pi^2/3} = 28.9\%$$

This means that 28.9% of unexplained variations in individual severity were resulted from between-crash variance, which strongly suggests the usefulness of the model specification of hierarchical structure. If an OBL mode was implemented without considering the random effects between crashes, the results will be biased and inaccurate.

Model comparison using DIC further strengthened this argument. DIC values for fitted OBL model (Equation (5.7)) and HBL model (Equation 5.5)) are given in Table 6.4. Results show that $\overline{D(\gamma)}$ of HBL model (1984.5) is less than one third of that obtained in OBL model (6165.5). After penalized by p_D, the DIC value for HBL model (3067.9) is also hugely less than that in OBL model (6191.9). This further proves that the use of crash-level random effects in HBL model can substantially improve the model fit.

TABLE 6.4 Results of Model Comparison using DIC

	$\overline{D(\gamma)}$	$D(\bar{\gamma})$	p_D	DIC
Ordinary logistic model	6165.5	6139.1	26.4	6191.9
Hierarchical logistic model	1984.5	901.1	1083.4	3067.9

6.4 DISCUSSIONS ON SIGNIFICANT RISK FACTORS

Summary statistics for the posterior samples of fixed effects of significant covariates are presented in Table 6.3. In the final HBL model, 9 variables are identified as significant judged by 95% BCI. They are: 1) *Time of day*, 2) *Intersection type*, 3) *Nature of lane*, 4) *Street lighting*, 5) *Presence of red light camera*, 6) *Pedestrian involved*, 7) *Vehicle type*, 8) *Driver age*, 9) *Involvement of offending party*. The detailed interpretations for these significant risk factors are offered in the following.

Time of Day

The time of crash occurrence is classified into 3 periods, i.e. day time (10am – 5pm), night time (8pm – 7am), and peak time (7am – 10am or 5 pm – 8 pm). Compared with crash occurrences during day time, crashes which occur at night time have 19% higher odds of high severity (IS(A)) (O.R. 1.19, 95% BCI (1.04, 1.39)). This finding is consistent with Simoncic (2001) who found crashes at night were more serious than those during daytime. This may be expected since speeding and alcohol use resulting in higher crash severity are more likely in these hours. Moreover, at night the effect of street lighting comes into play and this was also found to be significant in this study. The high probability of IS(A) in night time is consistent with previous studies for

severities of motorcycle crashes (Quddus et al., 2002) and single vehicle crashes (Rifaat et al., 2005) in Singapore. Furthermore, individuals involved at crashes in peak time (O.R. 0.41, 95% BCI (0.12, 0.85)) are also found to have reduced odds of being IS(A) by 60%. It can be reasoned that due to the higher traffic volume, the vehicle speeds during peak time are substantially reduced compared to off-peak time, hence resulting in lower crash severity. This is consistent with Zhang et al. (2000), in which the odds of fatality in crashes that occurred in 70-90 kph zones were almost six times more than those in crashes occurring in zones with slower speeds.

Intersection Type

It is found that crashes occurring at T/Y type intersections (O.R. 1.20, 95% BCI (1.02, 1.36)) increase the odds of being IS(A) by 20%, in contrast to other type of intersections. Results indicate that, though insignificant, X type intersections may have an averagely positive effect on reducing the crash severity. Vehicles on the minor road at T/Y type intersections, merging into the major road, have a higher probability to be seriously collided by the going-through vehicles on the major road. This is similar to the right-turn traffic (left-driving) at X type intersections. In addition, a shorter sight distance, commonly associated with a T/Y type intersections, may also be a factor causing more severe crashes.

Nature of Lane

Another significant geometric factor is *Nature of Lane*, where the right-most (left driving) lane (O.R. 1.26, 95% BCI (1.07. 1.83)) is identified to be significant on increasing the odds of severe crashes by 26%, compared with central lane. This result is consistent with the Khorashadi et al. (2005) who found that for right driving, if the location of collision is on the left lane, the likelihood of injury severity increased by 268.1%. The higher severity risk may be caused by higher speed on right-most lane than on other lanes. According to Bedard et al. (2002), traveling at speeds exceeding 112 kph was independently associated with a 164% increase in the odds of a fatality compared with speeds less than 56 kph.

Street Lighting

Street Lighting is identified as a significant factor (O.R. 0.31, 95% BCI (0.14, 0.59)). The odds ratios value indicates that a bad street lighting condition can increase the odds of severe crash by about 69%. This result is generally expected because drivers may have more reaction time and better perception ability on crash risk in good street lighting environments. Yau (2004) also found that street lighting condition affects the crash severity for the single vehicle crashes in Hongkong. This finding implies that improving the street lighting can substantially improve the safety condition at intersections.

Presence of Red Light Camera

Results show that among the highly significant risk factors, *Presence of red light camera* (O.R. 2.08, 95% BCI (1.68, 2.53)) is associated negatively with crash severity. In other words, the presence of red light camera is associated with higher severity level. In the sites with red light camera, the odds of being IS(A) increase by 108%. This may seem surprising compared to findings in many studies in which the red light camera has been proved to be useful in reducing the violation and crash frequencies, as well as relieving the crash severity. In a recent driver behavior study in Singapore, Huang and Chin (2006) have found that the presence of a red light camera is effective in curbing the red light running as well as reducing crash risk in angle crashes. Although red light camera itself may not increase the risk of severe crashes, it is associated with high risk sites. Specifically, intersections with red light camera may have already been placed in sites with more severe crashes since traffic authorities always install cameras at extraordinarily hazardous sites. Moreover, this reinforces the findings by Chin and Quddus (2003), where the presence of a surveillance camera was found to be associated with an increase in the total crash frequency at intersections. These results imply that, keeping other covariates unchanged, some unmeasured factors may have effects on the relative severity.

Pedestrian Involved

The variable *Pedestrian involved* is a significant factor affecting driver severity (O.R. 0.38, 95% BCI (0.14, 0.92)). The involvement of pedestrians substantially reduces the odds of being IS(A) by about 62%. This is intuitively reasonable since pedestrians,

rather than the drivers, are much easier to be injured seriously in the collisions. It is also supported by Chang and Wang (2006), who found that pedestrians were more likely to have higher risks of being injured than other types of vehicle drivers in traffic crash. Crash severity statistic also confirms this finding that of driver-vehicle units involved in the crashes of "vehicle against pedestrian" type, only 3.4% were injured severely and/or damaged extensively, compared with the overall rate of 16.5% as shown in Table 6.1.

