Traffic safety at road junctions

Aine Kusumawati

2008

Aine Kusumawati. (2008). Traffic safety at road junctions. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/40638>

<https://doi.org/10.32657/10356/40638>

Traffic Safety at Road Junctions

Aine Kusumawati

School of Civil & Environmental Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2008

ACKNOWLEDGEMENTS

This research was carried out in the School of Civil and Environmental Engineering, Nanyang Technological University. The financial support given by the University is gratefully acknowledged. I would also thank the University for granting me a Teaching-Assistantship during the last few months of my study.

I highly appreciate Institute Technology Bandung and Directorate General of Higher Education-Ministry of Education, Indonesia for granting me leave of absence during my study period.

I would like to express my sincere gratitude to A/P Wong Yiik Diew, for his support, never ending guidance, valuable advice and encouragement throughout the duration of my research. The effort and time that he devoted to me has enabled me to complete my thesis.

I would also like to show my appreciation to Prof. Henry Fan, A/P Gopinath Menon, A/P Piotr Olszewski, A/P Lum Kit Meng, A/P Lam Soi Hoi, A/P Tan Yan Weng, A/P Francesco Cavalero and Ms. Lakshmy Bhaskar for giving me lectures, as part of my study requirements, which greatly support my research.

I highly appreciate the Transportation Laboratory for offering me opportunities to work in favourable conditions. My thank to the Transportation Laboratory's technical staff members: Mr. Choi Siew Pheng, Mr. Liew Kai Liang, Mrs. Ng-Ho Choo Hiang and Mrs. Tan-Ma Ju An for their enthusiastic and non-conditional assistances.

I would like to thank the Land Transport Authority, especially Mr. Simon Ho, Mr. Koon Chng Tang and Ms. Mary Tan, for providing the GLIDE traffic data and some of the junctions' geometric data, and also Mr. David Wanty, Ms. Hau Lay Peng and Ms. Koh Puay Ping for their useful information on the accident information system data. I would also like to thank the Traffic Police Department

for providing the traffic accident data. Without the help rendered by them, this project would not have progressed as smoothly.

Special thanks to Iwan Soemarjo Halim, Thu Ya Chit, Min Min Naing, Abeyweera Herrath M. Buddhi Anupama, Chia Pei Ling and Leong Mun Yi for their helps in field-collection of junctions' geometric and traffic data. Many thanks also to all my other friends at Transportation Laboratory for their kindness, friendship and encouragements.

I am thankful to my mother, Anny Pratondo and my father-in-law, Chaeruddin Mustafa for their sincere prayer and never ending support that enable me to get through my difficult times. Finally, my deepest gratitude to my dear husband, Ali Akbar Mustafa, for his full support, empathy and encouragement which have motivated me on numerous occasions and enabled me to complete my thesis and for taking care of my children Vanya Karina Akbar, Zakiyya Anandita Akbar and Tristan Makarim Akbar.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	ii
TABLE OF CONTENTS	iv
SUMMARY	vii
LIST OF TABLES	ix
LIST OF FIGURES	xi
LIST OF SYMBOLS	xiii
CHAPTER 1 INTRODUCTION	1
1.1 Background	1
1.2 Road safety situation in Singapore	3
1.3 Research objectives	8
1.4 Research scope	8
1.5 Report structure	10
CHAPTER 2 LITERATURE REVIEW	11
2.1 General	11
2.2 Road safety and traffic accident	11
2.3 Accident prediction models	17
2.3.1 Cross-sectional models	19
2.3.2 Time series models	24
2.3.3 Combined cross-sectional/trend models	24
2.4 Bayesian methods	26
2.5 Safety at road junction	28
2.6 Summary	37
CHAPTER 3 DATA DESCRIPTION	40
3.1 Introduction	40
3.2 Junction sample	40
3.3 Road traffic accident data	42
3.4 Junction geometric and operational data	57
3.5 Traffic flow data	60
3.6 Summary	65
CHAPTER 4 PROBABILITY DISTRIBUTION OF ACCIDENT COUNTS	66
4.1 Introduction	66
4.2 Background theory	66
4.3 Review of goodness-of-fit tests for discrete distributions	72
4.3.1 General goodness-of-fit tests	72
4.3.2 Goodness-of-fit tests for Poisson distribution	74

4.3.3	Goodness-of-fit tests for generalised Poisson (GP) distribution	76
4.4	The data	76
4.5	Testing the validity of Poisson assumption	77
4.6	Stationary Poisson process vs non-stationary Poisson process	80
4.7	Poisson distribution or generalised Poisson (GP) distribution?	82
4.7.1	Application of generalised Poisson (GP) distribution to accident data	82
4.7.2	Goodness-of-fit tests	83
4.7.3	Sums of squares of error (SSE) and χ^2 statistics	84
4.8	Summary	86
CHAPTER 5 METHODOLOGY FOR DEVELOPMENT OF ACCIDENT PREDICTION MODELS		88
5.1	Introduction	88
5.2	Background theory	88
5.2.1	Poisson regression model	88
5.2.2	Negative binomial (NB) regression model	89
5.2.3	Generalised Poisson (GP) regression model	90
5.2.4	Zero inflated Poisson (ZIP) regression model	91
5.2.5	Zero inflated negative binomial (ZINB) regression model	92
5.2.6	Assessment of model goodness-of-fit	94
5.2.7	Identification of outliers	95
5.3	Model formulation	98
5.4	Modelling strategy	101
5.5	Model explanatory variables	104
5.6	Modelling procedure	108
5.7	Summary	112
CHAPTER 6 ACCIDENT PREDICTION MODELS		114
6.1	Introduction	114
6.2	Long-term models	114
6.2.1	Long-term models for X-signalised junctions	115
6.2.2	Long-term models for T-signalised junctions	122
6.3	Short-term models	128
6.3.1	Short-term models for X-signalised junctions	129
6.3.2	Short-term models for T-signalised junctions	135
6.3.3	Short-term models for X-unsignalised junctions	144
6.3.4	Short-term models for T-unsignalised junctions	148
6.4	Comparison of models' performance	156
6.4.1	Comparison of X-signalised junctions models' performance	159
6.4.2	Comparison of T-signalised junctions models' performance	162
6.4.3	Comparison of X-unsignalised junctions models' performance	165
6.4.4	Comparison of T-unsignalised junctions models' performance	165
6.5	Sensitivity analysis	167
6.6	Factors associated with accident occurrences	169
6.6.1	Factors associated with accident occurrences at signalised junctions	170

6.6.2	Factors associated with accident occurrences at unsignalised Junctions	172
6.6.3	Factors associated with severity of accidents	174
6.7	Summary	176
CHAPTER 7 ESTIMATION OF SAFETY BY EMPIRICAL BAYES METHOD		180
7.1	Introduction	180
7.2	Empirical Bayes (EB) method	180
7.3	Modified empirical Bayes (MEB) method	182
7.4	Properties of EB and MEB estimates	186
7.4.1	Properties of EB estimate	187
7.4.2	Properties of MEB estimate	188
7.5	Procedures for safety estimation and prediction	193
7.5.1	A calculation procedure for estimation of safety	193
7.5.2	A calculation procedure for prediction of safety	195
7.5.3	Calculation example	196
7.6	Results	202
7.7	Validation of the results of prediction	210
7.8	Summary	215
CHAPTER 8 IDENTIFICATION AND RANKING OF HAZARDOUS JUNCTIONS		216
8.1	Introduction	216
8.2	Review on current methods to identify and rank hazardous sites	216
8.2.1	Methods to identify hazardous sites	217
8.2.2	Methods to ranking hazardous sites	218
8.2.3	Conclusions on review	220
8.3	The application of MEB estimate for identifying and ranking of hazardous junctions	221
8.4	Analysis and results	222
8.5	Summary	231
CHAPTER 9 CONCLUSIONS AND RECOMMENDATION		233
9.1	Conclusions	233
9.2	Research contributions	237
9.3	Recommendations	239
REFERENCES		242
APPENDICES (on CD-ROM)		
APPENDIX A	Road traffic safety measures (1972-2005)	
APPENDIX B	List of study junctions	

APPENDIX C	Accident report forms
APPENDIX D	Results of fitting to Poisson distribution
APPENDIX E	Correlation analysis of model variables
APPENDIX F	Results of safety estimation and prediction by EB method
APPENDIX G	Results of safety estimation and prediction by MEB method

SUMMARY

The research was undertaken with the aim of estimating safety at road junctions. Accident prediction models were developed in the form of associative relationships between accident occurrences and roadway characteristics, traffic and operational factors at the junctions. The developed models took account of different driving situations encountered by drivers through classifying the junctions according to X-signalised, T-signalised, X-unsignalised and T-unsignalised junctions. Outputs from the accident prediction models were used for safety estimation by means of the original empirical Bayes (EB) method and a modified empirical Bayes (MEB) method. Procedures for developing accident prediction models and safety estimation methods were established and applied to Singapore road junctions.

The accident prediction models for (X- and T-) signalised and (X- and T-) unsignalised junctions were developed using a negative binomial regression model. In general, two types of accident prediction models were developed in this research, which were long-term models (based on long-term 1992-2003 data) and short-term models (based on short-term 1999-2003 data). The long-term models (for signalised junctions only), which consisted of a simple long-term model and a full long-term model, were formulated by including long-term trend in accident occurrences during 1992-2003 period. The short-term models, which were developed by varying the modelling period between 1 year and 5 years, were formulated by considering constant accident rate during the 1999-2003 period.

The original EB method was modified that entailed using the generalised Poisson (GP) distribution, which was shown to be a better fitting model than Poisson distribution, to model the accident count data of an entity (instead of the Poisson distribution used in the original EB method). The EB/MEB method basically estimates an entity's safety by combining the safety of a reference population, as estimated from the accident prediction models, with additional information from the historical accident count data of the entity. The main advantage of MEB over EB method is that the MEB method does not require implicit assumption regarding the

variation of the historical accident count data as the GP distribution can cater to both under-dispersed/over-dispersed data as well as equi-dispersed data whereas EB method requires the accident count data to be Poisson-distributed (equi-dispersed). More importantly, the use of MEB method can reduce the likelihood of making improper decision regarding the hazard level of a site as the actual variability of accident count history is used.

The research has shown that, in general, the use of the MEB with λ estimated from the longest available period (MEB1) method is able to improve the prediction of safety as compared to the EB method. The MEB1 method performed better than the EB method in which the MEB1 method gave relatively more sites (2.0, 7.2, 8.9, 3.3, 25.3, 0.4 percentage points more, respectively), for X-signalised (based on long-term model), T-signalised (based on long-term model), X-signalised (based on short-term model), T-signalised (based on short-term model), X-unsignalised, T-unsignalised junctions, respectively, with smaller individual square of error (SE) values. Moreover, the MEB1 method gave the smallest sum-squares of error (SSE) and root-mean square error ($RMSE$) between predicted and observed accident frequencies in 2004-2006 (for all types of junctions) than the original EB method and also than the other methods.

A procedure for identification and ranking of hazardous sites based on the MEB method was established. When the method was applied to the study junctions, 59 hazardous junctions were identified and then ranked according to three criteria, which were MEB safety estimate ($E(m|K)$), potential for safety improvement (PSI) and level of hazard (H). Evaluation on the three criteria revealed that ranking based on the $E(m|K)$ and PSI criteria were preferred.

LIST OF TABLES

Table 3.1	Junction sample	40
Table 3.2a	Summary statistics of annual accident counts at X-signalised junctions	43
Table 3.2b	Summary statistics of annual accident counts at X-unsignalised junctions	43
Table 3.2c	Summary statistics of annual accident counts at T-signalised junctions	44
Table 3.2d	Summary statistics of annual accident counts at T-unsignalised junctions	44
Table 3.3a	Annual number of accidents by severity and nature of collision at X-signalised junctions	45
Table 3.3b	Annual number of accidents by severity and nature of collision at X-unsignalised junctions	46
Table 3.3c	Annual number of accidents by severity and nature of collision at T-signalised junctions	47
Table 3.3d	Annual number of accidents by severity and nature of collision at T-unsignalised junctions	48
Table 3.4	Accident frequency in study area	51
Table 3.5	Significance (P-value) of trend coefficient estimates	56
Table 3.6	Geometric and operational variables	57
Table 3.7	Summary statistics of geometric/operational variables (1)	58
Table 3.8	Summary statistics of geometric/operational variables (2)	59
Table 3.9	Annual growth rate of traffic	63
Table 3.10	Functional classification of unsignalised junctions	64
Table 3.11	Number of signalised junctions in vicinity of unsignalised junctions	64
Table 4.1	Number of samples for checking the probability distribution of accident count data	77
Table 4.2	Results of goodness-of-fit tests to Poisson distribution	79
Table 4.3	Pattern of rejection to be Poisson distributed on rejected samples	79
Table 4.4	Trends examination	80
Table 4.5	Result of goodness-of-fit test to GP distribution	84
Table 4.6	Comparison of overall SSE and χ^2 statistics of Poisson and GP distributions	85

Table 4.7	Comparison of individual SSE and χ^2 statistics of Poisson and GP distributions	86
Table 5.1	Explanatory variables for signalised junctions	106
Table 5.2	Explanatory variables for unsignalised junctions	107
Table 6.1	Long term accident-exposure models for X-signalised junctions	116
Table 6.2	Full long-term accident prediction model for X-signalised junctions	120
Table 6.3	Simple long-term accident prediction model for X-signalised junctions	120
Table 6.4	Long-term accident prediction model for T-signalised junctions	123
Table 6.5	Full long-term accident prediction model for T-signalised junctions	126
Table 6.6	Simple long-term accident prediction model for T-signalised junctions	127
Table 6.7	Short-term (5-year) accident-exposure models for X-signalised junctions	130
Table 6.8	Short-term (5-year) accident prediction model for X-signalised junctions	132
Table 6.9	Short-term (4-year) accident prediction model for X-signalised junctions	133
Table 6.10	Short-term (3-year) accident prediction model for X-signalised junctions	134
Table 6.11	Short-term (2-year) accident prediction model for X-signalised junctions	134
Table 6.12	Short-term (1-year) accident prediction model for X-signalised junctions	135
Table 6.13	Short-term (5-year) accident-exposure models for T-signalised junctions	137
Table 6.14	Short-term (5-year) accident prediction model for T-signalised junctions	139
Table 6.15	Short-term (4-year) accident prediction model for T-signalised junctions	140
Table 6.16	Short-term (3-year) accident prediction model for T-signalised junctions	141
Table 6.17	Short-term (2-year) accident prediction model for T-signalised junctions	141

Table 6.18	Short-term (1-year) accident prediction model for T-signalised junctions	142
Table 6.19	Short-term (5-year) accident prediction model for X-unsignalised junctions	145
Table 6.20	Short-term (4-year) accident prediction model for X-unsignalised junctions	147
Table 6.21	Short-term (3-year) accident prediction model for X-unsignalised junctions	147
Table 6.22	Short-term (5-year) accident prediction model for T-unsignalised junctions	150
Table 6.23	Short-term (4-year) accident prediction model for T-unsignalised junctions	151
Table 6.24	Short-term (3-year) accident prediction model for T-unsignalised junctions	152
Table 6.25	Short-term (2-year) accident prediction model for T-unsignalised junctions	152
Table 6.26	Short-term (1-year) accident prediction model for T-unsignalised junctions	153
Table 6.27	Short-term accident prediction model (with trend) for T-unsignalised junctions	155
Table 6.28	Performance of X-signalised junctions models	157
Table 6.29	Performance of T-signalised junctions models	157
Table 6.30	Performance of X-unsignalised junctions models	158
Table 6.31	Performance of T-unsignalised junctions models	158
Table 6.32	Sensitivity analysis of short-term X-signalised junctions model	168
Table 6.33	Sensitivity analysis of short-term T-signalised junctions model	168
Table 6.34	Sensitivity analysis of long-term X-signalised junctions model	168
Table 6.35	Sensitivity analysis of long-term T-signalised junctions model	168
Table 6.36	Accident severities at signalised junctions	175
Table 6.37	Accident severities at unsignalised junctions	176
Table 6.38	Summary of accident prediction models	179
Table 7.1	EB and MEB safety estimation over 1999-2003 period	203
Table 7.2	EB and MEB safety prediction for 2004-2006 period	204

Table 7.3	Data points with MEB estimation (based on short-term models) problems	205
Table 7.4	Comparison of methods using <i>SSE</i> and <i>RMSE</i>	211
Table 7.5	Performance comparison of EB and MEB1 methods	212
Table 7.6	Performance comparison of EB and MEB2 methods	212
Table 8.1	List of hazardous X-signalised junctions identified by MEB method	223
Table 8.2	List of hazardous T-signalised junctions identified by MEB method	224
Table 8.3	List of hazardous X-unsignalised junctions identified by MEB method	224
Table 8.4	List of hazardous T-unsignalised junctions identified by MEB method	224
Table 8.5	Ranking of hazardous junctions	227
Table 8.6	Evaluation results for the 5 most hazardous junctions	230
Table 8.7	Evaluation results for the 10 most hazardous junctions	230
Table 8.8	Evaluation results for the 5 most hazardous X-signalised junctions	230
Table 8.9	Evaluation results for the 10 most X-signalised hazardous junctions	230
Table 8.10	Evaluation results for the 5 most hazardous T-signalised junctions	230
Table 8.11	Evaluation results for the 10 most T-signalised hazardous junctions	230
Table 8.12	Evaluation results for the hazardous X-unsignalised junctions	231
Table 8.13	Evaluation results for the 5 most hazardous T-unsignalised junctions	231
Table 8.14	Evaluation results for the 10 most hazardous T-unsignalised junctions	231