Vehicle Type

Vehicle type is categorized as three groups in this study, i.e. two-wheel vehicle, light vehicle, and heavy vehicle. By taking the most common light vehicle as reference, the other two dummy variables for two-wheel vehicle (O.R. 3.63, 95% BCI (2.53, 5.75)) and heavy vehicle (O.R. 0.13, 95% BCI (0.11, 0.23)) were all found to have significant effects on individual severity. Compared with light vehicle, two-wheel vehicle increased the odds of being IS(A) by 263%, representing the most significant factor in the model. The severity risk in two-wheel vehicle (e.g. motorcycles) is expected as two-wheel riders have not the facility of safety protections that are available in light vehicle (e.g. cars), such as seatbelt, airbag etc. Again the two-wheeler driver may be thrown off from the vehicle at the time of collision while in the case of car crashes this may rarely happen. Kockelman and Kweon (2002) found that riding a motorcycle is causing more severe injury than driving a car. Again heavy vehicle reduces the odds of being IS(A) by 87%. It is not surprising that as the vehicle weight increases, the risks of being injured or damaged decrease substantially, even though other driver-vehicle units involved in the same crash may be more vulnerable to be injured or damaged.

This finding is also supported by Levine et al. (1999), who reported that every 454 kg (1000 lbs) increase in vehicle weight was equivalent to the driver's ability to withstand front impact crashes of 10 more kph (6 mph) before being fatally injured. However, it is interesting to notice that as found in Rifaat et al. (2005), the truck crashes in single vehicle crashes are more likely to result in serious injuries and fatalities. This contradiction can be explained by the different collision types between intersection crash and single vehicle crash. In contrast to intersection crash, more severe crashes may be caused by higher energy exchange for trucks with roadside objects in single vehicle crashes. Moreover, as found in Rifaat and Chin (2005), the higher relative fatality risk was associated with truck crashes mainly on high speed roads such as expressway rather than other highway types where signalized intersections are located.

Driver Age

The demographic variable, *Driver age*, is found to be significant on individual severity, in which both young group (O.R. 1.16, 95% BCI (1.02, 1.43)) and aged group (O.R. 1.70, 95% BCI (1.03, 3.74)) are identified to have effects on increasing the odds of being IS(A). Odds ratios indicate that a 16% increase of the IS(A) odds is associated with young drivers while 70% for aged drivers. It is likely because young drivers drive more recklessly (Rifaat and Chin, 2005; Kocklelman et al., 2002) while aged drivers have relatively weak risk detecting and reacting abilities. Again Hilakivi et al. (1989) also showed that young drivers as well as older drivers are more at risk of being involved in severe crashes. Another reason for young drivers to be involved with severe crashes may be that they represent a large proportion of riders of two-wheel vehicles, which have been proven to be associated with a higher risk of being involved

in more severe crashes (Rifaat and Chin, 2005; Quddus et al., 2002). Furthermore, as indicated by Rifaat and Chin (2005), decrease of visual power, deterioration of muscle strength and reaction time may be responsible for the aged drivers to be involved in severe crashes.

Involvement of Offending Party

Involvement of Offending Party affects crash severity significantly (O.R. 1.63, 95% BCI (1.21, 2.14)). The at-fault driver-vehicle unit has 63% higher odds to be IS(A) than the not at-fault party. This provides a more convincible evidence for educating drivers to keep away from risk-taking maneuvers.

6.5 SUMMARY

This study developed a Bayesian HBL model to identify the risk factors on individual severity of driver injury and vehicle damage at urban intersections. It is helpful to account for the severity correlation of driver-vehicle units involved in the same multi-vehicle crashes. The estimation of random effects using ICC showed that 28.9% of unexplained variation in severity level was resulted from between-crash variance. Model comparison with ordinary logistic model using DIC further ensured the suitability and model-improving effectiveness of introducing the crash-level random effects. This means, if ordinary logistic model were used, 28.9% residual variance could not be explained by this model, which might result in inaccurate coefficient estimates of risk factors. The Bayesian hierarchical modeling approach also showed flexibilities to explicitly explore the hierarchical data structure in traffic safety field.

Of the covariates including various geometric features, traffic conditions, and driver-vehicle characteristics, 9 variables were identified as significant using 95% BCI. Among these, the crash-level significant factors are *Time of day*, *Intersection type*, *Nature of lane*, *Street lighting*, *Presence of red light camera*, and *Pedestrian involved*. In particular, it was found that crashes occurring in peak time, in good street lighting condition, and in the case of pedestrians involved are associated with lower severity, while those occurring in night time, at T/Y type intersections, on right-most lane, in the presence of red light cameras have larger odds of being severe. *Vehicle type*, *Driver age* and *Involvement of offending party* were also found to affect severities of driver injury and vehicle damage significantly. Specifically, results indicated that heavy vehicles have a better resistance on serious injury or extensive damage, while two-wheel vehicles, young or aged drivers, with the involvement of offending party have a higher risk of being high severity.

This study of CSPM has a great potential in traffic safety discipline, especially when the correlation exists in the data set. This study illustrated a way to analyze the potential within-crash correlations in severity study using the hierarchical modeling technique. It also proved and emphasized the importance of accounting for this kind of within-cluster correlation in yielding reliable and accurate effect estimates for various risk factors.

CHAPTER SEVEN

CONCLUSIONS AND RECOMMENDATIONS

7.1 CONCLUSIONS AND RESEARCH CONTRIBUTIONS

Crash prediction model (CPM) is one of the most important techniques in investigating the relationship of road traffic crash occurrence and various risk factors. Traditional models using generalized linear regression are incapable of taking into account the within-cluster correlations, which extensively exist in crash data generating or collecting process.

To overcome the problem, this study developed a Bayesian hierarchical method to analyze the traffic crash frequency and severity. It demonstrated the flexibilities and effectiveness of the Bayesian hierarchical modeling approach in explicitly modeling the multilevel structure and excess zeros in traffic safety data. Furthermore, this study also explored a theoretical framework to determine the suitability of applying various statistical safety models in predicting traffic crash frequency and severity. The proposed method has a great potential in traffic safety discipline. While most previous studies ignored the multilevel structure in traffic crash data, this study proved and emphasized the importance of accounting for the within-cluster correlation in yielding reliable and accurate effect estimates for various risk factors.

7.1.1 Crash Frequency Prediction Model (CFPM)

To account for the multilevel data structure and excess zeros in crash frequency prediction model (CFPM), this study innovatively developed zero-inflated model with location-specific random effects (REZIP). The results showed that REZIP could be used as an alternative to the ordinary random effect Poisson model (REP) or zero-inflated Poisson model (ZIP).

A methodological framework using Bayesian analysis with Markov Chain Monte Carlo (MCMC) algorithm for model specification was proposed. A computing programme using BUGS language was, in the first time, developed to calibrate the REZIP model. Bayesian credible interval (BCI) was used to examine the significance of estimated parameters. This framework was also shown to provide a reliable measure to fit various flexible models. A cross-validation assessment method in the Bayesian framework, i.e. cross-validation predictive densities, was innovatively adopted to evaluate the suitability of the models. Several utility functions including the mean predictive square error (MPSE) and disaggregate predictive probability-based utilities, as well as their BCI measures were developed to analyze the cross-validation results. The assessment measures proved to be useful and reliable to examine the predictive performance of the whole model as well as the realization of individual observations in the data, for instance, “zero” occurrence in crash data.

Using intersection data in Singapore, the illustrative results indicated that REZIP could significantly perform better in terms of predictive abilities over the other candidate models (REP and ZIP). Specifically, judged by the criteria of MPSE ($\bar{u}$), models

accounting for excess zeros have been demonstrated to have a significant improvement in predictive abilities ($\bar{u}_{ZIP}$=1.97, $\bar{u}_{REZIP}$=1.89). The consideration of location-specific random effects in the ZIP model, resulting in REZIP model, yields the smallest predictive square errors. And REZIP model also has the smallest credible interval width around the observations ($\bar{u}_{REZIP(\alpha=0.95)}$ =3.06). Using the probability-based predictive utility $u(f)$ for the whole dataset, the result implies that, compared to REP model, REZIP model can increase the predictive accuracy as the overall percentage of model fitness by about 15%, i.e. from 29% to 44%.

As for the parameter estimation, a small but significant decreasing time trend of crash occurrence was identified in the model. Several factors were found to be significant in affecting the crash frequency including *Total approach volume, Conflicting approach volume, Number of lanes* and *Presence of median, Sight distance, Distance of the bus stop from the intersection, Number of phases per cycle, Red duration in pedestrian crossing.* The differences of parameter estimations between different models also imply that careful model development and assessment should be conducted since different specifications could result in quite different effect estimates as well as in their credible intervals.

7.1.2 Crash Severity Prediction Model (CSPM)

In crash severity prediction model (CSPM), a hierarchical binomial logistic model (HBL) was developed to identify the risk factors on individual severity in traffic crashes. It is capable of accounting for the severity correlation of driver-vehicle units involved in the same multi-vehicle crashes. A full Bayesian method with MCMC

algorithm was employed for model calibration to explicitly model the two-level data structure, i.e. crash-level and individual-level. A computing programme was specifically developed using BUGS language to realize the proposed algorithm. Intra-class Correlation Coefficient (ICC) was employed to assess the random effects, and the Deviance Information Criterion (DIC) was developed for model comparison.

Using Singapore crash data, a CSPM on individual severity of driver injury and vehicle damage at signalized intersections was developed to illustrate and validate the proposed method. The estimation of random effects using ICC showed that 28.9% of unexplained variation in severity level was resulted from between-crash variance. Model comparison with ordinary binomial logistic model (OBL) using DIC further ensured the suitability and model-improving effectiveness of introducing the crash-level random effects ($DIC_{HBL} = 6191.9$, $DIC_{OBL} = 3067.9$).

Of the covariates including various geometric features, traffic conditions, and driver-vehicle characteristics, 9 variables were identified as significant using 95% Bayesian Credible I. Among these, the crash-level significant factors are *Time of day*, *Intersection type*, *Nature of lane*, *Street lighting*, *Presence of red light camera*, and *Pedestrian involved*. In particular, it was found that crashes occurring in peak time, in good street lighting condition, and in case of pedestrian involved are associated with lower severity, while those occurring in night time, at T/Y type intersections, on right-most lane, and in the presence of red light camera have larger odds of being severe. *Vehicle type, Driver age* and *Involvement of offending party* were also found to affect severities of driver injury and vehicle damage significantly. Specifically, results indicated that heavy vehicles have a better resistance on serious injury or extensive

damage, while two-wheel vehicles, young or aged drivers, and the involvement of an offending party have a higher risk of a more serious injury or damage.

7.2 RECOMMENDATIONS FOR FUTURE RESEARCH

The Bayesian hierarchical methodology developed in this study has great potentials of extension as well as application for future research in traffic crash analysis. Three major directions are outlined in this section, i.e. multilevel structure in traffic safety data, other possible model formulations, and Bayesian updating function for CPMs.

7.2.1 Multilevel Structure in Traffic Safety Data

Multilevel data structures are commonly ignored in the traffic safety studies. This study developed the Bayesian hierarchical method to model the within-location correlation in crash frequency prediction and the within-crash correlation in crash severity analysis. But the multilevel data structure in traffic data is not only limited in location-specific and crash-specific correlation in CPMs. A more general form can be proposed for traffic safety study to be a $5\times T$ -level hierarchy, i.e. geographic region – traffic site – crash – driver-vehicle unit – occupant, as shown in Figure 7.1. The involvement and emphasis for different sub-groups of these levels depend on different research purposes and also rely on the heterogeneity examination on crash data employed. Generally, macro-analysis focus on the former three levels, i.e. geographic region level, traffic site level, and crash level, while micro-analysis concern the later three levels, i.e. crash level, driver-vehicle unit level, and vehicle occupant level. The Bayesian hierarchical modeling method provides us with a flexible and reliable model calibration and assessment measure for these potential explorations and applications.