LIST OF FIGURES

Figure 1.1	Road traffic accident fatality rates for some developed countries in 2002 and 2003	4
Figure 1.2	Road traffic accident fatality rates of selected cities in 2002 and 2003	5
Figure 1.3	Singapore's road traffic accident fatality rates	6
Figure 1.4	Relationship between traffic safety and personal safety	6
Figure 1.5	Casualty/fatal accident rates	7
Figure 2.1	Classification of accident prediction models	18
Figure 3.1	Study area	41
Figure 3.2	Distribution of accidents at study area by severity during 1999-2006 period	49
Figure 3.3	Distribution of accidents at study area by nature of collision during 1999-2006 period	49
Figure 3.4	Distribution of between moving vehicles accidents at study area according to the type of collision during 1999-2006 period	50
Figure 3.5	Percentage of sites with given number of accidents during 1999-2006 period	50
Figure 3.6	Accident trends at study junctions (1999-2006)	52
Figure 3.7	Accident trends at study junctions (1992-2006)	52
Figure 3.8	Accident trends at signalised junctions in study area (1999-2006)	52
Figure 3.9	Accident trends at signalised junctions in study area (1992-2006)	53
Figure 3.10	Accident trends at unsignalised junctions in study area (1999-2006)	53
Figure 3.11	Accident trends at unsignalised junctions in study area (1992-2005)	53
Figure 3.12	Accident trends over 1999-2006 at different types of study junctions	54
Figure 3.13	Accident trends over 1992-2006 at different types of study junctions	55
Figure 3.14	Junction layout	58
Figure 5.1	Relationship between X-signalised junctions' flow with number of approach lane	103

Figure 5.2	Relationship between T-signalised junctions' flow with number of approach lane	104
Figure 5.3	Modelling framework	109
Figure 5.4	Modelling procedure	110
Figure 6.1	Cook's distances in long-term X-signalised junctions model	118
Figure 6.2	Leverage values in full long-term X-signalised junctions model	118
Figure 6.3	Cook's distances in simple long-term X-signalised junctions model	119
Figure 6.4	Leverage values in simple long-term X-signalised junctions model	119
Figure 6.5	Variance function of full long-term X-signalised junctions model	121
Figure 6.6	Variance function of simple long-term X-signalised junctions model	121
Figure 6.7	Cook's distances in full long-term T-signalised junctions model	124
Figure 6.8	Leverage values in simple long-term T-signalised junctions model	125
Figure 6.9	Cook's distances in simple long-term T-signalised junctions model	125
Figure 6.10	Leverage values in simple long-term T-signalised junctions model	126
Figure 6.11	Variance function of full long-term T-signalised junctions model	127
Figure 6.12	Variance function of simple long-term T-signalised junctions model	128
Figure 6.13	Cook's distances in short-term (5-year) X-signalised junctions model	131
Figure 6.14	Leverage values in short-term (5-year) X-signalised junctions model	132
Figure 6.15	Variance function of short-term (5-year) X-signalised junctions model	133
Figure 6.16	Variance function of short-term (4-year) X-signalised junctions model	136
Figure 6.17	Variance function of short-term (3-year) X-signalised junctions model	136

Figure 6.18	Variance function of short-term (2-year) X-signalised junctions model	136
Figure 6.19	Variance function of short-term (1-year) X-signalised junctions model	136
Figure 6.20	Cook's distances in short-term (5-year) T-signalised junctions model	138
Figure 6.21	Leverage values in short-term (5-year) T-signalised junctions model	139
Figure 6.22	Variance function of short-term (5-year) T-signalised junctions model	140
Figure 6.23	Variance function of short-term (4-year) T-signalised junctions model	143
Figure 6.24	Variance function of short-term (3-year) T-signalised junctions model	143
Figure 6.25	Variance function of short-term (2-year) T-signalised junctions model	143
Figure 6.26	Variance function of short-term (1-year) T-signalised junctions model	144
Figure 6.27	Cook's distances in short-term (5-year) X-unsignalised junctions model	144
Figure 6.28	Leverage values in short-term (5-year) X-unsignalised junctions model	145
Figure 6.29	Variance function of short-term (5-year) X-unsignalised junctions model	146
Figure 6.30	Variance function of short-term (4-year) X-unsignalised junctions model	147
Figure 6.31	Variance function of short-term (3-year) X-unsignalised junctions model	148
Figure 6.32	Cook's distances in T-unsignalised junctions model	149
Figure 6.33	Leverage values in T-unsignalised junctions model	149
Figure 6.34	Variance function of short-term (5-year) T-unsignalised junctions model	150
Figure 6.35	Variance function of short-term (4-year) T-unsignalised junctions model	153
Figure 6.36	Variance function of short-term (3-year) T-unsignalised junctions model	154
Figure 6.37	Variance function of short-term (2-year) T-unsignalised junctions model	154

Figure 6.38	Variance function of short-term (1-year) T-unsignalised junctions model	154
Figure 6.39	Variance function of short-term (5-year) T-unsignalised junctions model with trend	156
Figure 6.40	Performance of the short-term X-signalised junctions models at estimation years	160
Figure 6.41	Performance of the short-term X-signalised junctions models at prediction years	160
Figure 6.42	Performance of the long-term X-signalised junctions models	161
Figure 6.43	Performance of the simple long-term, short-term (5-year) and short-term (3-year) models for X-signalised junctions	162
Figure 6.44	Performance of the short-term T-signalised junctions models at estimation years	163
Figure 6.45	Performance of the short-term T-signalised junctions models at prediction years	163
Figure 6.46	Performance of the full long-term, simple long-term and short-term (5-year) models for T-signalised junctions	164
Figure 6.47	Performance of the short-term models for X-unsignalised junctions	165
Figure 6.48	Performance of the short-term T-unsignalised junctions models without trend at estimation years	166
Figure 6.49	Performance of the short-term T-unsignalised junctions models without trend at prediction years	166
Figure 6.50	Performance of the short-term (5-year) T-unsignalised junctions models with trend and without trend	167
Figure 7.1a	MEB estimates, $E(m)=0.233$ and $Var(m)=0.012$	189
Figure 7.1b	MEB estimates, $E(m)=0.233$ and $Var(m)=0.012$	189
Figure 7.2a	MEB estimates, $E(m)=2.047$ and $Var(m)=0.905$	189
Figure 7.2b	MEB estimates, $E(m)=2.047$ and $Var(m)=0.905$	190
Figure 7.3a	MEB estimates, $E(m)=4.949$ and $Var(m)=2.370$	190
Figure 7.3b	MEB estimates, $E(m)=4.949$ and $Var(m)=2.370$	190
Figure 7.4a	MEB estimates, $E(m)=9.097$ and $Var(m)=8.006$	191
Figure 7.4b	MEB estimates, $E(m)=9.097$ and $Var(m)=8.006$	191
Figure 7.5	MEB estimates, $E(m)=14.191$ and $Var(m)=19.484$	191

Figure 7.6	MEB estimates, $E(m)=24.071$ and $Var(m)=56.058$	192
Figure 7.7	T272, T501 and T577 MEB estimate analysis, $K=2$, $E(m)=0.199$ and $Var(m)=0.042$	206
Figure 7.8	X39 MEB estimate analysis, $K=12$, $E(m)=5.676$ and $Var(m)=13.401$	207
Figure 7.9	T7 MEB estimate analysis, $K=6$, $E(m)=1.860$ and $Var(m)=2.130$	207
Figure 7.10	T233 MEB estimate analysis, $K=4$, $E(m)=0.356$ and $Var(m)=0.136$	208
Figure 7.11	T521 MEB estimate analysis, $K=2$, $E(m)=0.125$ and $Var(m)=0.017$	208

LIST OF SYMBOLS

a	Gamma probability density function parameter
A	Annual number of accidents
AIC	Akaike Information Criterion
b	Gamma probability density function parameter
c	Number of classes or number of cells
$CITY$	Qualitative variable for type of landuse
d	Kolmogorov-Smirnov statistic
D_i	Cook's distance
$E(m)$	Safety of population (mean of m in the reference population)
$E(m K)$	Mean of m in the subpopulation that recorded K accidents
E_x	Expected frequencies of the value x
G^2	Scaled deviance statistic
G_i	Geometric variables
h_{ii}	Leverage statistic
H	Level of hazard
I^p	Cressie-Read (CR) statistic
$JTYPE$	Qualitative variable for junction type
k	Accident prediction model parameter
K	Total number of accidents of an entity during a certain period
L_H	Log-likelihood test ratio statistic
LTF	Left-turn flow
LTP	Left-turn type
m	Expected number of accidents occurring on a group of entities during a specified time period
MED	Qualitative variable for the presence of median
$MERG$	Qualitative variable for presence of downstream merger
n	Number of data points (sample size)
N	Accident sequence duration
$NAPLN$	Quantitative variable for total number of approach lane

<i>NEXLN</i>	Quantitative variable for total number of exit lanes
<i>NOMED</i>	Qualitative variable for the non-presence of median
<i>NOSLR</i>	Qualitative variable for the non-presence of sliproad
<i>NPHASE</i>	Quantitative variable for number of signal phases
O_x	Observed frequencies of the value x
p	Number of parameters that are estimated in a model
<i>PAR</i>	Potential for accident reduction
<i>PEDCROSS</i>	Qualitative variable for the presence of pedestrian crossing
<i>PSI</i>	Potential for safety improvement
q_i	ZIP model parameter
Q	Total junction inflow
Q_l	Sum of left-turning flow from stem and right-turning flow into stem of T-junction
Q_r	Sum of left-turning flow into stem and right-turning flow from stem of T-junction
Q_h	Two-way flow across head of T-junction
Q_{mj}	Major road inflow
Q_{mjt}	Major road inflow in year t
Q_{mn}	Minor road inflow
Q_{mnt}	Minor road inflow in year t
Q_p	Total junction inflow at primary direction
Q_s	Total junction inflow at secondary direction
Q_T	Total junction's inflow (sum of entering vehicles)
r	The largest positive integer for which $\theta + \lambda w > 0$ when $\lambda < 0$ in Generalised Poisson distribution
r_{P_i}	Pearson residual
$r_{SP_i}^2$	Studentized Pearson residual
<i>RLCAM</i>	Qualitative variable for the presence of red light camera
<i>RTLANE</i>	Quantitative variable for number of exclusive right-turn lanes
<i>RTF</i>	Right-turn flow

$\bar{R}$	Average accident rate for all road locations
R_c	Critical accident rate for particular location
s	Sample standard deviation
s^2	Sample variance
$SHLANE$	Quantitative variable for number of shared lanes
SLR	Qualitative variable for the presence of sliproad
$SPLIM$	Qualitative variable for speed limit
$SRLANE$	Number of lanes at sliproad
STF	Straight-thru flow
t	Counter for time (year)
T	Summation of all observation
$TOTLW$	Total leg width (m)
$TOTAPW$	Total approach width (m)
$TOTEXW$	Total exit width (m)
$TOWN$	Qualitative variable for type of landuse
$UTURN$	Qualitative variable for the presence of U-turn
V	Vuong statistic
$Var(m)$	Variance of m in the reference population (group of entities)
$Var(m K)$	Variance of m in the subpopulation that recorded K accidents
VT	Variance test (dispersion index test) statistic
X^2	Pearson statistic
$YBOX$	Qualitative variable for the presence of yellow box
α	Dispersion parameter
$\hat{\alpha}$	Maximum likelihood estimator of α
α^*	Moment estimator of α
α^0	Mean and zero-class frequency estimator of α
β_i	Accident prediction model parameter
δ	Level of statistical significance
ρ	Coefficient of correlation
λ	Generalised Poisson distribution parameter, a measure of departure from Poisson process.

$\hat{\lambda}$	Maximum likelihood estimator of λ
λ^*	Moment estimator of λ
λ^0	Mean and zero-class frequency estimator of λ
θ	Generalised Poisson distribution parameter; an indicator of the intensity of the natural Poisson process
$\hat{\theta}$	Maximum likelihood estimator of θ
θ^*	Moment estimator of θ
θ^0	Mean and zero-class frequency estimator of θ
τ	Zero inflated Poisson model parameter
μ	Mean
$\hat{\mu}$	Maximum likelihood estimator of μ
μ_t	Expected accident number at year t (where $t = 1, 2, 3, \dots$),
γ, γ_i	Accident prediction model parameters
α_1	Accident prediction model parameter
ω	Accident prediction model parameter
σ^2	Variance

CHAPTER 1

INTRODUCTION

1.1 Background

Road safety has nowadays become a major concern everywhere in the world. According to WHO (2004), the number of annual fatalities due to road traffic accidents worldwide is estimated to be almost 1.2 millions while the number of injured people could reach 50 millions. The numbers will continue to rise if insufficient attempts are made to improve road safety. In this regard, the estimation of road safety (in terms of accident frequencies, as used in this thesis) is important as it provides a valuable tool for identifying accident hot-spots so that remedial actions can be undertaken to minimise the unwanted effects of road accidents. Besides, safety estimation knowledge into accident occurrences is also useful for assessing the effect of remediation works.

Road safety is commonly associated with accident occurrences from road traffic activities and their harmful effects. Accidents are essentially multifactor events and their occurrences are a result of breakdown of a system involving inter-twined factors of roadway characteristics, traffic and operational conditions as well as vehicle conditions and human factors. Definitive root “causes” of road accidents are generally regarded as being extremely difficult to determine with certainty. A limited few in-depth studies (Sabey and Staughton, 1975; Treat *et al.*, 1977; Grime, 1987) have pointed to the human factors as the principal “causatory” factors in the majority of accidents. However, given the variable nature of human characteristics and behaviours, it is not easy to study the effect of the human factors on accident occurrences nor are there ready remediation actions that can be applied effectively on the road users. A common accident remediation strategy has thus been to apply engineering measures on the road infrastructure in order to reduce the likelihood of road users making errors (e.g. by having a median to reduce head-on collision) and to make the road environment to be more error-tolerant (e.g. by providing recovery

space for errant vehicles). Hence, road safety can be tackled by studying the associative relationship between accident occurrences and the more controllable accident contributing factors such as roadway characteristics, traffic and operational conditions, in the quest to make the road infrastructure as adaptive (in minimising scope for errors) and as forgiving (be error-tolerant) as possible. Of course, such a strategy, which is the approach adopted in this study, necessarily entails some degree of empiricism that also involves a substantial amount of data assembly for mass data analysis.

Various studies (for example, by Maycock and Hall, 1984; Pickering *et al.*, 1986; Lau and May, 1988; Bonneson and McCoy, 1993; Maher and Summersgill, 1996; Turner and Nicholson, 1998; Kumara and Chin, 2003; Greibe, 2003) have been carried out to investigate the relationship between accident occurrences and roadway characteristics as well as traffic and operational conditions at road junctions and road links and the resulting accident prediction models were then used to estimate road safety.

A recent approach of safety estimation is by using the empirical Bayes (EB) method as proposed by Hauer (Hauer *et al.*, 1989; Hauer, 1992 and 1997). The EB method combines the estimates of safety from multivariate modelling with information of historical accident counts so that a more precise estimation of an entity's safety could be obtained. In the EB method, the historical accident counts for an entity are assumed to be Poisson-distributed but many researchers have questioned the universality of this assumption. For example, Nicholson (1985) found that there are many variations in the accident count data, which often results in the accident counts to be either over-dispersed or are under-dispersed. Therefore, a generalised Poisson (GP) distribution, which can cater to data that are over-dispersed or under-dispersed or are Poisson distributed, is introduced in this research. The EB method is accordingly modified to incorporate the GP distribution.

The present research focuses on safety at road junctions which are the more problematic locations of the road network in terms of safety and capacity

considerations. Statistics show that in Singapore about 40% of road traffic accidents and 35% of road traffic fatalities occur at road junctions. This research entails the tasks pertaining to identification of accident contributory factors and estimation of road safety at the road junctions. Accident prediction models are developed using (non-linear) multivariate modelling techniques, with emphasis on relating accident occurrences with roadway characteristics, traffic and operational factors at the junctions. The results of multivariate modelling are input to the modified empirical Bayes (MEB) method (using GP distribution as an alternative to Poisson distribution) for estimation of road safety. The synthesised models are applied to identify and rank accident-prone locations. It is envisaged that the overall results using the modified techniques can give a better estimation of safety at the road junctions.

1.2 Road safety situation in Singapore

An examination of historical accident data shows that road safety situation in Singapore has remained relatively stable over the period 1988-2006. The safety condition in the last decade and half has notably improved when compared to the early eighties. This can be attributed to various safety-related measures that have been introduced in Singapore over the years which include education and persuasion of road users, strengthening of traffic legislation and enforcement as well as improvements of road/junction design and applications of sophisticated traffic management measures (see Appendix A).

The fatality rates from road traffic accidents in 1999-2006 have remained relatively stable at 4.0 to 5.3 fatalities per 100,000 population. The corresponding rate for 2003 was 5.0 fatalities per 100,000 population which was low compared to the 2003 rates of some Asia Pacific developed countries like Australia (8.2) and New Zealand (11.9). However, the fatality rate per 10,000 vehicles in Singapore was relatively high due to a low degree of motorisation. Singapore's fatality rate in 2003 was 3 fatalities per 10,000 vehicles while the corresponding rates in Australia (2002) and New Zealand (2002) were 1.4 and 1.5, respectively.

Figure 1.1 shows the position of Singapore among a number of countries in terms of road traffic accident fatality rate per 100,000 population and fatality rate per 10,000 vehicles in 2002 and 2003. It should be noted that in Singapore and the countries included in the comparison, a person is recorded as killed if he died within 30 days of the accident. Figure 1.2 shows the position of Singapore among major cities of the world in terms of road traffic accident fatality rate per 100,000 human population in 2003. It can be seen that Singapore's fatality rate was relatively high when compared to major Asian cities like Hong Kong, Macau, Seoul and Tokyo. The position of Singapore's road traffic safety among major countries and cities in Asia Pacific as shown in Figures 1.1 and 1.2 has indicated that there is still more room for Singapore to improve its road safety condition.

Figure 1.1 Road traffic accident fatality rates for some developed countries in 2002 and 2003
Source: TPD (2003 and 2004)

Figure 1.2 Road traffic accident fatality rates of selected cities in 2002 and 2003
Source: TPD (2003 and 2004) and Hong Kong Transport Department (2003 and 2004)

Figures 1.3 and 1.4 present Singapore’s road traffic accident fatality rates, in the form of personal safety rate and traffic safety rate over the 1982 to 2006 period. Personal safety rate, which is associated with the risk of being killed in a road traffic accident (RTA), is obtained by normalising the number of fatal RTA casualties with the human population. Traffic safety rate, which is associated with the safety of road traffic, is obtained by normalising the number of fatal RTA casualties with the vehicle population.

Figure 1.3 indicates that the fatality rates have been on a decreasing trend over the years; the personal safety rate improved from 11.6 fatalities per 100,000 population in 1982 to only 4.2 fatalities per 100,000 population in 2006, while the traffic safety rate improved from 7.3 fatalities per 10,000 vehicles in 1982 to 2.4 fatalities per 10,000 vehicles in 2006. Both rates experienced a decrease of about 64%-67% over this period.

Figure 1.3 Singapore’s road traffic accident fatality rates

Figure 1.4 Relationship between traffic safety and personal safety

Figure 1.4 shows that there were five periods in which the road safety rates can be grouped into relatively stable clusters; these periods were 1982-1984, 1985-1986, 1987-1994, 1995-1998, 1999-2006. The road safety situation in Singapore has basically improved over the years, with the road safety in 1999-2006 being distinctly better than in the early eighties. There was a lengthy period during 1987-1994 when the road safety situation remained relatively unchanged in which traffic safety rates hovered at around 4.2 to 4.7 and personal safety rates of 7.4 to 7.9.

Moreover, though the vehicle population has been increasing from year to year, the number of casualty accidents (accidents with fatality/injury) has been decreasing thereby resulting in a significant decrease in the casualty accident rate (see Figure 1.5). The corresponding rates in 1982 were 23.3 casualty accidents per 1,000 registered vehicles and 6.9 fatal accidents per 10,000 registered vehicles. In 2006, the rates had dropped to 9.7 casualty accidents (59%) per 1,000 registered vehicles and 2.3 fatal accidents (67%) per 10,000 registered vehicles.

Figure 1.5 Casualty/fatal accident rates

On the whole, the above analysis of road safety situation in Singapore has indicated that the period of 1999 to 2005 was a relatively stable period; the casualty accident and the associated fatality rates did not fluctuate widely thus providing a fairly stable period for the present research into accident occurrences at road junctions.

1.3 Research objectives

In general, the research is aimed at developing a sound and reliable methodology that can be used for road safety estimation, with certain emphasis on possible applications in Singapore road junctions. In detail, the objectives of the research are:

1. Setting up a sound procedure for building accident prediction models and developing accident prediction models for road junctions in the study area by a multivariate statistical modelling technique.
2. Examination of validity of Poisson assumption and assessing the application and suitability of generalised Poisson (GP) distribution on (historical) accident count data.
3. Reviewing the empirical Bayes (EB) method for road safety estimation and modification of the EB method by incorporating the GP distribution into the method, constitutes a principal focus of this research.
4. Application of the modified EB (MEB) method for estimating and predicting safety of the study junctions and comparing the results of safety prediction between the EB and MEB methods.
5. Application of the developed accident prediction models and the MEB method for identifying and ranking of hazardous sites, which shall be useful for improving road safety condition, particularly for the case of Singapore road junctions.

1.4 Research scope

The research shall involve:

1. A detailed literature review on issues related to traffic safety at road junctions, safety estimation techniques and modelling techniques.
2. Assembling a large database of accident data and associated road, traffic and operational factors. Data assembly shall entail collecting information from relevant agencies (Traffic Police Department (TPD), Land Transport

Authority (LTA)), and making on-site inspection of each junction covered in the database.

3. Data analysis, which shall include:
 - Profiling road junctions' accident characteristics in Singapore
 - Examining traffic count data captured by GLIDE system (a computerised traffic control and management system) and estimating the annual daily traffic from the GLIDE data
 - Examination of validity of Poisson assumption using historical accident count data
 - Application of generalised Poisson distribution on historical accident count data
4. Setting up procedure for building accident prediction models and applying the procedure for developing accident prediction models, by a multivariate statistical modelling technique, which shall be performed using statistical software packages such as LIMDEP and SAS. Several functional forms shall be examined to find the most appropriate model. The proposed models involve the Poisson, negative binomial (NB), generalised Poisson (GP), zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) regression models. Relevant issues on accident modelling as highlighted in the literature review shall be taken into account in the modelling process.
5. Estimation of safety at road junctions, which shall include the following steps:
 - Reviewing the original EB method for safety estimation
 - Proposing a modification of the method
 - Application of the modified method to estimate safety at study junctions
 - Comparing the results of the modified method with results by the original method
 - Evaluation of the modified method
6. Application of the developed accident prediction models and MEB method to identify and rank hazardous junctions
7. Conclusions and recommendations

1.5 Report structure

The report consists of nine chapters. The first chapter is the introductory part, which describes the background of research, research objectives, scope of research and the report structure. The second chapter deals with a literature review on various issues related to traffic safety at the road junctions as well as modelling and estimation techniques, as relevant to the scope of this research. However, the literature review of certain topics is embedded elsewhere in the chapters dealing specifically with these topics. The third chapter describes the data that have been collected for which the compiled database was subsequently used for developing accident prediction models and road safety estimation. The fourth chapter examines the validity of Poisson assumption in accident counts analysis and the appropriateness of applying generalised Poisson distribution as an alternative to the Poisson distribution. Chapter Five deals with setting up the procedure for the development of accident prediction models through multivariate modelling techniques. The chosen models and factors associated with accident occurrences at road junctions are identified as presented in Chapter Six. Chapter Seven covers the estimation of safety by the EB method as well as the development of the MEB method. The application of the MEB method to identify and rank hazardous junctions is then presented in Chapter Eight. Finally, conclusions regarding this research and recommendations for future research are highlighted in the last chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 General

The literature review comprises five main parts. The first part focuses on various aspects of road safety and traffic accidents. The relationship between road safety and traffic accident is explained in this part, which includes various factors contributing to accident occurrences and the concept of accident exposure. The second part reviews available techniques to formulate accident prediction models as a tool to estimate road safety while the third part describes Bayesian method as an alternative approach to estimate road safety. In the fourth part, various accident prediction models that have been developed to estimate safety at road junctions are reviewed. A conclusion regarding this review and its relevance to the present research are given in the last part.

2.2 Road safety and traffic accident

According to Hauer (1997), ‘road safety is a manifest in the occurrences of accident and their harm’. A broader definition of road safety is ‘the number of traffic accidents (crashes), or accident consequences, by kind and severity, expected to occur on the entity during a specified period’. The entity may be a road, a road section, a junction, etc. These expected numbers change over time. Safety is not to be equated with the fluctuating count of accidents but it should be equated with expected accident frequency as it is a more stable property which is the nature of long-term average. Counts are always non-negative integers; expected numbers can be any non-negative numbers, integers or not. The count of accidents is a reflection of the underlying expected number, a hint that enables us to estimate what the expected number or frequency at some point in time is or was.

Traffic accidents are rare and random events, which means the number of accidents is low when compared to the number of traffic flow through the site and also, one can never be sure when and where the next accident will occur. The number of accident occurrences in a given time period at a site is usually assumed to be Poisson distributed. However, the assumption is open to contention, because the Poisson process requires that the variance and mean of the accident count sequence to be about equal while actual accident count data have been found to be inconsistent with the assumption for some data series (Nicholson, 1985; Ng *et al.*, 1995); it was found that there was too much discrepancy in the variance-to-mean ratio for some sites. To overcome this problem, Nicholson suggested using the binomial distribution if variance-to-mean ratio is distinctly less than unity, and to use the negative binomial distribution if variance-to-mean ratio is distinctly greater than unity.

Road traffic accidents arise from the interaction of exposure and risk. Generally, exposure can be defined as the number of potential situations that accident of a certain type would occur in a certain time and certain area while risk is the probability of an accident occurring given that a potential accident situation has arisen. The number of accidents can be reduced by reducing the accident exposure and also reducing the accident risk.

At road junctions, accident exposure is usually represented by traffic flow while accident risk may be a function of various interactive factors such as human factor (human errors), vehicle factor (vehicle defects) and road environment factor (unfavourable road conditions), and is subject to both temporal and spatial random variations. Studies have shown that the human factor is present in 95% of road traffic accidents, and is a sole contributor in about 65% of the accidents (Grime, 1987). Vehicle defects, alone or in combination, are present in 8% of the accidents, and unfavourable road conditions, alone or in combination, are present in about 28% of the accidents. Although human factors are the principal causatory factors in majority of accidents, however, given the variable nature of human characteristics and behaviours, it is not easy to study the effect of the human factors on accident

occurrences nor are there ready remediation actions that can be applied effectively on the road users. A common accident remediation strategy has thus been to apply engineering measures on the road infrastructure in order to reduce the likelihood of road users making errors (e.g. by having a median to reduce head-on collision) and to make the road environment to be more error-tolerant (e.g. by providing recovery space for errant vehicles).

Some of the accident contributory factors are examined further in the following sections.

a. Human factors and situational factors

A number of studies investigated the influence and association of human factors (driver characteristics such as age, gender and driving experience) and situational factors (such as time of the day when accident happened) with accident occurrences. There was a general agreement in the literature that factors of age, driving experience and gender were correlated with aggregated accident risk of the driver populations (Storie, 1977; Polus *et al.*, 1988; Massie *et al.*, 1995; Stamatiadis and Deacon, 1995; Abdel-Aty and Radwan, 2000). A number of studies showed that night-time travel was more risky than daytime travel for all driver groups and that the risk was disproportionately high for young drivers (Williams, 1985; Massie *et al.*, 1995; Doherty *et al.*, 1998).

b. Vehicle factors

A fair amount of literature exists on the effects of vehicle factors on accident frequency; however, little of it involves statistical modelling technique such as multivariate regression. The most common vehicle factor examined was vehicle age (Joksch, 1971; Joksch and Wuerdemann, 1973; Dutt *et al.*, 1977). Older vehicles were more likely to be involved in accidents resulting from mechanical breakdown. Other vehicle-related factors such as size, weight and model type have been found

to be related to accident frequency; however, their influence on accident severity was more pronounced.

A study of Singapore's road traffic accident characteristics (Wong *et al.*, 2000) revealed that defective brakes and wheels was the major vehicle factor that contributed to accidents, especially in motorcycles, buses, cars, and light goods vehicles. Overloading and insecure loading was most prevalent for heavy goods vehicles. Other vehicle factors found to be associated with accident occurrences were defective vehicle lighting and defective braking system.

c. Road environment and operational factors

Some of the accident-related road environment and operational factors investigated in previous studies, especially those in relation with accidents at road junctions, are discussed in the following paragraphs.

Ogden *et al.* (1994) attempted to relate land use at the junction with accident frequency. It seemed that there was some tendency for safer junctions to be associated with industrial land uses, and for less safe junctions with residential land use. However, no possible cause was offered for the observed pattern. Poch and Mannering (1996) found that road junctions in central business district (CBD) area had a lower likelihood of rear-end accidents. This may be due to signal progression typically provided for through vehicles in the CBD of their study area which reduces the amount of stops and starts vehicles have to make thus decreasing rear-end accident potential. They also found that junctions with some approaches classified as local streets tended to have lower accident frequencies. This is because local streets have the lowest amount of traffic flow in a road network which translates to a lower opportunity for accident to occur. If all approaches of the junctions were local streets, the number of accidents would be even lower.

The effect of the existence of shared movement lanes at a junction was investigated by Poch and Mannering (1996) who found that left-through lane sharing (under right-side driving system as in USA) increased accident potential. One possible explanation is that turning vehicles must stop for gap to manoeuvre through the junction and this causes a high potential for rear-end accidents with straight-through vehicles that approach in the same lane at prevailing speed. Ogden *et al.* (1994) made a comparison (under left-side driving system) between the percentages of accidents occurring at signalised junctions with exclusive right-turn lanes and at signalised junctions with shared right-through lanes, after adjusting for the effect of traffic flow (exposure) on accidents. They plotted the average accident counts per site at 76 signalised junctions during 1987-1991 against total traffic flow entering the junctions. The best fit line and the lines for one accident above/ below the best fit line were plotted. Based on the plots, the junctions were grouped into low, medium and high accident counts. All junctions with average accident counts within one accident above/below the best fit line were categorized as 'medium'. A junction in the 'low' category had fewer accidents than what would be expected on the basis of their traffic flow. A junction in the 'high' category had more accidents than what would be expected on the basis of their traffic flow. They concluded that there was a clear tendency for sites with exclusive right-turn lanes to be safer than those with shared right-through lanes.

Poch and Mannering (1996) also indicated that signal control at a junction had the tendency to decrease total and angle collisions with the exception of eight-phase signalised junctions. When the number of phases was only two, one would expect higher accident numbers as a two-phase system allows permissive filtering of left-turn movements (on right-side driving system), which required vehicles to yield to oncoming traffic. However, it was found that the overall effect of two-phase signalised junctions relative to unsignalised junctions was a lower accident number. On the other hand, signalised junctions with eight signal phases were associated with higher accident numbers, even when compared to unsignalised junctions. This observation is consistent with that made by Taylor (1995). In her study to investigate factors affecting accidents at urban signalised junctions, she found that

complicated signalling arrangement can result in increased number of accidents, in particular pedestrian accidents. However, Golias (1997) concluded that the signalisation of a priority controlled four-arm junction did not always have the same effect on safety; it may decrease the expected number of accidents at the junction, but it may also leave the junction safety unchanged. Poch and Mannering (1996) found that the presence of protected left-turn phasing, which allows vehicle to turn left without conflict, decreased the number of accidents.

More conflict points are often associated with more traffic accidents at junctions. Thus, having fewer conflict points would reduce the possibility of being in a situation that has a higher risk of accident occurrence. This implies higher accident rates can be expected at X-junctions than at T-junctions, as supported by the findings of the Hakkert and Mahalel model (Del Mistro, 1981) and the research of Poch and Mannering (1996).

Another important study on the effect of road environment factors was that of Wong and Nicholson (1992) who observed that there was a strong association between adverse road geometrics and high accident locations. Others like Hammerslag *et al.* (1982), Okamoto and Koshi (1989), Persaud (1991) and Shankar *et al.* (1995) have also found empirical relationships between road traffic accidents and highway geometrics when investigating accident involvement rate, accident probability, safety-related geometric design variables and the accident reduction potential of geometric improvements.

d. Traffic flow

The frequency of an accident occurrence can theoretically be computed from the product of exposure and risk but such an approach has been found to be extremely difficult in practice. Road safety researchers tend to relate accidents directly with only traffic flow in developing accident prediction model, without considering exposure and risk separately, on the basis that accident exposure and accident risk

are functions of traffic flow which imply the number of accidents is also a function of traffic flow.

One common index of accident exposure at the road junctions is the product of conflicting flows. The amount of exposure is estimated from summing over pairs of conflicting movements. Alternative forms of movement interaction include product (or the square root of the product) of approach (or conflicting) flows (Tanner, 1953; Colgate and Tanner, 1967; Leong, 1973; Maycock and Hall, 1984; Pickering *et al.*, 1986; Hauer *et al.*, 1989; Bonneson and McCoy, 1993; Taylor, 1995; Mountain *et al.*, 1998; Greibe, 2003), or simply the sum of the junction entry flows (Jadaan and Nicholson, 1992; Belanger, 1994; Chin and Quddus, 2003). More details regarding the models proposed by these researchers are presented in Section 2.5.