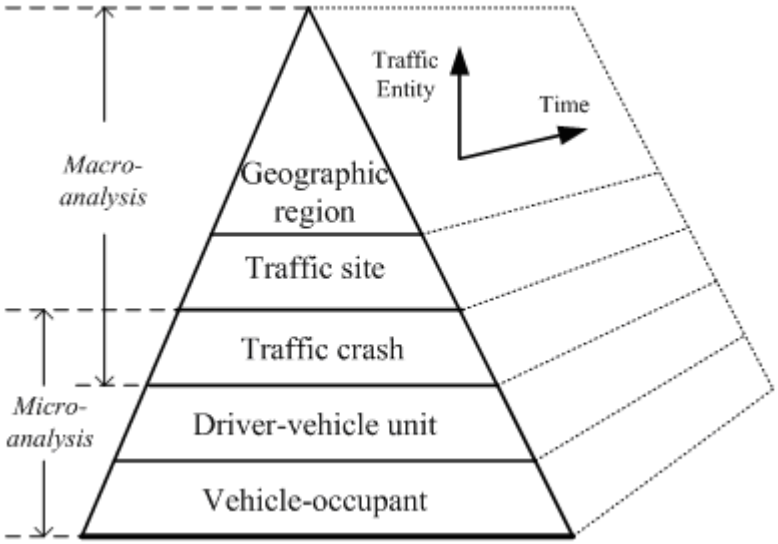

Figure 7.1 A $5 \times T$ -Level Hierarchy in Traffic Safety Data

7.2.2 Other Possible Model Formulations

It should be noted that all the considerations and treatments in this study are aimed at accounting for the possible sources of over-dispersion in crash data. In particular, while hierarchical models take the physical data collection scheme into consideration, zero-inflated models assume a dual-state data-generating process to explain the excess zeros. Hence, in crash frequency prediction, a natural extension of the proposed methodology is to negative binomial (NB) model, which even allows within-cluster dispersion. Intrigued by the proposed REZIP model, the NB model accounting for both random effect and zero-inflation can also be investigated. And in the crash severity prediction, the hierarchical multinomial models as well as ordered models can also be developed to account for the special characteristics of dependent variables representing crash severity levels. These non-nested complicated models can be implemented and

compared in the proposed Bayesian framework which provides a fairly flexible and reliable tool for model specification, model calibration as well as suitability assessment.

On the other hand, while this study only considered the random intercept in the regression equations, the random effects on the covariate coefficients can also be examined with careful specifications, resulting in random slope models. In the random slope models, the cross-level interaction between covariates could be appropriately specified and estimated.

7.2.3 Bayesian Updating Function for CPM

From the practical perspective, the Bayesian statistics can accumulate evidences in favor of any model. In Bayesian modeling technique, specifying the prior amounts to introducing extra information or data based on accumulated knowledge, and the posterior estimate in being based on the combined sources of information (prior and likelihood) therefore has greater precision. Moreover, within Bayesian framework, that data may be analyzed sequentially, with no loss of information. A model can be fit to data at any time, resulting in posterior distribution for all parameters of a model. If additional data become available generated by the same process, then the posteriors from the first analysis serve as the priors for the second analysis, and the result is the same as if the two sets of data were estimated simultaneously.

All the practical properties of Bayesian technique mentioned above naturally make us expect its possible application to innovatively improve the development of CPMs. A

special property of the CPMs among most the traffic safety problems is that the data is difficult to collect and gradually available along the time scale, e.g. year by year. And furthermore, there are many possible variations for the prediction models itself as the outcome of changes of some influential factors, e.g. the installation of red light camera, or the adjust of amber interval time. This means that, to make the models valid, we need update them periodically with the coming of new data. Fortunately, the Bayesian algorithm provides a quite flexible and reliable measure to realize this updating requirement. In Bayesian context, the previous model could be used as the prior knowledge of the updated model, in other words, the posterior distributions of model parameters are used as the prior distributions of the parameters in new model, which will be updated only by the newly-collected data to obtain the new posterior distributions of the model parameters. Hence, a Bayesian updating system could be developed for CPMs in future study. The recently-developed computational methods along with improvements in computing speed have made it possible to compute Bayesian inference for more complicated models on larger datasets.

REFERENCES

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accident Analysis and Prevention 37(3), 417-425.

Abdel-Aty, M., Radwan, E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention 32(5), 633-642.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In Proc 2nd Int. Symp. Information Theory, Budapest: Akademiai Kiado, 267-281.

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34(6), 729-741.

Angers, J.F., Biswas, A., 2003. A Bayesian analysis of zero-inflated generalized Poisson model. Computational Statistics and Data Analysis 42, 37-46.

Barron, D.N., 1998. The analysis of count data: overdispersion and autocorrelation. Sociological Methodology 22, 179-219.

Bedard, M., Guyatt, G..H., Stones, M.J., Hirdes, J.P., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis and Prevention 34(6), 717-727.

Ben-Akiva, M., Lerman, S. R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, MA: MIT Press.

Box, E.P., Hunter, W.G., Hunter, J.S., 1978. Statistics for Experimenters. John Wiley and Sons, New York.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulation. Journal of Computational and Graphical Statistics 7(4), 434 -455.

Chang L.Y., Mannering, F., 1999. Analysis of Vehicle Occupancy and the Severity of Truck- and Non-Truck-Involved Accidents. Accident Analysis and Prevention 31(5), 579-592.

Chang, L.Y., Wang H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38(5), 1019-1027.

Chin, H.C., 2007. Cost of traffic accidents in Singapore. Presented in "Global Road Safety Week 2007, Road Safety Conference, Towards Better Road Safety for All." Organized by LTA Academy, Singapore.

Chin, H.C., Haque, M.M., Yap H.J., 2006. An estimate of road accident costs in Singapore. In Proceedings of International Conference on Road Safety in Developing Countries, 28-35, Dhaka, Bangladesh.

Chin, H.C., Quddus, M.A., 2003a. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis and Prevention 35(2), 253-259.

Chin, H.C., Quddus, M.A., 2003b. Modeling count data with excess zeros. Sociological Methods and Research 32(1), 90-116.

Chin, H.C., Quek, S.T., 1997. Measurement of traffic conflicts. Safety Science 26(3), 169-185.

Congdon, P., 2003. Applied Bayesian Modeling. John Wiley & Sons, Ltd.

Evans, L., 1992. Car size or car mass: which has greater influence on fatality risk? American Journal of public Health 82(8), 1105-1112.

Evans, L., 1993. Mass ratio and relative driver fatality risk in two-vehicle crashes. Accident Analysis and Prevention 25(2), 213-224.

Evans, L., Frick, M., 1994. Car mass and fatality risk: has the relationship changed? American Journal of Public Health 84(1), 33-36.

Gelfand, A.E., Dey, D.K. 1994. Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society, Series B, 56, 501-514.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398-409.

Gelman, A., Carlin, J.B., Stern, H.S., 2003. Bayesian Data Analysis, 2nd edition. Chapman & Hall, New York.

Ghosh, S.K., Pabak, M., Lu, J.C., 2006. Bayesian analysis of zero-inflated regression models. Journal of Statistical Planning and Inference 136, 1360-1375.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1995. Markov Chain Monte Carlo Methods in Practice. Chapman & Hall, New York.

Goldstein, H., 2003. Multilevel Statistical Models, 3rd Edition. Edward Arnold.

Greene, W.H., 1995. LIMDEP Version 7.0 User's Manual. Econometric Software, Inc.

Greene, W.H., 1997. Econometric Analysis. Prentice Hall, New Jersey.

Haight, F.A., 1967. Handbook of the Poisson Distribution. John Wiley & Sons.

Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56, 1030-1039.

Hauer, E., 1992. Traffic conflicts and exposure. Accident Analysis and Prevention 14(5), 359-364.

Hauer, E., 2006. The frequency-severity indeterminacy. Accident Analysis and Prevention 38(1), 78-83.

Hausman, J.C., Hall, B.H., Griliches, Z., 1984. Econometric models for count data with an application to the patents - RandD relationship. Econometrica 52(4), 909-938.

Hilakivi, I., Veilahti, J., Asplund, P., Sinivuo, J., Laitinen, L., Koskenvuo, K., 1989. A sixteen-factor personality test for predicting automobile driving accidents of young drivers. Accident Analysis and Prevention 21(5), 413-418.

Hinde, J., 1982. Compound Poisson regression models. In GLIM 82: Proceedings of the International Conference on Generalized Linear Models, R. Gilchrist (ed), 109-121, New York: Springer-Verlag.

Huang, H.L., Chin, H.C., Heng, H.H., 2006. Effect of red light camera on accident risk at intersections. Transportation Research Record 1969, 18-26.

James, J.L., Kim, K.E., 1996. Restraint use by children involved in crashes in Hawaii, 1986-1991. Transportation Research Record 1560, 8-11.

Jones, B., Jansen, L., Mannering, F.L., 1991. Analysis of the frequency and duration of the freeway accidents in Seattle. Accident Analysis and Prevention 23(4), 239-55.

Jones, A.P., Jorgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. Accident Analysis and Prevention 35(1), 59-69.

Jones, I.S., Whitfield, R.A., 1988. Predicting injury risk with “New Car Assessment Program” crashworthiness ratings. Accident Analysis and Prevention 20(6), 411-419.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accidents rate and involvements using linear and Poisson regression models. Transportation Planning and Technology 15(1), 41-58.

Jovanis, P., Chang, H., 1986. Modeling the relationship of accident to mile traveled. Transportation Research Record 1068, 42-51.

Karen, C.H. Yip, Kelvin, K.W. Yau, 2005. On modeling claim frequency data in general insurance with extra zeros. Insurances: Mathematics and Economics 36, 153-163.

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an explanatory analysis. Accident Analysis and Prevention 37(5), 910-921.

Kim, D.G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accident Analysis and Prevention 39(1), 125-134.

Kim, D.G., Washington, S., 2006. The significance of endogeneity problems in crash models: an examination of left-turn lanes in intersection crash models. Accident Analysis and Prevention 38(6), 1094-1100.

Kim, D.G., Washington, S. Oh, Jutaek, 2006. Modeling crash types: new insights into the effects of covariates on crashes at rural intersections. Journal of Transportation Engineering 132(4), 282-292.

King, G., 1989. Event count models for international relations: generalization and applications. International Studies Quarterly 33, 123-147.

Kocklelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34(3), 313-321.

Kulmala, R., 1995. Safety at rural three- and four-arm junctions: development and application of accident prediction models. Technical Research Center at Finland, VTT Publications, Espoo.

Kumara, S.S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized T-intersections with special emphasis on excess zeros. Traffic Injury Prevention. 3(4), 53-57.

Lambert, D., 1992. Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics 34, 1-14.

Land, K.C., McCall, P.L., Nagin, D.S., 1996. A comparison of Poisson, negative binomial and semiparametric mixed Poisson regressive models with empirical applications to criminal careers data. Sociological Methods and Research 24, 387-442.

Lawless, J. F., 1987. Negative binomial and mixed Poisson regressions. Canadian Journal of Statistics 15, 209-225.

Lee, A.H., Stevenson, M.R., Wang, K., Kelvin, K.W. Yau, 2002. Modeling young driver motor vehicle crashes: data with extra zeros. Accident Analysis and Prevention 34, 515-521.

Lee, J., Mannering, F.L., 2002. Impact of roadside features on the frequency and severity of run-off-road accidents: an empirical analysis. Accident Analysis and Prevention 34(2), 349-361.

Lenguerrand, E., Martin, J.L., Laumon, B., 2006. Modeling the hierarchical structure of road crash data: application to severity analysis. Accident Analysis and Prevention 38(1), 43-53.

Levine, E., Bedard, M., Molloy, D.W., Basilevsky, A., 1999. Determinants of driver fatality risk in front impact fixed object collisions. Mature Medicine Canada 2, 239-242.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models for motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37(1), 35-46.

Lord, D., Washington, S.P., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accident Analysis and Prevention 39(1), 53-57.

Long, J.S., 1997. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA, Sage Publications.

Lui, K.J., McGee, D., Rhodes, P. and Pollock.D., 1988. An application of a conditional logistic regression to study the effects of safety belts, principal impact points and car weights on drivers' fatalities. Journal of Safety Research 19, 197- 203.

Mannering, F.L., Grodsky, L.L., 1995. Statistical analysis of motorcyclists' perceived accident risk. Accident Analysis and Prevention 27(1), 21–31.

Maycock, G., Hall, R.D., 1984. Accident at 4-Arm Roundabouts. Berks, UK: Transport and Road Research Laboratory.

McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. Frontiers of Econometrics, 105-142. New York: Academic Press.

McFadden, D. Econometric model of probabilistic choice. Structural Analysis of Discrete Data, 198-272. Cambridge, MA: MIT Press. 1981.

Mercier, C.R., Shelley, M.C., Rimkus, J., Mercier, J. M., 1997. Age and gender as predictors of injury severity in head-on highway vehicular collisions. Transportation Research Record 1581, 37-46.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road section: Poisson versus negative binomial regression. Accident Analysis and Prevention 26(4), 471-482.

Miaou, S.P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention 25(6), 689-709.

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39(3), 459-468.

O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis and Prevention 28(6), 739-753.

Peden, M. et al. Edited, 2004. World report on road traffic injury prevention. World Health Organization.

Poch, M., Mannering, F. L., 1996. Negative binomial analysis of intersection accident frequencies. Journal of Transportation Engineering 122(2), 105-113.