Accident exposure can change from period to period. Such changes will have an influence on safety. Changes in exposure should be taken into consideration especially when one estimates the safety of 'before' and 'after' periods of an entity. In this regard, Hauer (1997) applied correction factors to take account of changes in exposure. If there were no changes in exposure (or other changes in the entity) over the 'before' and 'after' periods (of equal durations), the number of accidents expected in the 'after' period would be the same as that in the 'before' period. When the two periods have the same exposure but of different durations, then the number of accidents expected in the 'after' period is multiplied by the duration ratio of the two periods. If the periods are different and traffic flows have also changed, then the number of expected accidents in the 'after' period should be corrected by the duration ratio of the two periods and traffic flow correction factor. The traffic flow correction factor can be calculated by assuming a safety performance function that relates the expected number of target accidents with traffic flow.

2.3 Accident prediction models

Accident prediction models (APM), as an important tool for estimating road safety, may be formulated based on spatial variation (cross-sectional models), temporal

variation (time-series models), or both (combined cross-sectional/time-series models). Figure 2.1 shows further classifications of the three models.

Figure 2.1 Classification of accident prediction models

Cross-sectional models involve modelling accidents that occur in a group of entities observed at a similar time period. The entity could be a physical object such as a person and a vehicle, or a geographically defined unit such as a road section or a junction. In time series models, the trends of accident counts that occur in an entity over a certain period are examined. The combined cross-sectional/time-series models incorporate the trends of accident counts in the modelling to estimate the

expected accident numbers of an entity. More detailed explanations regarding these models are covered in the following sub-sections.

2.3.1 Cross-sectional models

Cross-sectional models relate accident counts with the varying characteristics of the entities. According to the procedures currently used to build the APM, cross-sectional models can be divided into two categories, which are multivariate models and regression tree models. A multivariate model may be formulated using ordinary linear regression or non-linear regression (generalised linear model). Regression tree models, which use non-parametric techniques, consist of models based on methods called classification and regression tree (CART) and hierarchical tree-based regression (HTBR).

Multivariate modelling methods

Multivariate modelling is the most popular technique that has been used in various safety studies. Different researchers adopted different model forms in their safety studies, which may include ordinary linear regression models and nonlinear regression models. The nonlinear regression models commonly applied include Poisson, negative binomial (NB), zero inflated Poisson (ZIP), and zero inflated negative binomial (ZINB) regression models. The nonlinear models are also known as the family of generalised linear models. Each model has its own advantages and disadvantages with respect to the strength and weaknesses under different circumstances, which are further elaborated in the following paragraphs.

Linear regression models have been widely used in various safety studies (Tanner, 1953; Colgate and Tanner, 1967; Leong, 1973; Lau and May, 1988; Jadaan and Nicholson, 1992; Abbas, 2004). This technique is based on the assumption of errors being normally-distributed with zero mean and of constant variance. In particular, the use of linear regression models has been popular due to its straight-forward approach. However, Jovanis and Chang (1986) among others, stated that ‘the

ordinary linear regression for road safety studies should be used with caution because of problems associated with the non-negativity of the accident frequency and error terms which have unequal variance’.

Furthermore, Maycock and Hall (1984) and also Turner and Nicholson (1998) have concluded that the ordinary linear model was not appropriate for predicting accident frequency. There were various reasons supporting this assertion. First of all, the accident frequency is discrete, and hence does not follow the normal distribution. Furthermore, the variance in the accident frequency is not constant, but tends to increase as the flow increases. The number of accidents also cannot be negative while a normally-distributed error structure implies a substantial probability of a negative number of accidents, especially when the flow is small and the expected number of accidents is also small.

Similarly, Miaou and Lum (1993) evaluated the statistical properties of two conventional linear regression models to determine their suitability for modelling traffic accident and road geometric relationships and they concluded that the linear models lacked the distributional properties to adequately describe the random, discrete, non-negative and sporadic vehicle accidents. Hence, the ordinary linear regression models are not appropriate to make probabilistic statements about accidents on the road, and the parameter estimates derived from these linear models are questionable.

The unsatisfactory properties of the linear regression models have led to the investigation of the Poisson and NB regression models in recent years. For example, Maycock and Hall (1984), Bonneson and McCoy (1993), Miaou (1994), Poch and Mannering (1996), Turner and Nicholson (1998), Berhanu (2004) and Hiselius (2004) have empirically demonstrated the superiority of the Poisson or NB regression over the standard linear regression models in analysing accident frequencies. The Poisson model is appropriate when the mean and the variance of the accident count data are approximately equal. The NB model is appropriate when

the data are over-dispersed, i.e. the variance of the data is significantly greater than the mean.

The Poisson regression model is attractive in that there is only a single parameter to be estimated but it does have its limitations, especially when the variance of the accident data is constrained to be equal to the mean. Accident data were often found to have a variance greater than the mean (a case of over-dispersion), which can result in biased model coefficients and erroneous standard errors. To overcome this problem of over-dispersion, the NB regression model is recommended.

The NB distribution is suitable for describing discrete, non-negative events like frequency of traffic accidents. It has advantage over the Poisson distribution in that it does not require the mean to be equal to the variance. The appropriateness of using NB model over Poisson model can be determined by the statistical significance of an estimated coefficient α which is a parameter related to the degree of over-dispersion. When α is significantly different from zero (as measured by the t -statistic), then the suitability of the NB model is confirmed. Otherwise, the Poisson model would prevail (Poch and Mannering, 1996).

The Poisson regression model can be regarded as a limiting model of the NB regression model as α approaches zero. However, Miaou (1994) stated that although the NB regression model is more general than the Poisson regression model, it requires more extensive computations to estimate model parameters and to generate inferential statistics than the Poisson regression model.

One of the problems arising in traffic accident count analysis is that sometimes there are a large number of entities (e.g. road sections, junctions) that have no reported accident during some observed period of time. For example, in a study by Miaou and Lum (1993), over 80% of the road sections had no reported accidents in a one-year period (though some proportion of these sites had accidents in another period). This situation may lead to an incorrect conclusion that all such entities are safe when, in fact, only some of these zero-accident entities are truly safe (in zero-

accident state) while others are actually unsafe but fortuitously have zero accident count (in a statistical sense) during the period of observation (Shankar *et al.*, 1997; Kumara and Chin, 2003). It should be noted that the zero-accident state may be truly a zero-accident state or an accident state whereby the accidents are not reported; the problem of under-reporting is dependent upon the integrity of the database which, in turn, is related to the data collection system.

The use of standard Poisson and NB regression models do not account for the distinction between truly safe sites and fortuitously zero-accident sites, and thus the estimated model will be biased by over-representation of zero-accident observations in the sample. Such estimation can also incorrectly suggest over-dispersion in the data which indicates that a NB regression model is appropriate when a Poisson regression is correct. To overcome such problems, the ZIP or ZINB regression models should be used.

Thus, when the data have excessive zeros, then the zero inflated regression models are preferable to NB regression model as it allows for the over-dispersion in the data due to excess zeros. However, Miaou (1994) concluded that the interpretation of the zero inflated models could be difficult as accident frequency in zero inflated models is related to other accident-related variables in a much more complicated way than in standard models.

Regression tree methods

The application of regression tree methods in safety study was apparently first reported by Lau and May (1988). They built an injury accident prediction model for signalised junctions using a grouping and classifying technique called classification and regression tree (CART). CART analyses the residuals of the base model and then groups the junctions with similar accident pattern. The CART is basically a three-level procedure with a 'tree' structure for easier interpretation and application. Level I involves generation of a base model, Level II is concerned with analysis of residuals and Level III makes adjustment for individual accident history. Levels I

and II predictions cover group estimations whereas Level III prediction is aimed at the safety estimation of individual junctions.

Stewart (1996) reviewed several applications of the CART technique. Different applications of CART in road safety studies were explored: classification tree models, regression tree models, and using CART to define categorical variables in regression modelling. In the first two applications, CART was used as a standalone model, in the sense that CART provided the desired output directly. Basically, classification tree models are used when the response variable is a discrete (categorical) variable whereas regression tree models are used when the response variable is a continuous variable. The third application involved using CART to generate intermediate results as inputs to other models.

Another tree-based regression method for accident analysis was the one introduced by Karlaftis and Golias (2002) which is called a hierarchical tree-based regression (HTBR) method. A relationship between rural geometric characteristics and accident numbers was developed by this method which was then used to predict road safety. HTBR splits the accident data recursively using optimal splitting rules based on geometric characteristics of the entity to form a tree structure. The HTBR method was applied by Abdel-Aty and Keller (2005) to investigate factors affecting accident severity levels at signalised junctions and by Abdel-Aty *et al.* (2006) to analyse factors affecting the types of accidents at signalised junctions.

The regression tree method, unlike other types of regression models, is a non-parametric technique, therefore the method does not require prior knowledge of the underlying distribution of the accident data. Further, the method is able to handle multicollinearity between variables as well as missing observations. In essence, the tree structure is constructed by splitting the data into nodes based on the values of the independent variables and calculating the mean values of the response variable within the nodes. Once the tree has been built, it can be easily used to predict the expected accident number or to establish a typical combination of the entity's characteristics that is associated with a high number of accidents.

2.3.2 Time series models

The time series models are generally used to study safety at a macro level based on long-term past accident data, for example to study the trend of accident rates at the national level. Two kinds of methods are available, which are those that use simple time-series models and those that use advanced time-series models.

In the simple time series method, safety, usually in the form of accident rates, is estimated directly from historical accident count data and by using linear trend analysis without considering factors that influence safety (El-Sadig *et al.*, 2002; Scuffham and Langley, 2002; Evans, 2003).

The advanced time-series models incorporate explanatory variables in the model besides the historical accident count data. Various researchers for example, Wagenaar (1984), Lassarre (1986), Scott (1986) and Reinfurt *et al.* (1991) have examined the relationships between economic conditions and other explanatory variables and motor vehicle crash involvement by using various time-series models such as ARIMA (Auto Regressive Integrated Moving Average), Box-Jenkins, dynamic regression time-series and structural time series models. Application of this kind of time-series modelling is not very popular in safety studies and it is rarely used nowadays. This may be due to the difficulty in finding significant statistical relationship in advance time-series modelling than conventional regression technique as noted by Wagenaar (1984).

2.3.3 Combined cross-sectional/trend models

Accident counts on an entity vary from year to year due to various factors such as traffic growth, economic growth and the effects of road safety policies and programmes. Combined cross-sectional/trend models incorporate these trends when modelling the accident occurrences and exposure in order to estimate the safety of an entity. Lord and Persaud (2000) classified the methods which incorporate trends

into accident modelling according to the outcome sought, which are marginal models, transition models and random-effects models.

Marginal model is a model in which regression of the response on explanatory variables is modelled separately from within-site correlation. In this model, the regression parameters have the interpretation of population average. In a transition model, past responses are treated as explanatory variables so within-site (temporal) correlation exists in the model. In a random-effects model, there is a natural heterogeneity across entities in the regression coefficients which can be represented by a probability distribution. The regression coefficients thus measure the more direct influence of explanatory variables on the responses in heterogeneous entities. The random-effects model can cater to both between-site and within-site variations. More details regarding these models can be found in Diggle *et al.* (1994).

The common approach is to use the marginal model, which appears to be the most appropriate model for accident estimation. Maher and Summersgill (1996) and Mountain *et al.* (1998) built similar functional forms of the marginal model, but unfortunately the form of the models was not in standard, additive linear form which means the model parameters cannot be estimated directly using the generalised linear modelling technique. Maher and Summersgill (1996) and Mountain *et al.* (1998) proposed an iterative approach using constructed variables to find the estimates of the parameters.

Lord and Persaud (2000) introduced the application of generalised estimating equations (GEE) procedure to overcome the difficulty in developing unbiased marginal models as GEE can handle temporal correlation in disaggregated data. Several accident prediction models were developed by using 6-year data from 868 signalised cross junctions. They concluded that although the annual trend was not statistically significant, the GEE model incorporating the time trend performed better than models that do not accommodate trend. However, this procedure has limitations as it requires each site to have observations for every variable in each year.

The GEE model was also used by Wang *et al.* (2006) and Abdel-Aty and Wang (2006). Wang *et al.* (2006) fitted 3-year accident data using GEE models with NB link function to model temporal correlation of longitudinal accident data while Abdel-Aty and Wang (2006) fitted 2-year spatially correlated accident data using similar model to model spatially correlated data. Four different correlation structures (independent, exchangeable, autoregressive and unstructured) were examined in modelling the temporal correlation of the accident data while only three correlation structures (independent, exchangeable and autoregressive) were examined in modelling the spatial correlation of the accident data. They concluded that the autoregressive correlation structure is a better GEE model to model both temporally and spatially correlated data.

Hauer *et al.* (1994) proposed a multinomial maximum-likelihood function to estimate the parameters of what is classified as a transitional model. The model linked year-to-year change in the expected accident counts to the change in its covariates. Thus, the number of parameters that represented the year-to-year change to be estimated was equal to the number of periods with historical accident counts data.

Shankar *et al.* (1998) and Chin and Quddus (2003) used a random-effects NB model to incorporate temporal variation of accident counts. They used maximum-likelihood method to estimate the model parameters. Based on the results, Shankar *et al.* (1998) postulated that the random-effects NB model was suitable only in the situation that the spatial and temporal effects were totally unobserved; when spatial and temporal effects were explicitly included, the regular NB model was statistically adequate.

2.4 Bayesian methods

Apart from the above accident prediction models, another approach to estimate road safety is by using Bayesian methods, which consist of empirical Bayes (EB) method and hierarchical Bayes (HB) method.

Hauer (1986, 1992, 1997) introduced an EB method to estimate the safety of an entity based on the postulation that the observed count of accidents for an entity is a biased estimate of its expected number of accidents. The method was built under assumptions that accident counts at an entity are Poisson distributed and the expected accident number in the reference population is Gamma distributed. The EB method is related to multivariate modelling method in that the multivariate models are used to estimate the safety of the reference population and their results are used as one of the inputs needed in the EB method.

Mountain *et al.* (1996) showed that the EB method gave the best result in a comparison of several nonlinear accident prediction models. Elvik (1988) reviewed the EB method and highlighted inherent difficulties in defining the reference populations for estimating safety of road sections: “whether a long road section should be treated as a sample of several smaller sections from a population of small sections or as a sample of one section from a population of large sections”. However, the problems do not arise for distinctive entities such as individual junction or driver. Elvik (1988) rationalised that a group of entities cannot form a population unless they are similar among themselves to a minimal degree, which can be shown by the ability of fitting a probability distribution to the accident counts defined for the group of entities. Hence, to overcome the difficulties in defining the reference population, Hauer (1997) suggested using the multivariate modelling method to estimate the safety of the reference population.

More recently, hierarchical Bayesian approaches have been used to tackle various problems in road safety. These methods can handle the uncertainty and the great variability of accident data and produce a probabilistic ranking of accident sites as well as estimate the accident number. For example, Schluter *et al.* (1997) developed a Gamma-Poisson hierarchical Bayesian model to estimate the accident number and then rank and select the accident sites. In their method, informative and non-informative hyperpriors were considered together with guidelines on how to determine the type of prior densities and their parameters.

Tunaru (2002) proposed a multiple response model based on hierarchical Bayesian approach for ranking accident sites based on bivariate Poisson-lognormal distribution. This approach was extended by Brijs *et al.* (2003) by considering a more realistic model for the accident behaviour that takes into account the number of accidents, the number of fatalities and the number of light and severely injured casualties for a given time period for each site. The model was developed using a three-variate Poisson distribution which allows for covariance between the variables. MacNab (2003) also introduced a hierarchical Poisson random effects spatio-temporal model based on a Bayesian hierarchical methodology for modelling and analysing accident and injury surveillance data.

2.5 Safety at road junction

Safety at the road junction has been studied for many years. Various techniques were used by different researchers, but regression analysis has become the most popular technique for many decades. In general, researchers attempted to build a relationship between accident frequency and various indicators of accident exposure by which safety at road junction were estimated. Most studies (for example, Jadaan and Nicholson, 1992; Bonneson and McCoy, 1993; Turner and Nicholson, 1998) were concerned with relating accident frequency to traffic flow while only a few (Maycock and Hall, 1984; Pickering *et al.*, 1987; Taylor, 1995; Poch and Mannering, 1996; Kumara and Chin, 2003) made an attempt to include junction geometrics or operational factors as model variables. However, it seems that their attempts were only limited in the development of the accident prediction model without making any rigorous validation regarding the prediction power of the models.

As reviewed by Jadaan and Nicholson (1992), Tanner (1953) developed a model which related accidents and traffic flows at rural three-legged junctions (T-junctions) in the United Kingdom. The model was of the form:

$$A = 0.0045Q_r^{0.56} Q_h^{0.62} + 0.0075Q_l^{0.36} Q_h^{0.88} \quad (2.1)$$

where A is the annual number of accidents, Q_h the two-way flow across the head of T-junction, Q_r the sum of left-turning flow into stem and right-turning flow from stem of T-junction, Q_l the sum of left-turning flow from stem and right-turning flow into stem of T-junction. Tanner further suggested that the above relationship can be simplified into:

$$A = 0.0045\sqrt{Q_r Q_h} + 0.0075\sqrt{Q_l Q_h} \quad (2.2)$$

The relationship was corroborated by a subsequent study by Colgate and Tanner (1967) and further simplified by Leong (1973) to:

$$A = k(Q_{mj} Q_{mn})^\gamma \quad (2.3)$$

where Q_{mj} and Q_{mn} are major road inflow volumes and minor road inflow volumes respectively, k is a constant and γ is a parameter with values ranging from 0.21 to 0.45.