Porter, B. E., England, K. J., 2000. Predicting red-light running behavior: a traffic safety study in three urban settings. Journal of Safety Research 31(1), 1-8.

Qin, X, Ivan, J.N., Ravishanker, N., Liu J.F., 2005. Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov Chain Monte Carlo Modeling. Journal of Transportation Engineering 131(5), 345-351.

Quddus, M.A., Noland, R.B., Chin, H.C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. Journal of Safety Research 33(4), 445-462.

Raftery, A.E., 1986. Choosing models for cross-classifications (Comment on Grusky and Hauser). American Sociological Review 51, 145-46.

Raftery, A.E.,1995. Bayesian model selection in social research. In P. Marsden (Ed.), Sociological Methodology, Washington, DC: The American Sociological Association, 111-163..

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Longford, I., Lewis, T., 2000. A User's Guide to MLwiN, second ed. Institute of Education, London.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., Congdon, R., 2001. HLM 5: Hierarchical Linear and Nonlinear Modeling, second edition, Scientific Software International, Chicago.

Retting, R. A., Ulmer, R. G., Williams, A. F., 1999. Prevalence and characteristics of red right running crashes in the United States. Accident Analysis and Prevention 31(6), 687-694.

Rifaat, S.M., Chin, H.C., 2005. Analysis of severity of single-vehicle crashes in Singapore. In: TRB 2005 Annual Meeting CD-ROM, Transportation Research Board, National Research Council, Washington D.C.

Shankar, V.N., Albin, R.B., Milton, J.C., Mannering, F.L., 1998. Evaluation of median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effect negative binomial model. Transportation Research Record 1635, 44-48.

Shankar, V.N., Mannering, F., 1996. An exploratory multinomial Logit analysis of single-vehicle motorcycle accident severity. Journal of Safety Research 27(3), 183-194.

Shankar, V. N., Mannering, F. L., Barfield, W., 1995. Effect of roadway geometric and environmental factors on rural freeway accident frequencies. Accident Analysis and Prevention 27 (3), 371-389.

Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accident Analysis and Prevention 28(3), 391- 401.

Shankar, V. N., Milton, J.C., Mannering, F.L., 1997. Modeling accident frequencies as zero-altered probability process: an empirical enquiry. Accident Analysis and Prevention 29 (6), 829-837.

Shankar, V.N., Ulfarsson, G..F., Pendyala, R.M., Neberagal, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41(7), 627-640.

Shibata, A., Fukuda, K., 1994. Risk factors of fatality in motor vehicle traffic accidents. Accident Analysis and Prevention 26(3), 391- 397.

Simoncic, M., 2001. Road fatalities in Slovenia involving a pedestrian, cyclist or motorcyclist and a car. Accident Analysis and Prevention 33(2), 147-156.

Skinner, C.J., Holt, D., Smith, T.M.F., 1989. Analysis of Complex Surveys. Wiley, Chichester, UK.

Smith, B.J., 2001. Bayesian Output Analysis Program (BOA), Version 1.0.0 for S-PLUS and R. available at http://www.public-health.uiowa.edu/boa.

Snijders, A.B., Bosker, R.J., 2000. Multilevel Analysis, An Introduction to Basic and Advanced Multilevel Modeling. SAGE Publications, London.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lunn, D., 2003a. WinBUGS version 1.4.1 User Manual. MRC Biostatistics Unit, Cambridge, UK.

Speigelhalter, D. J., Best, N. G., Carlin, B. P., Linde, V. D. 2003b. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B, 64(4), 583-616.

Tanner, M., Wong, W., 1987. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association 82, 528-550.

Vehtari, A., Lampinen, J., 2002. Bayesian model assessment and comparison using cross-validation predictive densities. Neural Computation 14, 2439-2468.

Vogt, A., Bared, J., 1998. Accident models for two-lane rural segments and intersections. Transportation Research Record 1635, 18-29.

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica 57(2), 307-333.

Wahba, G., 1990. Spline Models for Observational Data. Philadelphia: Society for Industrial and Applied Mathematics.

Wang, K., Lee A.H., Kelvin, K. W. Yau, Philip, J.W., 2003. A bivariate zero-inflated Poisson regression model to analyze occupational injuries. Accident Analysis and Prevention 35, 625-629.

Washington, S., Congdon, P., Karlaftis, M., Mannering, G., 2005. Bayesian multinomial logit models: exploratory assessment of transportation applications. In: TRB 2005 Annual Meeting CD-ROM, Transportation Research Board, National Research Council, Washington D.C.

Xiao, Q., John, N.I., Nalini, R., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis and Prevention 36, 183-191.

Xie, M., He B., Goh, T.N., 2001. Zero-inflated Poisson model in statistical process control. Computational Statistics and Data Analysis 38, 191-201.

Yang, C. MacNab, 2003. A Bayesian hierarchical model for accident and injury surveillance. Accident Analysis and Prevention 35(1), 91-102.

Yau, K., 2004. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accident Analysis and Prevention 36(3), 333–340.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G., Mao, Y., 2000. Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. Accident Analysis and Prevention 32(1), 117-125.

APPENDICES

Appendix A

This appendix indicates the list of the intersections selected in the chapter four, as shown in Table A.1

TABLE A.1 The List of Signalized Intersections Within Study Area

Intersection ID	Name of the Connecting Roads
1	Commonwealth Avenue West, Clementi Avenue 3, Clementi Avenue 4
2	Commonwealth Avenue West, Clementi Avenue 2
3	Commonwealth Avenue West, Clementi Road
4	Commonwealth Avenue West, North Bouna Vista Road
5	Clementi Road, West Coast Road, Pasir Panjang Road
6	Commonwealth Avenue, Queensway
7	Commonwealth Avenue, Alexandra Road
8	Alexandra Road, Delta Road, Lower Delta Road
9	Clementi Road, West Coast Highway
10	Clementi Road, West Coast Road, Pasir panjang Road
11	Clementi Avenue 2, West Coast Road, Clementi west Street 2
12	Jurong East Ave 1, Jurong East State 32
13	Jurong East Ave 1, Jurong Town Hall Road
14	Jurong East Ave1, Toh Guan Road, Jurong East Central
15	Jurong East Central, Boon Lay Way
16	Jurong East Central, Jurong East Street 13
17	Jurong East Central, Jurong East Street 21
18	Jurong East Central, Jurong Town Hall Road
19	Jurong East Street 11, Jurong Town Hall Road