Pickering *et al.* (1986) attempted to model accidents at rural T-junctions by relating the annual number of accidents to traffic flow and junction geometry. The basic form of the model was:

$$A = k(f(Q))^\gamma \exp(\sum \beta_i G_i) \quad (2.4)$$

where A is the annual number of accidents, $f(Q)$ the relevant flow functions, G_i the geometric variables and k , γ , β_i are parameters to be estimated. They classified accidents into 21 types, which comprised 13 types for accidents occurring within 20 m from the junction (0-20 m accidents) and 8 types for accidents occurring within 20-100 m from the junction (20-100 m accidents). They utilised ten types of flow functions in their models, such as total junction inflow, major road inflow, minor road inflow, crossing flow product, major minor product, etc. One of their best fitting models (significant at confidence level higher than 99.9%) for predicting the annual frequency (A) of 0-20 m accidents as a function of flow alone was:

$$A = 0.24(Q_{mj} Q_{mn})^{0.49} \quad (2.5)$$

where Q_{mj} is major road inflow and Q_{mn} is minor road inflow (measured in thousand of vehicles/day). As similar to the results of Maycock and Hall (1984) who made a study of roundabouts, Pickering *et al.* (1986) found that an increase in traffic flow was associated with an increase in accident frequency and for some types of accidents, the accident frequency was much more closely related to the square root of the product of conflicting flows (following the 'square root law' by Colgate and Tanner (1967)) than to the product or sum of the conflicting flows.

A different approach in developing accident prediction models for signalised junctions was that proposed by Lau and May (1988). They used CART as a building block for developing injury accident models. Macroscopic-type models for injury accidents per year, which consisted of model based on traffic intensity, model based on junction characteristics and model based on information that includes individual accident history, were derived. The factors found to be significant in the models included traffic intensity, proportion of cross street traffic, junction type, signal type, number of lanes on cross streets and main streets, and left turn arrangement (under USA driving system). However, the developed models had relatively low coefficients of correlation, which was only 0.475 for model based on traffic intensity and 0.580 for model based on junction characteristics. No correlation analysis was done for model based on information that included individual accident history as the model represented future accident potential.

Hauer *et al.* (1989) estimated safety at 145 signalised junctions in Toronto based on traffic flows and accident history. They suggested that a logical accident model should relate accident frequency with traffic flows to which the colliding vehicles belong and not to the sum of the entering flows. They categorized accidents into 15 patterns which were based on the movements of the vehicles before the collisions. An equation was given to estimate the expected number of accidents and the variance of estimation using the relevant traffic flows for each of the accident patterns.

Jadaan and Nicholson (1992) established a relationship between road accidents that consisted of link accidents as well as junction accidents, and traffic flows in an urban network. The relationships for junction accidents were investigated using a simple entering flow functions with linear regression models. However, the developed models were not well fitting given that the resulting R^2 values were 0.29 and 0.06 for unsignalised junctions and signalised junctions, respectively.

Bonneson and McCoy (1993) estimated safety at two-way stop-controlled junctions on rural highways by applying the generalised linear modelling approach. They developed the following model for predicting the annual expected accident frequency for two-way stop-controlled junctions on rural highways:

$$E(m) = 0.692 \left(\frac{Q_{mj}}{1000} \right)^{0.256} \left(\frac{Q_{mn}}{1000} \right)^{0.831} \quad (2.6)$$

where $E(m)$ is the expected accident frequency, Q_{mj} the major road inflow and Q_{mn} the minor road inflow. This model suggested that mean accident frequency increased in a nonlinear fashion with increasing major or minor road traffic flows. The ratio of the observed accident count to the predicted accident count indicated a close agreement with each other.

Brude and Larsson (1993) used linear regression technique to build models for predicting accidents at junctions where pedestrians and cyclists are involved. The R^2 values of the models for predicting annual number of pedestrian and cyclist accidents varied between 0.09 to 0.35 and 0.04 to 0.12, respectively, which were quite small. The R^2 values increased to 0.46 and 0.31 for pedestrian and cyclist accidents, respectively, when the models were aggregated over the study period (1983-1988). They concluded that the number of accidents involving unprotected road users (pedestrians and cyclists) per unprotected road user increased with increasing number of motor vehicles but decreased with increasing number of pedestrians and cyclists.

Belanger (1994) used NB regression model to develop accident prediction model at four-legged unsignalised junctions. The EB method was then used to estimate safety of the junctions. Total junction accident models and accident models by pattern were developed for three categories of speed. He also examined the effect of some operational/junction factors which included presence of flashing beacon, sight distance and turning lanes. Various functional forms of traffic flow were investigated and the best model obtained for total junction accidents was:

$$Acc / day = 5.59 \times 10^{-6} (Q_{mj})^{0.42} (Q_{mn})^{0.51} \quad (2.7)$$

where Q_{mj} is the major road inflow and Q_{mn} is the minor road inflow. Belanger (1994) concluded that the total junction models should only be used if the only available data were in the form of total number of accidents and entering vehicles on each approach. However, when accidents by pattern and traffic flows by movements are available, the use of pattern models is preferred because they can provide a detailed identification of abnormal situations.

Taylor (1995) investigated factors affecting accidents at 221 three-legged urban signalised junctions across Great Britain. She concluded that the best flow function for predicting vehicle-only accidents, pedestrian accidents and total accidents at the junctions was the 'vehicle-encounter product' function. This function was the sum of the product of flows in all pairs of crossing, merging and diverging streams. The best pedestrian flow function for predicting both pedestrian accidents and total accidents was a weighted sum of the major road and minor road pedestrian crossing flows. The models indicated that the presence of a pedestrian crossing phase at the junction did not contribute to a reduction in pedestrian accidents at the junctions, but more complicated signalling arrangements resulted in increased pedestrian accidents. Other results from the study also indicated that more complex signalling arrangements and increasing number of exit lanes on minor arm were correlated with more accidents.

Poch and Mannering (1996) investigated interactions between junction geometry and traffic-related elements and accident frequency (for a road network in USA).

Some of the factors that were found to be associated with lower accident frequency at road junctions besides lower traffic flows were signalisation (number of signal phases should be less than eight phases), protected left-turn phasing and presence of local-street approach(es). On the other hand, signalised junctions with eight phases, restricted sight distance, shared left-through lanes and higher speed limit were associated with higher accident frequencies.

Turner and Nicholson (1998) reviewed accident models which relate accident frequency to traffic flows and developed a generalised linear model for predicting individual accident types at junctions in New Zealand. Their sample of junctions comprised priority-X, roundabout-X, signalised-X, and uncontrolled/priority-T types of junctions. Basically, they investigated the relationships between total accidents at the junctions and three forms of traffic flows, namely the sum of junction's inflow (type 1 model), the product of junction's inflow on the approach roads (type 2 model) and the product of conflicting inflow (type 3 model). The type 1 model was built using ordinary linear regression while the other two were built using generalised linear modelling. The type 3 was the principal model while the other two models were used for comparison purposes only. They concluded that the type 3 model was the best model among the three although it did not provide more accurate estimates than type 1 or type 2 models. The conclusion was made based upon factors other than the goodness-of-fit of the model; one of them was the fact that the type 3 model was the only model that enabled estimation of individual accident types. The model indicated that while the number of accidents increased as flow increased, the rate of increase tended to reduce with increasing flow.

Mountain *et al.* (1998) investigated the effect of accident trends over time in the estimates of safety at junctions. Generalised linear modelling was used to develop the accident model which allows for the possibility of accident risk varying over time. The form of the model was:

$$E(m) = \sum \mu_i = \mu_0 \left(\sum \varphi' Q_{mjt}^{\beta_1} Q_{mnt}^{\beta_2} \right) \quad (2.8)$$

where $E(m)$ is the expected accident counts at a junction over the study period (between years t_1 and t_2 , where t_1 and t_2 may vary from junction to junction), Q_{mjt} is the major inflow in year t , Q_{mnt} is the minor inflow in year t and φ , β_1 , β_2 are parameters to be estimated. The models cannot be fitted using standard generalised linear modelling and an iterative approach was used instead. They showed that the accident risk at the study junctions declined annually by 6% and concluded that the inclusion of trend in the model increased the accuracy of accident prediction.

Lord and Persaud (2000) applied GEE procedure to develop accident prediction model that incorporated trend at signalised cross junctions. Six-year data from 868 junctions from CBD and non-CBD area in Toronto-Canada were used. The accident prediction model with trend was compared with that without trend. They concluded that although the trend was not statistically significant but it would still be beneficial to keep the presence of the trend in the model. The conclusion was based on consideration that if the model was applied in a 'before-after' study then the model without trend underestimated accident in the 'before treatment' period while the model with trend overestimated accident in the 'after treatment' period such that it could have underestimated the effectiveness of treatment.

Kumara and Chin (2003) used a ZINB regression model to identify the geometric and traffic factors that influence accident occurrences at 104 signalised T-junctions in Singapore over the 1992-2000 period. They found that accident occurrences were significantly affected by total approach flow and left turn flow, and accident occurrences increased with increasing flow. Moreover, accident occurrences also increased with a greater number of signal phases, the existence of uncontrolled left-turn slip roads, a horizontal curve and a permissive right-turn phase. However, the presence of left-turn lane(s) and median railings and having more than +5% gradient at the approaches were associated with fewer accident occurrences. It should be noted that the developed model contained some explanatory variables with coefficient estimates that were not significant at 95% confidence level.

Greibe (2003) developed accident models individually for X-signalised, X-unsignalised, T-signalised and T-unsignalised junctions since there are considerable differences in accident pattern and risk for different types of junctions. The models were developed using Poisson regression model because of its simplicity and because the values of the model parameters were almost similar with those obtained using a more complicated (negative binomial) regression model. Explanatory variables that were found to be significant in the models included traffic flow and number of lanes. The estimated model was:

$$E(m) = 1.13 \times 10^{-4} Q_p^{0.561} Q_s^{0.47} \exp(\beta_1 NLANE_1 + \beta_2 NLANE_2) \quad (2.9)$$

where $E(m)$ is the expected accident frequency, Q_p the inflow from primary direction, Q_s the inflow from secondary direction, $NLANE_1$ the number of lanes in primary direction, $NLANE_2$ the number of lanes in secondary direction, β_1 (equal to 1.29 if $NLANE_1 = 2$ and equal to 1 if $NLANE_1 > 2$) the estimated coefficient for $NLANE_1$ and β_2 (equal to 0.83 if $NLANE_2 = 2$ or 3 and equal to 1 if $NLANE_2 > 3$) the estimated coefficient for $NLANE_2$. It was found that the above model had a percentage explained variance of 60%, which was only 1% better than a similar model with only traffic flow as the explanatory variable. It was then concluded that simple accident models with only traffic flow as the explanatory variable were found to be almost as well fitting as more complicated models that had other explanatory variables besides traffic flow. Regarding the number of data points that should be included in the models, Greibe (2003) found that the estimated parameter values did not change considerably when more than 100 junctions were used.

Chin and Quddus (2003) used a random-effects NB model to investigate the relationship between accident occurrence and the geometric, traffic and control characteristics of signalised junctions in Singapore. Data from 52 X-signalised junctions over the 1992-1999 period were used. The model treats the data in a time-series cross-section panel aimed at dealing with the spatial and temporal variations in the data. The developed model had sufficient explanatory and predictive power, as indicated by log-likelihood ratio index of 0.318. The results showed that total approach flows, number of phases per cycle, existence of uncontrolled left-turn lane

and presence of surveillance camera significantly affected the safety at the junctions.

Abdel-Aty and Keller (2005) investigated factors affecting accident severity levels at signalised junctions using HTBR method. A 2-year (2000-2001) accident data from 832 junctions in Central Florida, USA were collected. The accident severity levels were categorised into five levels: no-injury, possible injury, non-incapacitating injury, incapacitating injury, fatal injury. Four regression tree models were developed for injury accidents while two models were developed for no-injury accidents (based on type of database: complete or restricted). They found that each model had different significant factors; the number of lanes on minor road, the number of exclusive right turn lanes, the ADT on the minor road and the ADT on the major road were the most significant factors for possible injury, non-incapacitating injury, incapacitating injury and no-injury, respectively.

Abdel-Aty *et al.* (2005) analysed significant factors affecting types of accidents at signalised junctions using HTBR method. They used the same database as in Abdel-Aty and Keller (2005). One of their findings noted that traffic volume along the major road was the most important factor only for predicting right-turn accidents. They concluded that the prediction of expected accident number from aggregated accident models may be inaccurate since different collision types have different variables.

Wang *et al.* (2006) investigated significant factors affecting accidents at signalised cross junctions using GEE models with NB link function. A 3-year accident data from 208 signalised cross junctions in Central Florida was used. Several functional forms of traffic flow were checked by Cumulative Residual method as to find the most appropriate functional form of traffic flow at junction and it was concluded that total traffic inflow per lane was the better form than other common functional forms such as total traffic inflow. They also found that junctions with heavy traffic, larger total number of lanes, larger number of phases, higher speed limits and in high population areas were correlated with higher accident number whereas

junctions with higher number of exclusive right-turn lanes and having partial left-turn protection phase had lower accident number.

In Abdel-Aty and Wang (2006), significant factors affecting accidents at signalised junctions were also examined. A 2-year accident dataset for 476 signalised junctions from 41 corridors in Florida was used. They found that junctions with heavy traffic, larger total number of lanes, shorter signal spacing, larger number of phases and higher speed limits along corridors were correlated with higher accident number whereas junctions having 3 legs, with exclusive right-turn lanes on both roadways, having a protected left-turn phase for left-turning traffic from a corridor and located in open county or residential areas had lower accident number.

2.6 Summary

This chapter has described various works that have been done in traffic safety area, specifically in the estimation of safety at road junctions. The review provides the necessary background information to understand the ‘state of the art’ in the estimation of safety and hence the scope of research potential. The literature review also allows the problem being investigated to be placed in a proper perspective.

The review has pointed out the importance of studying the interaction between accident occurrences and roadway characteristics. Although human factors are found to be present in 95% of road traffic accidents, and are a sole contributor in about 65% of the accidents, it is not easy to study the effect of the human factors on accident occurrences nor are there ready remediation actions that can be applied effectively on the road users given the variable nature of human characteristics and behaviours. Therefore, another way of improving road safety is by studying the relationship between accident occurrences and controllable accident contributing factors such as roadway characteristics, and traffic and operational conditions in order to make the road infrastructure as adaptive (in minimising scope for errors) and as forgiving (be error-tolerant) as possible.

It seems that a popular approach to estimate road safety entity is based on cross-sectional multivariate modelling of accident data by using various modelling techniques. However, if a long period of data is used in building the model, the result of prediction may be poor as accident risk may have changed over the years due to various factors such as traffic growth, economic growth and the effects of road safety policies and programmes. Therefore, one should be vigilant on assessing whether or not to incorporate accident trend in building accident prediction model.

A simple but reliable way to take account of accident trend in accident prediction model is by including trend (time) factor as one of the variables in multivariate modelling. This method, which is capable of handling year to year variation in accident data, has the limitation in that it requires historical accident data and exposure to be available. By and large, one should pay full attention in deciding on the form of trend variable in the model because sometimes the form cannot be solved by standard modelling techniques. In addition, when an inappropriate form of trend is used then the modelling result may not be satisfactory.

The alternative safety estimation technique involves a Bayesian approach that consists of empirical Bayes (EB) method and hierarchical Bayesian method. The EB method has been shown to be able to give a better estimation of safety when compared to multivariate modelling method. Here, the safety estimation is based on mixing two clues on safety: that derived from information about the characteristics of the entity and the accident counts of its reference population, and that from the accident counts of the entity of which its safety is being estimated. The hierarchical Bayesian method, which is a relatively new technique, can handle the uncertainty and the great variability of accident data and produce a probabilistic ranking of accident sites as well as estimate the accident number. However, the technique requires complicated calculations.

From the review, the estimation of safety by the EB method seems promising when combined with the multivariate models that incorporate accident trend. However, the EB method is necessarily dependent upon the assumption used to derive the model. As

noted earlier, Hauer (1997) assumed that the count of accidents K at a site is Poisson-distributed while many researchers have indicated that accident counts need not always be Poisson-distributed. In this regard, the EB method may be enhanced if the underlying variability is better handled say by using a more generalised distribution. This shall form a central focus of this research involving modification of the original EB method for safety estimation. Pertinent details about the development of the modified EB method are described in subsequent chapters.

CHAPTER 3

DATA DESCRIPTION

3.1 Introduction

This chapter describes the data used in the development of selected accident prediction models for estimating traffic safety at the road junctions. The data include information on accidents, traffic flows and geometric/operational characteristics pertaining to junctions (X-signalised/unsignalised and T-signalised/unsignalised) located in the study area.

3.2 Junction sample

In building up the database, on-site visits were made to more than a thousand junctions in the western part of Singapore over the period of 2003 up to mid 2004. At each site, information was captured on the geometric lay-out as well as the operational features. These junctions were also checked against various records to ensure there were no major geometric/operational changes during the part of the study period (1992-2006). Altogether, a total of 1032 junctions were assembled in the final junction sample. The list of the 1032 junctions is presented in Appendix B. Table 3.1 shows the sample distribution of the junctions and Figure 3.1 shows the map of the study area.