20	Clementi Road, Kent Ridge Crescent
21	Boon Lay Drive, Boon Lay Avenue
22	Jalan Ahammed Ibrahim, Jalan Boon Lay
23	Jalan Ahammed Ibrahim, Jurong Pier Road
24	Corporation Road, Jalan Ahammed Ibrahim
25	Jurong Port Road, Jalan Buroh
26	Ayer Rajah Expressway, Jurong Town Hall Road, Jln Ahamed Ibrahim
27	Jalan Ahammed Ibrahim, Jurong Port Road
28	Pan-Island Expressway, Jurong Town Hall Road, Bukit Batok Road
29	Boon Lay Way, Jurong Town Hall Road
30	Boon Lay Way, Jalan Boon Lay
31	Corporation Road, Corporation Drive
32	Boon Lay Way, Jurong West Street 51, Yung Ching Road
33	Boon Lay Way, Corporation Road
34	Bukit Timah Road, Caneagh Road
35	Bukit Timah Road, Clementi Road
36	Bukit Timah Road, Selegie Road
37	Bukit Timah Road, Stevens Road
38	Bukit Timah Road, Farrer Road
39	Dunearn Road, Adam Road, Whittey Road
40	Holland Road, Six Avenue
41	Alexandra Road, Tanglin Road
42	Commonwealth Drive, Tanglin Halt Road
43	Alexandra Road, Pasir Panjang Road, Telok Blangah Road
44	Alexandra Road, Queensway, Jalan Bukit Merah
45	Dover Road, North Bouna Vista Raod, AYE Avenue
46	Lower Kent Ridge, North Bouna Vista Road, From AYE
45	Corporation Drive, Ho Chin Road

46	Lower Delta Road, Jalan Bukit Merah
47	Lower Delta Road, Ayer Rajah Expressway
48	Lower Delta Road, Tiong Bahru Road
49	Henderson Road, Jalan Bukit Merah
50	Henderson Road, Tiong Bahru Road
51	Boon Lay Avenue, Jalan Boon Lay
52	Commonwealth Avenue, Commonwealth Drive, Holland Avenue

Appendix B

This appendix illustrates a portion of sample crash data, as shown in Table B.1

Table B.1 A Part of the Crash Data File Consisting All the Fields

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	IPNO	ACDTE	ACTIME	ACCLASS	SPF	CBDIN	CBDH	EX/	TYPELOC	GRIDCD1	GRIDCD2	STRCODE1	STRCODE2	CONS	HITRU
2	98149977	3/1/98	1823	3	E	N	N	8	01	496	443	CAR20		CA	N
3	98149984	3/1/98	2029	3	C	N	N	8	08	549	463	PAI01		MP	N
4	98149984	3/1/98	2029	3	C	N	N	8	08	549	463	PAI01		MP	N
5	98150010	3/1/98	2157	3	C	N	N	8	09	518	485	CEE01		PS	N
6	98150010	3/1/98	2157	3	C	N	N	8	09	518	485	CEE01		PS	N
7	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
8	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
9	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
10	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
11	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
12	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
13	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
14	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
15	98150027	3/1/98	2117	3	F	N	N	2	06	479	588	SER05	GAA01	SB	N
16	98150058	3/2/98	0046	3	D	N	N	8	08	410	457	COA03		CL	N
17	98150058	3/2/98	0046	3	D	N	N	8	08	410	457	COA03		CL	N
18	98150058	3/2/98	0046	3	D	N	N	8	08	410	457	COA03		CL	N
19	98150058	3/2/98	0046	3	D	N	N	8	08	410	457	COA03		CL	N
20	98150109	3/2/98	0517	3	J	N	N	8	01	368	465	AYR04		JR	N
21	98152894	3/2/98	0737	3	C	N	N	8	10	583	473	PAI01		KB	N
22	98152894	3/2/98	0737	3	C	N	N	8	10	583	473	PAI01		KB	N

	P	Q	R	S	T	U	V	W	X	Y	Z	A	AE
1	RDENGRNG	TYPETRAF	TYPEROAD	NATLANE	SPEEDLTD	SURVCAM	RDSURFAC	NATACC	NOMVEH	TYPECOLL	T	T	N
2	5	2	2	1	3	N	1	2	03	3			00
3	5	4	3	1	5	N	1	2	02	2			00
4	5	4	3	1	5	N	1	2	02	2			00
5	5	4	3	2	5	N	1	2	02	3			00
6	5	4	3	2	5	N	1	2	02	3			00
7	5	3	2	4	3	N	1	2	09	3			00
8	5	3	2	4	3	N	1	2	09	3			00
9	5	3	2	4	3	N	1	2	09	3			00
10	5	3	2	4	3	N	1	2	09	3			00
11	5	3	2	4	3	N	1	2	09	3			00
12	5	3	2	4	3	N	1	2	09	3			00
13	5	3	2	4	3	N	1	2	09	3			00
14	5	3	2	4	3	N	1	2	09	3			00
15	5	3	2	4	3	N	1	2	09	3			00
16	5	3	2	3	3	N	1	2	04	3			00
17	5	3	2	3	3	N	1	2	04	3			00
18	5	3	2	3	3	N	1	2	04	3			00
19	5	3	2	3	3	N	1	2	04	3			00
20	3	4	3	1	5	N	1	1	01				00
21	5	4	3	2	5	N	1	2	02	5			00
22	5	4	3	2	5	N	1	2	02	5			00

	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	A	AP
1	VEHCNTY	VEHTYPES	VEHMAKE	VEHCLY	VEHSCLAS	VEHMANO	VEHDAMAG	OFNGPRTY	CAUSACC1	CA	CA	CA	C	HDLGHTON
2	SG	06	T15		22	10	3	2						N
3	SG	05	Y01	0250	98	10	3	2						Y
4	SG	06	N06		98	08	3	1	101					Y
5	SG	06	D05		98	10	3	1	117					Y
6	SG	06	A38		98	10	3	2						Y
7	SG	08	Z99		24	03	3	2						Y
8	SG	10	F15		26	10	2	1	117					Y
9	SG	06	N06		98	03	3	2						Y
10	SG	06	Z99		98	03	3	2						Y
11	SG	06	Z99		98	03	3	2						Y
12	SG	06	T15		22	03	3	2						Y
13	SG	06	T15		23	03	3	1	117					Y
14	SG	06	T15		22	03	3	2						Y
15	SG	06	T15		23	03	3	2						Y
16	SG	13	Z99		24	03	3	2						Y
17	MY	06	Z99		98	03	3	2						Y
18	SG	06	Z99		98	10	2	1	117					Y
19	SG	06	T15		22	03	3	2						Y
20	SG	23	L07		01	08	3	1	101					Y
21	SG	06	T15		22	10	3	1	116					N
22	SG	01			98	10	3	2						N