Table 3.1 Junction sample

No	Junction Type		Number of sample	Proportion of sample
1	X-junction	Signalised	203	19.6%
		Unsignalised	71	6.9%
	Total X-junction		274	26.5%
2	T-junction	Signalised	201	19.5%
		Unsignalised	557	54.0%
	Total T-junction		758	73.5%
Total			1032	100%

Figure 3.1 Study area
Source: Mighty Minds, 2006

3.3 Road traffic accident data

The road traffic accident data were extracted from a computerised accident database obtained from Singapore Traffic Police Department (TPD). The database is compiled based on accident report forms that are used by a police officer at the scene of accident to record all relevant details regarding the accident event, which include information on the time and location of occurrence, accident severity (fatal, serious injury, slight injury), accident nature (single vehicle, between moving vehicles, between vehicle and stationary object, between vehicle and pedestrian), type of collision (head on, head to side, side swipe same direction, side swipe opposite direction, etc) and also particulars of each of the vehicles, drivers, passengers and pedestrians involved in the accident. The information is captured using three types of accident report form, namely General Data Form, Vehicle-Driver-Passenger Data Form and Pedestrian Data Form. Samples of these accident report forms are included in Appendix C.

It should be noted that the TPD accident records cover only reported fatal and injury accidents. Therefore, any minor accident not incurring any injury to person or involves only damage to property is not reflected in the accident database because such accidents may be settled between the accident parties, and are usually not reported to the TPD.

For the purpose of this research, the accident data were extracted using SAS application software package in order to obtain accident counts pertaining to respective junctions and all the available information related to the accidents. Summary statistics of the accident data from 1992-2006 period are presented in Tables 3.2a – 3.2d and 3.3a – 3.3d.

Table 3.2a Summary statistics of annual accident counts at X-signalised junctions

Statistics of annual accident counts	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Mean	1.617	1.672	1.751	1.591	1.595	1.526	1.406	1.719	1.660	1.547	1.404	1.635	1.660	1.360	1.532
Standard error of mean	0.180	0.150	0.156	0.149	0.137	0.122	0.116	0.151	0.157	0.133	0.145	0.142	0.147	0.126	0.131
Standard deviation	2.420	2.011	2.105	2.033	1.885	1.701	1.625	2.145	2.233	1.894	2.064	2.026	2.089	1.792	1.865
Sample variance	5.858	4.043	4.432	4.135	3.554	2.893	2.640	4.599	4.988	3.586	4.262	4.104	4.364	3.212	3.478
Range	0-16	0-10	0-14	0-10	0-9	0-7	0-9	0-14	0-12	0-10	0-15	0-12	0-13	0-10	0-9
Number of samples	180	180	181	186	190	194	197	203	203	203	203	203	203	203	203
Number of samples with zero accident	69	64	58	72	70	72	71	69	74	78	84	72	75	84	78

Table 3.2b Summary statistics of annual accident counts at X-unsignalised junctions

Statistics of annual accident counts	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Mean	0.197	0.254	0.296	0.296	0.352	0.282	0.296	0.338	0.239	0.310	0.268	0.380	0.352	0.155	0.310
Standard error of mean	0.052	0.072	0.097	0.071	0.102	0.086	0.079	0.102	0.062	0.086	0.085	0.132	0.100	0.059	0.089
Standard deviation	0.435	0.603	0.818	0.595	0.864	0.721	0.663	0.861	0.520	0.729	0.716	1.113	0.847	0.497	0.748
Sample variance	0.189	0.363	0.668	0.354	0.746	0.520	0.440	0.741	0.270	0.531	0.513	1.239	0.717	0.247	0.560
Range	0-2	0-3	0-5	0-3	0-5	0-4	0-3	0-5	0-2	0-4	0-3	0-8	0-5	0-2	0-3
Number of samples	71	71	71	71	71	71	71	71	71	71	71	71	71	71	71
Number of samples with zero accident	58	58	59	54	56	58	56	57	57	56	61	57	56	64	59

Table 3.2c Summary statistics of annual accident counts at T-signalised junctions

Statistics of annual accident counts	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Mean	0.740	0.737	0.744	0.623	0.707	0.783	0.607	0.587	0.682	0.672	0.612	0.537	0.587	0.502	0.711
Standard error of mean	0.089	0.086	0.081	0.066	0.076	0.087	0.071	0.076	0.081	0.086	0.080	0.071	0.071	0.063	0.081
Standard deviation	1.170	1.134	1.073	0.886	1.035	1.194	0.989	1.079	1.144	1.221	1.140	1.005	1.001	0.889	1.143
Sample variance	1.368	1.287	1.151	0.786	1.072	1.426	0.978	1.164	1.308	1.492	1.299	1.010	1.004	0.791	1.306
Range	0-6	0-5	0-5	0-4	0-5	0-7	0-4	0-5	0-5	0-7	0-6	0-6	0-5	0-5	0-7
Number of samples	173	175	176	183	184	189	196	201	201	201	201	201	201	201	201
Number of samples with zero accident	101	105	100	109	106	105	126	136	129	132	133	139	132	138	118

Table 3.2d Summary statistics of annual accident counts at T-unsignalised junctions

Statistics of annual accident counts	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Mean	0.125	0.151	0.132	0.153	0.121	0.139	0.135	0.151	0.135	0.129	0.106	0.095	0.124	0.124	0.144
Standard error of mean	0.018	0.021	0.020	0.019	0.017	0.019	0.019	0.018	0.019	0.016	0.015	0.015	0.017	0.017	0.018
Standard deviation	0.412	0.475	0.457	0.435	0.389	0.444	0.444	0.435	0.451	0.381	0.357	0.355	0.408	0.403	0.425
Sample variance	0.169	0.226	0.209	0.189	0.151	0.197	0.197	0.189	0.203	0.145	0.127	0.126	0.166	0.163	0.181
Range	0-4	0-4	0-4	0-3	0-4	0-3	0-3	0-3	0-4	0-3	0-2	0-3	0-5	0-3	0-4
Number of samples	536	536	536	542	544	547	550	557	557	557	557	557	557	557	557
Number of samples with zero accident	481	473	482	474	486	488	493	489	501	493	507	514	498	501	490

Table 3.3a Annual number of accidents by severity and nature of collision at X-signalised junctions

Year	Number of sites	Total number of accidents	Number of accidents, by severity			Number of accidents, by nature of collision			
			Fatal	Serious injury	Slight injury	Single vehicle	Between moving vehicles	Between vehicle-stationary object	Between vehicle-pedestrian
1992	180	291	8	28	255	11	247	7	26
1993	180	301	9	27	265	10	261	2	28
1994	181	317	10	17	290	20	265	3	29
1995	186	296	11	25	260	14	244	1	36
1996	190	303	4	21	278	17	254	6	27
1997	194	296	14	13	269	10	246	6	34
1998	197	277	11	9	257	7	248	3	19
1999	203	349	9	13	327	10	297	6	36
2000	203	337	5	20	312	14	290	6	29
2001	203	314	6	14	294	12	261	1	38
2002	203	285	5	4	276	8	240	5	33
2003	203	332	12	5	315	20	261	15	35
2004	203	337	7	4	326	14	288	2	33
2005	203	276	7	6	263	15	227	4	30
2006	203	311	6	3	302	9	270	5	27
Total (1999-2006)		2230	2541	57	69	2415	102	2134	44
Total (1992-2006)		4311	4622	124	209	4289	191	3899	72

Table 3.3b Annual number of accidents by severity and nature of collision at X-unsignalised junctions

Year	Number of sites	Total number of accidents	Number of accidents, by severity			Number of accidents, by nature of collision			
			Fatal	Serious injury	Slight injury	Single vehicle	Between moving vehicles	Between vehicle-stationary object	Between vehicle-pedestrian
1992	71	14	1	0	13	3	11	1	0
1993	71	18	0	1	17	1	15	1	1
1994	71	21	1	4	16	0	21	0	0
1995	71	21	2	1	18	0	20	0	0
1996	71	25	1	2	22	0	25	0	0
1997	71	20	1	0	19	2	18	0	0
1998	71	21	1	0	20	1	19	0	1
1999	71	24	0	0	24	1	23	0	0
2000	71	17	0	3	14	0	16	1	0
2001	71	22	0	1	21	0	22	0	0
2002	71	19	1	0	18	0	17	1	1
2003	71	27	0	1	26	0	27	0	0
2004	71	25	0	0	25	2	22	1	0
2005	71	11	0	0	11	1	10	0	0
2006	71	22	0	0	22	1	20	1	0
Total (1999-2006)		145	167	1	5	161	5	157	4
Total (1992-2006)		285	307	8	13	286	12	286	6

Table 3.3c Annual number of accidents by severity and nature of collision at T-signalised junctions

Year	Number of sites	Total number of accidents	Number of accidents, by severity			Number of accidents, by nature of collision			
			Fatal	Serious injury	Slight injury	Single vehicle	Between moving vehicles	Between vehicle-stationary object	Between vehicle-pedestrian
1992	173	128	3	11	114	13	95	4	16
1993	175	129	5	11	113	9	101	5	14
1994	176	131	4	12	115	9	102	8	12
1995	183	114	4	6	104	5	93	2	14
1996	184	130	0	8	122	8	108	1	13
1997	189	148	10	9	129	6	116	5	21
1998	196	119	2	5	112	4	105	0	10
1999	201	118	6	8	104	7	89	8	14
2000	201	137	1	9	127	11	115	2	9
2001	201	135	2	7	126	5	121	1	8
2002	201	123	2	3	118	10	100	1	12
2003	201	108	2	2	104	13	85	3	7
2004	201	118	4	2	112	8	98	1	11
2005	201	101	4	0	97	10	82	3	6
2006	201	143	4	1	138	4	125	2	12
Total (1999-2006)		840	983	25	32	926	68	815	21
Total (1992-2006)		1739	1882	53	94	1735	122	1535	46

Table 3.3d Annual number of accidents by severity and nature of collision at T-unsignalised junctions

Year	Number of sites	Total number of accidents	Number of accidents, by severity			Number of accidents, by nature of collision			
			Fatal	Serious injury	Slight injury	Single vehicle	Between moving vehicles	Between vehicle-stationary object	Between vehicle-pedestrian
1992	536	128	3	11	114	13	95	4	16
1993	536	129	5	11	113	9	101	5	14
1994	536	131	4	12	115	9	102	8	12
1995	542	114	4	6	104	5	93	2	14
1996	544	130	0	8	122	8	108	1	13
1997	547	148	10	9	129	6	116	5	21
1998	550	119	2	5	112	4	105	0	10
1999	557	118	6	8	104	7	89	8	14
2000	557	137	1	9	127	11	115	2	9
2001	557	135	2	7	126	5	121	1	8
2002	557	123	2	3	118	10	100	1	12
2003	557	108	2	2	104	13	85	3	7
2004	557	69	1	2	66	3	58	3	5
2005	557	69	0	0	69	12	51	1	5
2006	557	80	0	0	80	1	71	2	6
Total (1999-2006)		481	561	9	17	535	29	484	14
Total (1992-2006)		999	1079	15	53	1011	62	929	24

Figures 3.2 and 3.3 show the distribution of accidents according to the nature of collision and accident severity for the 1992-2006 period. Figure 3.4 shows the distribution of between moving vehicles accidents according to the type of collision for 1992-2006 period. Figure 3.5 summarises the distribution of sites by the number of accidents per site during the 1992-2006 period, expressed in terms of percentage of sites.

Figure 3.2 Distribution of accidents at study area by severity during 1992-2006 period

Figure 3.3 Distribution of accidents at study area by nature of collision during 1992-2006 period

Figure 3.4 Distribution of between moving vehicles accidents at study area according to the type of collision during 1992-2006 period

Figure 3.5 Percentage of sites with given number of accidents during 1992-2006 period

The above statistics show that around 89% of the accidents that occurred at the study sites during the 1992-2006 period were slight injury accidents, and only a small percentage of the accidents were fatal and serious injury in terms of severity. Around 84% of the accidents were between moving vehicles, 9% were vehicle-pedestrian accidents and the rest were for single vehicle accidents and between vehicle-stationary object accidents. Among the “between moving vehicle” accidents, 67% were head-to-side accidents, 15% were head-to-rear accidents and 4% were head-on accidents. The proportion of sites with zero accident over the 1992-2006 period was the largest for T-unsignalised junctions, followed by X-unsignalised, T-signalised and X-signalised junctions.

The non-inclusion of non-injury accidents in the accident database meant that the observed distributions would differ (by accident types) from the overall proportions for accidents that included non-injury cases. For example, Figure 3.4 indicates that head-to-side accidents were the most frequent accident type at the junctions in the study area, followed by head-to-rear accidents. If non-injury accidents were included the proportion of head-to-rear accidents would increase relative to head-to-side accidents since head-to-rear accidents collisions at junctions would be less damaging than head-to-side collisions.

Table 3.4 shows the frequency of fatal and injury accidents that occurred in the study area during the 1992-2006 period.

Table 3.4 Accident (fatal and injury) frequency in study area

Year	X-signalised	T-signalised	X-unsignalised	T-unsignalised	Total
1992	291	128	14	67	500
1993	301	129	18	81	529
1994	317	131	21	71	540
1995	296	114	21	83	514
1996	303	130	25	66	524
1997	296	148	20	76	540
1998	277	119	21	74	491
1999	349	118	24	84	575
2000	337	137	17	75	566
2001	314	135	22	72	543
2002	285	123	19	59	486
2003	332	108	27	53	520
2004	337	118	25	69	549
2005	276	101	11	69	457
2006	311	143	22	80	556

The trend of accidents is presented in Figures 3.6 to 3.13 for 1999-2006 period and 1992-2006 period with the significance (P-value) of the trend coefficient estimates presented in Table 3.5. It should be noted that the trend analysis for the 1992-2006 period was carried out by removing those sites that could be broadly identified to have been developed within that period.

Figure 3.6 Accident trends at study junctions (1999-2006)

Figure 3.7 Accident trends at study junctions (1992-2006)

Figure 3.8 Accident trends at signalised junctions in study area (1999-2006)

Figure 3.9 Accident trends at signalised junctions in study area (1992-2006)

Figure 3.10 Accident trends at unsignalised junctions in study area (1999-2006)

Figure 3.11 Accident trends at unsignalised junctions in study area (1992-2006)

Figure 3.12 Accident trends over 1999-2006 at different types of study junctions

Figure 3.13 Accident trends over 1992-2006 at different types of study junctions

Table 3.5 Significance (P-value) of trend coefficient estimates

Junction type	1999-2006	1992-2006
X-signalised	0.156	0.723
T-signalised	0.692	0.137
X- and T-signalised	0.199	0.349
X-unsignalised	0.646	0.376
T-unsignalised	0.334	0.334
X- and T-unsignalised	0.264	0.686
All types	0.186	0.361

Figure 3.6 shows that, generally, accident occurrences in the study junctions seem to have been decreasing in the 1999-2006 period. Nevertheless, regression analysis using a power trend showed that the decreasing trend was not significant at 5% level as the P-value of the trend coefficient estimate was 0.186, which was greater than the critical P-value of 0.05. The fitted regression line was also poor, with R^2 equal to 0.2706. The analysis over a longer period (1992-2006) also did not reveal a significant declining trend, as shown in Table 3.5 and Figures 3.7, 3.9, 3.11 and 3.13, which may indicate, at first, that there were no improvements in road safety situation during the period 1992-2006. However, the resulting profiles have not been corrected for exposure and risk. When they were corrected (through inclusion of trends in accident prediction models as presented later in Chapter 6), it turned out that the decreasing trends of the 1992-2003 period were significant at 5% level (for X-signalised and T-signalised junctions).

When the 1999-2006 data were disaggregated according to types of traffic control as shown in Figures 3.8 and 3.10, no distinct declining trend for each type of junction was also observed (with P-value of the trend coefficient estimate = 0.199 and 0.264 for signalised junctions and unsignalised junctions, respectively). When the data were further disaggregated according to the junction types as shown in Figure 3.12, there was also no clear trend observed.

3.4 Junction geometric and operational data

Junction geometric and operational data were obtained from field data collection. Twenty seven types of geometric and operational variables were collected for

signalised junctions while nineteen types of geometric and operational variables were collected for unsignalised junctions. The variables are listed in Table 3.6 while Figure 3.14 gives an illustrated explanation of some of the geometric variables. Summary statistics of the variables are presented in Tables 3.7 and 3.8.

Table 3.6 Geometric and operational variables

No	Geometric/operational variables	Signalised junctions	Unsignalised junctions
1	Total number of approach lanes (NAPLN)	√	√
2	Total number of exit lanes (NEXLN)	√	√
3	Lane configuration (LCON)	√	√
4	Total leg width (m) (TOTLW)	√	√
5	Total approach width (m) (TOTAPW)	√	√
6	Total exit width (m) (TOTEXW)	√	√
7	Number of exclusive right-turn lanes (RTLANE)*	√	√
8	Number of exclusive right-turn lanes with storage lane (RTSTOR)*	√	√
9	Number of exclusive right-turn lanes without storage lane (RTNOSTOR)*	√	√
10	Presence of slip road (SLROAD)	√	
11	Presence of slip road with storage lane (SRSL)	√	
12	Presence of slip road without storage lane (SRNOSL)	√	
13	Number of shared lanes (SHLANE)	√	√
14	Number of shared (left-thru) lanes (SHLTL)	√	√
15	Number of shared (right-thru) lanes (SHRTL)	√	√
16	Number of shared (left-right) lanes (SHLRL)	√	√
17	Number of shared (left-right-thru) lanes (SHLRTL)	√	√
18	Presence of downstream merger (MERG)	√	
19	Presence of median (MED)	√	√
20	Presence of traffic island (ISL)	√	√
21	Presence of U-turn (UTURN)	√	
22	Presence of pedestrian crossing (PEDCROSS)	√	
23	Presence of red light camera (RLCAM)	√	
24	Speed limit (SPLIM)	√	√
25	Presence of yellow box (YBOX)	√	√
26	Number of signal phases (NPHASE)	√	
27	Landuse type (LDUSE)	√	√

* right-turn lanes at minor approach of T-junction was, however, not considered as an exclusive right-turn lane

Figure 3.14 Junction layout

Table 3.7 Summary statistics of geometric/operational variables (1)

No	Geometric/operational variables	X-signalised			T-signalised		
		Range	Mean	Mode	Range	Mean	Mode
1	Total number of lanes (NLANE)	8-35	-	16	6-22	-	12
2	Total number of approach lanes (NAPLN)	4-20	-	10	3-12	-	7
3	Total number of exit lanes (NEXLN)	4-16	-	8	2-11	-	6
4	Total leg width (m) (TOTLW)	31-154	75.7	-	19.2-91.7	53.2	-
5	Total approach width (m) (TOTAPW)	14-84	39.6	-	12.8-42.8	25.3	-
6	Total exit width (m) (TOTEXW)	12.8-60	32.8	-	6.4-50.7	22.9	-
7	Number of exclusive right-turn lanes (RTLANE)	0-8	-	4	0-2	-	1
8	Number of exclusive right-turn lanes with storage lane (RTSTOR)	0-4	-	0	0-2	-	0
9	Number of exclusive right-turn lanes without storage lane (RTSTOR)	0-6	-	0	0-2	-	0
10	Number of shared lanes (SHLANE)	0-8	-	2	0-4	-	1
11	Number of shared left-thru lanes (SHLTL)	0-4	-	0	0-1	-	1
12	Number of shared right-thru lanes (SHRTL)	0-4	-	1	0-2	-	0
13	Number of shared left-right-thru lanes (SHLRTL)	0-1	-	0	0-1	-	0
14	Number of shared left-right lanes (SHLRL)	0-4	-	0	0-1	-	0
15	Number of signal phases (NPHASE)	2-7	-	4	2-4	-	3
No	Geometric/operational variables	X-unsignalised			T-unsignalised		
		Range	Mean	Mode	Range	Mean	Mode
1	Total number of lanes (NLANE)	4-20	-	8	3-18	-	6
2	Total number of approach lanes (NAPLN)	2-10	-	4	2-9	-	3
3	Total number of exit lanes (NEXLN)	2-10	-	4	1-9	-	3
4	Total leg width (m) (TOTLW)	18-77	36.8	20	12-71.1	53.2	18
5	Total approach width (m) (TOTAPW)	10-36.6	39.6	10	6-41.3	25.3	9
6	Total exit width (m) (TOTEXW)	7-34.3	32.8	10	5.4-32.4	22.9	9
7	Number of exclusive right-turn lanes (RTLANE)	0-2	-	0	0-2	-	0
8	Number of exclusive right-turn lanes with storage lanes (RTSTOR)	0-2	-	0	0-2	-	0
9	Number of exclusive right-turn lanes without storage lane (RTSTOR)	0-2	-	0	0-2	-	0

Table 3.8 Summary statistics of geometric/operational variables (2)

No	Geometric/operational variables	Frequency			
		X-signalised	T-signalised	X-unsignalised	T-unsignalised
1	Lane Configuration (LCON)				
	Same number of lanes for all legs	32	25	44	194
	Different number of lanes for one or more legs	171	176	27	363
2	Sliproads (SLROAD)				
	No sliproad for all legs	106	125	71	557
	Sliproads for some legs	34	18	0	0
	Sliproads for all legs	63	58	0	0
3	Sliproads with storage lane (SRSL)				
	No sliproad with storage lane or all legs	120	139	71	557
	Sliproads with storage lane for some legs	42	18	0	0
	Sliproads with storage lane for all legs	41	44	0	0
4	Sliproads without storage lane (SRNOSL)				
	No sliproad with storage lane or all legs	162	179	71	557
	Sliproads with storage lane for some legs	35	16	0	0
	Sliproads with storage lane for all legs	6	6	0	0
5	Downstream Mergers (MERG)				
	No downstream merger for all legs	182	192	71	557
	Downstream merger for at least one leg	21	9	0	0
6	Medians (MED)				
	No median for all legs	41	39	55	440
	Medians for some legs	77	80	16	108
	Medians for all legs	85	82	0	9
7	Traffic Island (ISL)				
	Absent	201	192	68	528
	Present	2	9	3	29
8	U-turns (UTURN)				
	No U-turn for all legs	152	175	71	557
	U-turn for at least one leg	51	26	0	0
9	Pedestrian Crossings (PEDCROSS)				
	No pedestrian crossing or exist only for some legs	37	111	71	557
	Pedestrian crossings for all legs	166	90	0	0
10	Red Light Camera (RLCAM)				
	No camera for all legs	175	181	71	557
	Camera for at least one leg	28	20	0	0
11	Speed Limit (SPLIM)				
	Same speed limit for all legs	190	191	70	472
	Different speed limits for one or more legs	13	10	1	25
12	Yellow Box (YBOX)				
	Absent	83	132	71	381
	Present	120	69	0	176
13	Type of Land Use (LDUSE)				
	City area	66	22	9	89
	Town area	108	143	11	191
	Other area	29	36	51	277

* only 1 X-signalised and T-signalised junctions do not have pedestrian crossing at all, while all unsignalised junctions do not have pedestrian crossing

The variables listed in Tables 3.7 and 3.8 serve to show the scope of the database from which suitable variables shall be selected as explanatory variables in developing accident prediction models. The set of possible explanatory variables

may have strong correlation among themselves which shall thus require additional analysis to choose the ‘best’ combination of variables for modelling. The topic of variable selection shall be explored later in Chapter 5.

3.5 Traffic flow data

The traffic flow data were obtained from the Green Link Determining System (GLIDE) records. The GLIDE is an intelligent traffic control system that controls traffic signal and manages traffic on Singapore’s road network. Data obtained from GLIDE included junction layout, phase diagram that indicates the phase sequence, and traffic counts at 15-minute intervals for each approach lane of the junction.

Traffic count data obtained from GLIDE, however, have some limitations. Firstly, data are recorded in the form of number of vehicles that passed a detector without classification by vehicle type. Secondly, the GLIDE detectors do not cover left-turn movements along the sliproads. Thirdly, a detector can at times be faulty due to various reasons and there would then be no traffic count data available for the affected lane(s).

Despite the limitations, this research still relied on the GLIDE to obtain information regarding traffic flows at the study junctions as the GLIDE provides the only practical source of traffic flow data on an area-wide basis. It would be overly resource-intensive to manually collect the data at the junctions otherwise. A study by Naing (2004) showed that the average deviation of the GLIDE count data as compared to manual count data was -7.3% (under-counting by GLIDE) when motorcycles were included into the analysis and $+3.3\%$ (over-counting by GLIDE) when motorcycles were excluded from the manual counts. Given the condition, a sensitivity analysis on the impact of the accuracy of GLIDE count data into the accident prediction model results was done, as presented later in Chapter 6. It has to be noted that no data were available for unsignalised junctions.

For this research purpose, the GLIDE data for each signalised junction in the study area were collected for a week during June 2004. The data were first checked to ensure that they were not abnormal; for example, checks were made to ensure that counts for corresponding periods on different weekdays should be fairly similar, and also should not deviate excessively among the different lanes of an approach. The daily average from the week's data at each junction was taken to be its average daily traffic (ADT). If a piece of traffic count data at a junction was found to be not available (due to faulty detector/s) during the period, then the data were collected at another available period.

To correct for unavailability of counts at the sliproads, several models were developed to estimate traffic flows at the sliproads. The models were developed through simple regression analysis using manual traffic count data from Singapore Land Transport Authority (LTA) and from field studies by Chia (2005) and Leong (2005). It was found that left-turn flow from a leg may be influenced by total (entry) flow (total flow without left-turn flow, which is unknown) of the respective leg, total flow attracted to the (left-turn) destination leg and number of lanes at the respective sliproad.

Three models were developed for estimating the left-turn sliproad flows at X-signalised junctions, at middle leg of T-signalised junctions, and at minor leg (stem) of T-signalised junctions, respectively. Firstly, correlation matrices were developed, then variables that were highly correlated with left turn volume were selected to be included in the basic models. Modelling was then carried out using linear regression technique. Variables were selected using a backward elimination procedure whereby the least significant variable (based on the t-value or P-value of its estimated parameter) at 95% confidence level was progressively eliminated one by one from the models. This process was continued until all remaining variables were statistically significant to be retained in the models. The resulting models are:

- Model for estimating left-turn flow of leg ($i = 1$) at X-signalised junction:

$$LTF_i = -446.874 + 0.187 STF_i + 0.164RTF_{i+2} + 0.179STF_{i+3} + 443.226SRLANE \quad (3.1)$$

- Model for estimating left-turn flow from middle leg ($i = 1$) of T-signalised junction:

$$LTF_i = 193.367 + 0.208RTF_{i+2} \quad (3.2)$$

- Model for estimating left-turn flow from minor leg ($i = 1$) of T-signalised junctions:

$$LTF_i = 0.377STF_{i+2} \quad (3.3)$$

where

LTF_i = left-turn flow from leg (i) to leg (i+1)

STF_i = straight-thru flow from leg (i)

STF_{i+2} = straight-thru flow from leg (i+2)

STF_{i+3} = straight-thru flow from leg (i+3)

RTF_i = right-turn flow from leg (i)

RTF_{i+2} = right-turn flow from leg (i+2)

$SRLANE$ = number of lanes at sliproads

$i, i+1, \dots$ = notation of legs (clockwise direction)

Model for estimating left-turn flow of leg ($i = 1$) at X-signalised junction had R^2 equal to 0.617, which implied that 61.7% of the variability in left-turn flow had been explained by the model. The f -test in the analysis of variance indicated that a significant amount of variability was explained, with P-value less than 0.0001 with all model's variables significant at 95% confidence level.

Model for estimating left-turn flow from middle leg ($i = 1$) of T-signalised junction had R^2 equal to 0.348, which implied that only 34.8% of the variability in left-turn

flow had been explained by the model. Nevertheless, the f -test in the analysis of variance indicated that a significant amount of variability was explained, with P-value equal to 0.0001. Based on the f -test result, the model was considered adequate although the R^2 of the model was very low.

Model for estimating left-turn flow from minor leg ($i = 1$) of T-signalised junction had R^2 equal to 0.700, which implied that the model was able to explain 70% of the variability in left-turn flow. The f -test in the analysis of variance indicated that a significant amount of variability was explained, with P-value less than 0.0001. However, the model contains no constant since it was found to be insignificant at 95% confidence level.

Since the traffic flow data were collected only for 2004, estimations were done to obtain the data for other periods. Although it is desirable to find the specific growth rate applicable to each junction there was insufficient resources to calibrate the individual rates. It was assumed that all the junctions in study area had similar annual traffic growth rate corresponding to the annual growth rate of vehicle population in Singapore as shown in Table 3.9.

Table 3.9 Annual growth rate of traffic

Year	No of Vehicles	Growth	Year	No of Vehicles	Growth
1992	557,584		2000	692,807	0.580%
1993	584,322	4.795%	2001	708,370	2.246%
1994	611,611	4.670%	2002	706,956	-0.200%
1995	642,129	4.990%	2003	711,043	0.578%
1996	668,304	4.076%	2004	727,395	2.300%
1997	683,204	2.230%	2005	754,992	3.794%
1998	681,081	-0.311%	2006	799,373	5.878%
1999	688,811	1.135%			

Source: LTA, 2006

As noted earlier, unsignalised junctions are not covered by the GLIDE system. Also, no other traffic flow data are available for the unsignalised junctions. Therefore, the traffic flow data for unsignalised junctions can only be obtained either through field survey or by estimation. It is, however, overly resource intensive to collect the traffic data directly by field survey.

Another way to get the traffic flow data for the unsignalised junctions is by estimating the flow using known flow data at nearby signalised junctions. However, as is often the case, most unsignalised junctions are located in the (low end) collector/access network, and most of these junctions are therefore surrounded by unsignalised junctions too. Table 3.10 shows that 73% of the unsignalised junctions in the study area are located at the collector/access network while Table 3.11 shows that the majority (87%) of the junctions are surrounded by other unsignalised junctions (58%) or linked to at most one signalised junction (29%).

Table 3.10 Functional classification of unsignalised junctions

Functional classification of junction's legs	X-unsignalised	T-unsignalised	Total	
P-P	0	0	0	0%
P-S or P-S-L	1	10	11	2%
P-L	4	155	159	25%
S-S or S-S-S-L	7	15	22	4%
S-L or L-S	10	143	153	24%
L-L	49	234	283	45%
Total	71	557	628	100%

Notes: P – primary (arterial) road ; S – secondary (collector) road ; L – local (access) road

Table 3.11 Number of signalised junctions in the vicinity of unsignalised junctions

Number of signalised junctions in vicinity	X-unsignalised	T-unsignalised	Total	
0	55	309	364	58%
1	11	172	183	29%
2	5	67	72	11%
3	0	9	9	1%
4	0	-	0	0%
Total	71	557	628	100%

Under the prevailing situation, the estimation of traffic flow using available flow data from nearest signalised junction(s) would be highly inaccurate. Moreover, the estimation procedures require the assumption that the amount of traffic generated or attracted along any given link between two adjacent junctions to be equal. Therefore, it is generally difficult to obtain a reasonable estimate of the traffic flow data at unsignalised junctions.

Another estimation technique is through traffic modelling at network level. Nevertheless, this technique requires a large amount of data such as geometric data of roads and junctions and origin-destination (OD) data. A comprehensive OD data for Singapore road network is presently not available which makes this technique not feasible for this research.

It is therefore very hard to obtain reliable traffic flow data for unsignalised junctions. No further attempt was thus made to obtain the traffic flow data for unsignalised junctions. However, due to the importance of the accident prediction model in assessing safety performance of junctions, the models for unsignalised junctions shall continue to be developed despite the limitations. Moreover, the majority of the junctions in Singapore are unsignalised. Thus, the accident prediction models for unsignalised junctions were developed without the traffic variable, which shall be further elaborated in Chapter 5.

3.6 Summary

This chapter describes mainly the accident data, geometric and operational characteristics and traffic flow data of the study junctions. The data for subsequent analysis and modelling were basically derived from the database described in this chapter.

Essentially, the accident data were extracted from TPD accident database for the 1992-2006 period. Several statistics on the accident data were then presented. A number of geometric and operational characteristics of the study junctions were collected through site visits and summary statistics were produced. Traffic flow data for signalised junctions were extracted from GLIDE records. However, the GLIDE does not capture the traffic flow data at the sliproads, and models were used to estimate the left-turn flows at signalised junctions. No traffic flow data were available for unsignalised junctions.

CHAPTER 4

PROBABILITY DISTRIBUTION OF ACCIDENT COUNTS

4.1 Introduction

Accident counts at a road entity are usually assumed to be Poisson-distributed. The validity of the Poisson assumption is examined in this chapter. Following this, a generalised Poisson distribution as an alternative distribution for the accident counts is introduced. A comparison of these two distributions is presented afterwards to determine their suitability for accident count analysis.

4.2 Background theory

It is common in accident analysis to assume that the occurrence of road accidents at a specific site in a given time period is governed by a Poisson process as accidents are rare and random events, and the accident counts are Poisson-distributed. Furthermore, the accident occurrence may be governed by a stationary Poisson process or by a non-stationary Poisson process. In a stationary Poisson process, the underlying true accident rate is constant over time while in a non-stationary Poisson process the underlying true accident rate is changing over time. Nevertheless, the underlying true accident rate (UTAR) for a site is not known with certainty; the UTAR is not the same as the observed accident rate but can be estimated from the observed accident counts.

The principle of complete randomness is not always applicable in accident analysis. Nicholson (1986) found that accident counts are not always random when analyzing the pattern of accident occurrences. Non-randomness in accident counts may come from a secular trend, a cyclic variation or a discontinuity in the data. The runs test, a non-parametric test often used for checking randomness, may be used to assess whether the observed accident counts are produced by a random process. Even then,

accident count sequence produced by a random process need not automatically be Poisson-distributed.

A number of researchers (Nicholson, 1985; Nicholson and Wong, 1993; Ng *et al.* 1995; Thomas, 1996) have shown that accident counts at individual sites are not always Poisson-distributed; these findings of non-constancy in the Poisson assumptions have been noted in a number of subsequent studies by latter authors. Many variations exist in the accident data, which results in the accident data being over-dispersed or under-dispersed. Nicholson (1985) recommended choosing one of several distributions depending upon the variability of the accident data indicated by a dispersion index defined as variance over mean ratio. Binomial distribution is to be used if the dispersion index is less than unity, Poisson distribution if the dispersion index is not significantly different from unity and negative binomial (NB) distribution if the dispersion index is greater than unity.