	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
1	DRVAGE	DRVRACE	DRVOCC	DRVSEX	DRVINJ	DLCLASS	DRVGSCH	LICTYPE	DRVDMPT	DRVGMTH	DRVDISQ	DRVSUSP
2	42	CN	01	M		3	5	1	00	99	N	N
3	29	CN	99	M	3	2A	2	1	00	01	N	N
4	30	CN	99	F		3	5	1	06	99	N	N
5	48	CN	99	M	3	3	5	1	00	99	N	N
6	20	CN	23	M		3	5	1	00	04	N	N
7	39	CN	99	M		3	5	1	00	99	N	N
8	39	CN	02	M	3	3	5	1	00	00	N	N
9	28	CN	99	M		3	5	1	00	57	N	N
10	48	MY	99	M		3	5	1	00	61	N	N
11	30	MY	99	M		3	5	1	00	00	N	N
12	47	CN	01	M		3	5	1	06	99	N	N
13	37	CN	01	M		3	5	1	14	32	N	N
14	34	CN	01	M		3	5	1	00	99	N	N
15	44	CN	01	M		3	5	1	06	99	N	N
16	27	IN	03	M	3	3	1	1	00	00		
17	44	CN	99	M		3	5	1	00	00	N	N
18	42	CN	99	M	3	3	1	1	00	99	N	N
19	45	CN	01	M		3	1	1	00	99	N	N
20	24	CN	99	M	3	3	1	1	06	08	N	N
21	52	CN	01	M		3	5	1	00	99	N	N
22	73	CN	99	M	3				06	00	N	N

	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM
1	FPAGE	FPSEX	FPINJ	OMPKILL	OMPSEINJ	OMPSLINJ	OFPKILL	OFPSEINJ	OFPSLINJ	DAYWEEK	PPLATE
2				00	00	00	00	00	00	0	N
3	28	F	3	00	00	00	00	00	00	0	N
4				00	00	00	00	00	00	0	N
5				00	00	00	00	00	00	0	N
6				00	00	00	00	00	00	0	N
7				00	00	00	00	00	00	0	N
8				00	00	00	00	00	00	0	N
9				00	00	00	00	00	00	0	N
10				00	00	00	00	00	00	0	N
11				00	00	00	00	00	00	0	N
12				00	00	00	00	00	00	0	N
13				00	00	00	00	00	00	0	N
14				00	00	00	00	00	00	0	N
15				00	00	00	00	00	00	0	N
16	40	M	3	00	00	04	00	00	00	1	N
17				00	00	00	00	00	00	1	N
18				00	00	00	00	00	00	1	N
19				00	00	00	00	00	00	1	N
20				00	00	00	00	00	00	1	N
21				00	00	00	00	00	00	1	N
22				00	00	00	00	00	00	1	N

CURRICULUM VITAE

Huang Helai (黄合来)

1996-2000 B.E. Department of Civil Engineering, Tianjin University, P.R. China

2000-2003 M.E. Department of Civil Engineering, Tianjin University, P.R. China

2003-2007 Research Scholar, Centre for Transportation Research, Department of Civil Engineering, National University of Singapore, Singapore

List of Publications

1. Chin H.C., **Huang H.L.,** 2008. Modeling multilevel data in traffic safety: a Bayesian hierarchical approach. Invited book chapter in *Transportation Accident Analysis and Prevention*. Editor-in-chief: Frank Columbus. Nova Science Publisher, Inc.
2. **Huang H.L.**, Chin H.C., Heng H.H., 2006. Effect of red light camera on accident risk at intersections. *Transportation Research Record* 1969, Page 18-36. (also presented in *TRB 2006 Annual Meeting*, Transportation Research Board, Washington D.C.)
3. **Huang H.L.**, Chin H.C., Haque M.M., 2007. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. *Accident Analysis & Prevention*. (forthcoming)
4. **Huang H.L.**, Chin H.C., 2007. Disaggregate propensity study of RLR accident using quasi-induced exposure method. *Journal of Transportation Engineering (ASCE)*. (forthcoming)
5. **Huang H.L.**, Chin H.C., 2007. Bayesian zero-inflated Poisson regression with location-specific random effects on crash prediction model. *Journal of Transportation Engineering (ASCE)*. (forthcoming).
6. **Huang H.L.**, Chin H.C., Haque M.M., 2008. Bayesian hierarchical analysis on crash prediction models. *Transportation Research Record* (forthcoming) (also presented in *TRB 2008 Annual meeting*, Transportation Research Board, Washington D.C.)
7. Haque, M.M., Chin H.C. **Huang H.L.,** 2008. Examining exposure of motorcycles at signalized intersections. *Transportation Research Record* (forthcoming) (also presented in *TRB 2008 Annual Meeting*, Transportation Research Board, Washington D.C.)
8. Haque, M.M., Chin H.C., **Huang H.L.**, 2008. Modeling fault among motorcyclists involved in crashes. *Traffic Injury Prevention*. (forthcoming)
9. Zhang S.R., **Huang H.L.,** 2002. GIS-based river related management information system. *Journal of Irrigation and Drainage* 21(6), Page 9-12. (in Chinese)
10. **Huang H.L.**, Chin H.C., 2007. Safety countermeasure development and evaluation for signalized intersections in Singapore. *14th International Conference on Road Safety on Four Continents*, Bangkok, Thailand. Organizer: Swedish National Road and Transport Research Institute (VTI).
11. Haque M.M., Chin H.C, **Huang H.L.,** 2006. Modeling random effect and excess zero in road traffic accidents prediction. Proceedings of *The Nineteenth KKCNN Symposium on Civil Engineering*, Page 245-248. Kyoto, Japan.
12. Chin H.C., **Huang H.L.,** 2006. A safety evaluation procedure on traffic treatments for using modified empirical Bayesian approach. Proceedings of *International Conference on Traffic Safety in Developing Countries*, Page 143-147. Dhaka, Bangladesh.
13. Chin H.C., **Huang H.L.,** 2004. A re-examination of instances of red light running. Proceedings of *The Seventeenth KKCNN Symposium on Civil Engineering*, Page 663-668. Ayutthaya, Thailand.