Rather than having a multitude of statistical distributions, it would be desirable to have a single probability distribution which can accommodate the variability inherent in the accident data. Thus, the generalised Poisson (GP) distribution, which is a generalisation of the Poisson distribution, is introduced as it has the advantage that it can cater to data that are more dispersed, or less dispersed, than required by the Poisson distribution.

A GP distribution probability function as defined by Consul (1989) is given by:

$$f(x|\theta, \lambda) = \frac{\theta(\theta + \lambda x)^{x-1} \exp(-(\theta + \lambda x))}{x!} \quad \text{for } x = 0, 1, 2, \dots \quad (4.1)$$

$$= 0 \quad \text{for } x > w, \text{ when } \lambda < 0$$

and zero otherwise, where:

$$\theta > 0$$

$$\max(-1, -\theta/w) < \lambda \leq 1$$

w , which is ≥ 4 , is the largest positive integer for which $\theta + \lambda w > 0$ when $\lambda < 0$

The mean (μ) and variance (σ^2) of the above GP distribution are:

$$\mu = \frac{\theta}{(1-\lambda)} \quad (4.2)$$

$$\sigma^2 = \frac{\theta}{(1-\lambda)^3} \quad (4.3)$$

The GP distribution has two parameters: θ and λ ; they are called the first and second parameter of GP distribution. The parameter λ need not be related to the parameter θ . The lower limit of λ is imposed to ensure that there are at least five classes with nonzero probability when λ is negative.

The parameter θ is an indicator of the intensity of the natural Poisson process, while the parameter λ may be regarded as a measure of departure from Poisson process. The parameter λ is the average rate of the restitution process of the subjects under study. A positive value of λ denotes that the subjects are predisposed towards accelerating the natural Poisson process and a negative value of λ indicates the reverse of retarding the process.

A restricted form of GP distribution can be obtained when it is known in advance that the second parameter λ is linearly proportional to the parameter θ . By putting $\lambda = \alpha\theta$ in the GP distribution of Equation (4.1), the probability distribution for the restricted GP distribution then becomes:

$$f(x|\theta, \alpha\theta) = \frac{(1 + \alpha\theta)^{x-1} \theta^x \exp(-\theta(1 + \alpha\theta))}{x!} \quad \text{for } x = 0, 1, 2, \dots \quad (4.4)$$

and zero otherwise.

Furthermore, Equation (4.2) can be rewritten as:

$$\mu = \frac{\theta}{1 - \alpha\theta} \quad (4.5)$$

$$\text{or} \quad \theta = \frac{\mu}{1 + \alpha\mu} \quad (4.6)$$

Substituting Equation (4.6) into Equation (4.4) yields:

$$f(x|\mu, \alpha) = \left(\frac{\mu}{1 + \alpha\mu} \right)^x \frac{(1 + \alpha\mu)^{x-1}}{x!} \exp\left(\frac{-\mu(1 + \alpha\mu)}{1 + \alpha\mu} \right) \quad (4.7)$$

where $-1/\mu < \alpha < 1$.

Equation (4.7) is a more convenient parameterization of the GP distribution with mean μ and dispersion parameter α (Wang and Famoye, 1997). The variance of the GP distribution then becomes:

$$\sigma^2 = \mu(1 + \alpha\mu)^2 \quad (4.8)$$

The GP distribution possesses the twin properties of over-dispersion and under-dispersion as $\alpha > 0$ or $\alpha < 0$, respectively. When the value of α is zero, the GP distribution becomes the standard Poisson distribution:

$$f(x|\mu) = \frac{\mu^x \exp(-\mu)}{x!} \quad (4.9)$$

with mean μ equal to variance σ^2 .

The parameter μ of the Poisson distribution can be estimated through maximum-likelihood estimator as:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \quad (4.10)$$

The moment estimator of the parameter μ of the Poisson distribution will yield the same result.

The GP distribution parameters can be estimated through maximum-likelihood (ML) estimator, moment estimator, and mean and zero frequency estimator. They all estimate the GP distribution parameters somewhat differently. The properties and complete derivations of the estimators can be found in Consul (1989) and a summary is presented in the following sections.

Maximum-likelihood (ML) estimator

Assuming that the observed frequencies in a random sample of size n from the GP distribution model $f(x|\theta, \lambda)$ as defined in Equation (4.1) are O_x , $x = 0, 1, 2, \dots, v$

for the different classes so that $\sum_{x=0}^v n_x = n$, where v is the largest of the observation.

The ML estimator of $\hat{\theta}$ is then:

$$\hat{\theta} = \bar{x}(1 - \hat{\lambda}) \quad (4.11)$$

where

$$\bar{x} = \frac{1}{n} \sum_{x=0}^v x n_x \quad (4.12)$$

The ML estimator of $\hat{\lambda}$ is given by the root of the equation:

$$\sum_{x=0}^v n_x \frac{x(x-1)}{\bar{x} + (x - \bar{x})\lambda} - n\bar{x} = 0 \quad (4.13)$$

Equation (4.13), which will have multiple roots for λ , does not give a value for λ when $v = 0$ or $v = 1$. Therefore, the ML estimators $\hat{\theta}$ and $\hat{\lambda}$ do not exist when $v = 0$ or $v = 1$.

For a random sample of $x_1, x_2, \dots, x_n$, the ML estimator $\hat{\alpha}$ of the GP distribution $f(x|\theta, \alpha\theta)$ as defined in Equation (4.4) is:

$$\hat{\alpha} = \frac{1}{\hat{\theta}} - \frac{1}{\bar{x}} \quad (4.14)$$

where $\hat{\theta}$ is the root of the Equation:

$$\sum_{i=1}^n \frac{x_i(x_i - 1)}{\theta(\bar{x} - x_i) + x_i\bar{x}} - n = 0 \quad (4.15)$$

The root of Equation (4.15) does not exist when all x_i 's are equal.

Moment estimator

The moment estimators θ^* and λ^* for the GP distribution model $f(x|\theta, \lambda)$ as defined in Equation (4.1) are:

$$\theta^* = \sqrt{\frac{\bar{x}^3}{s^2}} \quad (4.16)$$

$$\lambda^* = 1 - \sqrt{\frac{\bar{x}}{s^2}} \quad (4.17)$$

where $\bar{x}$ and s^2 are the sample mean and sample variance.

The moment estimators θ^* and α^* for the GP distribution model $f(x|\theta, \alpha\theta)$ as defined in Equation (4.4) are:

$$\theta^* = \sqrt{\frac{\bar{x}^3}{s^2}} \quad (4.18)$$

$$\alpha^* = \sqrt{\frac{s^2}{\bar{x}^3}} - \frac{1}{\bar{x}} \quad (4.19)$$

where $\bar{x}$ and s^2 are the sample mean and sample variance.

Mean and zero frequency estimator

This estimator can be used when the frequency for the zero class in the sample is larger than most other class frequencies. The mean and zero frequency estimators θ^0 and λ^0 for the GP distribution model $f(x|\theta, \lambda)$ as defined in Equation (4.1) are:

$$\theta^0 = \ln\left(\frac{n_0}{n}\right) \quad (4.20)$$

$$\lambda^0 = 1 - \frac{\theta^0}{\bar{x}} \quad (4.21)$$

The mean and zero frequency estimators θ^0 and α^0 for the GP distribution model $f(x|\theta, \alpha\theta)$ as defined in Equation (4.4) are:

$$\theta^0 = \ln\left(\frac{n_0}{n}\right) \quad (4.22)$$

$$\alpha^0 = \frac{1}{\theta^0} - \frac{1}{\bar{x}} \quad (4.23)$$

Consul (1989) suggested that moment estimator shall be used when $|\lambda| < 0.5$ and $\theta > 2$, but not when $(\alpha \cdot \theta) \approx 1$ whereas mean and zero frequency estimator is appropriate for small values of θ . ML estimator will not exist when all x_i 's are equal.

4.3 Review of goodness-of-fit tests for discrete distributions

4.3.1 General goodness-of-fit tests

There are several goodness-of-fit tests that can be used for discrete distributions. The most well-known one is the χ^2 test which uses the test statistic:

$$\chi^2 = \sum_{x=0}^k \frac{(O_x - E_x)^2}{E_x} \quad (4.24)$$

where O_x and E_x are the observed and expected frequencies of the random variable X . The test statistic approximates a χ^2 distribution with $(c - p - 1)$ degrees of freedom where c is the number of classes and p is the number of parameters being estimated. For the case of Poisson distribution, the degrees of freedom using this test statistic is the number of classes subtracted by two since there is only one parameter to be estimated. The calculated test statistic is compared to critical χ^2 value with respective degrees of freedom in order to assess whether the data come from the hypothesized distribution.

Although the χ^2 test statistic is easy to compute and can be used for testing a variety of distributions, it has the limitation that the statistic does not approximate a χ^2 distribution if the expected cell frequencies determined by the hypothesized

distribution are small such that the test statistic tends to be larger than the theoretical chi-square value it is supposed to estimate. One would then tend to reject the hypothesized distribution with a probability greater than the selected significance level (when actually it should not be rejected) if the expected cell frequencies are small (Horn, 1977). There are various rules regarding how large the expected cell frequencies should be. The most widely accepted rule is set by Cochran (1954) who stipulated that all expected cell frequencies should be greater than or equal to one and at least 80% of the expected cell frequencies should be at least five.

Cressie and Read (1984) developed a test that can be considered as a generalisation of the χ^2 test. The test statistic is defined as:

$$2nI^\phi = \frac{2}{\phi(\phi+1)} \sum_{i=1}^m O_i \left[\left(\frac{O_i}{E_i} \right)^\phi - 1 \right] \quad (4.25)$$

where n is equal to number of data and c is equal to number of cells. For all values of $\phi \in \mathbb{R}$, where $\mathbb{R}$ is the set of rational numbers, the statistic $2nI^\phi$ follows asymptotically a χ^2 distribution with $(c - p - 1)$ degrees of freedom where c is the number of cells and p is the number of parameters being estimated. For $\phi = 1$, Equation (4.25) reduces to the χ^2 test statistic. Cressie and Read (1984) showed that $\phi = \frac{2}{3}$ gives the best performance as long as $E_i \geq 1$ and $n \geq 10$.

Another test that can be used to assess the goodness-of-fit to discrete distributions is the Kolmogorov-Smirnov (KS) test. The KS test was originally developed for continuous distribution. If the standard KS test was used for discrete distribution then the approximation of p -values would be very conservative. Hence, Conover (1972, 1999) developed an exact KS test for discrete distribution and found that the exact p -values for discrete distributions are often only about one-third as large as their approximations from the standard KS test. The KS statistic for the two-sided test is:

$$d = \underset{\text{all } i}{\text{Max}} |F(x_i) - S(x_i)| \quad (4.26)$$

where

$F(x_i)$ = the theoretical cumulative distribution function specified in the null hypothesis
 $S(x_i)$ = the sample cumulative distribution function for any given observation x_i

The exact p -value of the test is computed as described in Conover (1972, 1999). However, as similar to the standard Kolmogorov-Smirnov test for continuous distribution, the parameters of the discrete distributions being tested must not be estimated from the sample.

4.3.2 Goodness-of-fit tests for Poisson distribution

The test commonly used to test the validity of Poisson assumption, besides the well-known χ^2 test, is the variance test or dispersion index test. The statistic used for this test is:

$$VT = \sum_{i=1}^n \frac{(x_i - x/n)^2}{x/n} = (n-1) \frac{s^2}{\bar{x}} \quad (4.27)$$

with s^2 being the sample variance of x and n is the sample size. The statistic is directly related to the variance-to-mean ratio (dispersion index), multiplied by $(n-1)$. The statistic is approximately χ^2 distributed with $(n-1)$ degrees of freedom. The observed dispersion index is compared to the critical χ^2 value with respective degrees of freedom. The accuracy of the χ^2 approximation improves when the sample size and x/n increase.

Rao and Chakravarti (1956) developed a likelihood test for testing an observed distribution to be Poisson. The test statistic used is:

$$L_H = \sum O_x \ln(O_x x!) \quad (4.28)$$

where O_x is the observed frequency of variate value x , $O = \sum O_x$ and $T = \sum xO_x$. A table of critical value is given for O value between 3 to 100 and T value between 3 to 10 at an exact level of significance ≤ 0.05 . If the observed value of the statistic is greater than or equal to the (tabulated) critical value, the null hypothesis that an observed distribution is Poisson is to be rejected at a level of significance ≤ 0.05 . When the parameters of the sample (i.e. O and T) are outside the critical range, then the following statistic can be used:

$$L_H = 2\left(\sum O_x \ln(O_x x!) + (T - O)\ln O + T - T \ln T\right) \quad (4.29)$$

The statistic is comparable to a χ^2 distribution with $(c-1)$ degrees of freedom, where c is the number of classes.

Campbell and Oprian (1979) developed tables of critical values of the Kolmogorov-Smirnov d statistic for testing the Poisson distribution when the Poisson parameter is unknown and is estimated from the sample. The critical values were developed by using Monte Carlo simulation. The critical values are tabulated for sample size as small as five for levels of significance of 0.20, 0.15, 0.10, 0.05 and 0.01. Campbell and Oprian (1979) concluded that this KS test is as powerful as the χ^2 test with the added advantage of being applicable for small sample sizes.

Nicholson and Wong (1993) constructed an exact statistical test for testing the Poisson assumption of accident count data. The test was derived based on combinatorial analysis and the lower and upper bounds of variance-to-mean ratio at 5% and 10% significance levels were tabulated. Nicholson and Wong suggested that the exact test should be used when N (sequence duration) is less than 21 and K (total number of accidents) is less than 34. For larger values of K and N , the dispersion index test is as accurate as the exact test. Although this test was originally developed by Nicholson and Wong (1993) for application in accident analysis, the application has been extended to other fields.

where O_x is the observed frequency of variate value x , $O = \sum O_x$ and $T = \sum xO_x$. A table of critical value is given for O value between 3 to 100 and T value between 3 to 10 at an exact level of significance ≤ 0.05 . If the observed value of the statistic is greater than or equal to the (tabulated) critical value, the null hypothesis that an observed distribution is Poisson is to be rejected at a level of significance ≤ 0.05 . When the parameters of the sample (i.e. O and T) are outside the critical range, then the following statistic can be used:

$$L_H = 2\left(\sum O_x \ln(O_x x!) + (T - O)\ln O + T - T \ln T\right) \quad (4.29)$$

The statistic is comparable to a χ^2 distribution with $(c - 1)$ degrees of freedom, where c is the number of classes.

Campbell and Oprian (1979) developed tables of critical values of the Kolmogorov-Smirnov d statistic for testing the Poisson distribution when the Poisson parameter is unknown and is estimated from the sample. The critical values were developed by using Monte Carlo simulation. The critical values are tabulated for sample size as small as five for levels of significance of 0.20, 0.15, 0.10, 0.05 and 0.01. Campbell and Oprian (1979) concluded that this KS test is as powerful as the χ^2 test with the added advantage of being applicable for small sample sizes.

Nicholson and Wong (1993) constructed an exact statistical test for testing the Poisson assumption of accident count data. The test was derived based on combinatorial analysis and the lower and upper bounds of variance-to-mean ratio at 5% and 10% significance levels were tabulated. Nicholson and Wong suggested that the exact test should be used when N (sequence duration) is less than 21 and K (total number of accidents) is less than 34. For larger values of K and N , the dispersion index test is as accurate as the exact test. Although this test was originally developed by Nicholson and Wong (1993) for application in accident analysis, the application has been extended to other fields.

4.3.3 Goodness-of-fit tests for generalised Poisson (GP) distribution

Consul (1989) used the χ^2 test to test the goodness-of-fit of the GP distribution, but as stated earlier the test is not appropriate for small sample sizes. There is another test specifically designed to test whether a GP distribution gives an improved fit over a standard Poisson distribution. This test was developed using a score statistic by Rayner and McIntyre (1985). The statistic used is:

$$W_u = \left(\frac{s^2}{\bar{x}} - 1 \right)^2 \frac{n}{2} \quad (4.30)$$

with s^2 being the sample variance of x and n the sample size. The statistic follows approximately a χ^2 distribution with one degree of freedom. However, this approximation is only valid when the mean and sample size are large.

4.4 The data

Sequences of historical accident count data from 1032 junctions between 1992-2003 period were assembled for testing the validity of Poisson assumption. However, sequences from 365 junctions were excluded since these junctions had zero accident count over the whole analysis period. Therefore, sequences from only 667 junctions were tested; of these, the majority (94%) had accident records for as long as 12 years (year 1992-2003) while some others (6%) only had data that varied between 5 to 11 years. The accident data sequences for the 667 junctions were accordingly tagged so that the sequence duration is equal to the duration of available historical data (up to year 2003) for every junction. It is noted that accident count data for year 2004, 2005 and 2006 were excluded from the analysis since the data were available only recently.

Apart from the full-duration (up to year 2003) analysis, the sequence duration of each junction was also fixed to a 5-year analysis period (year 1999-2003), which resulted in 507 out of 1032 junctions having no accident record during the 5-year period. Overall, a total of 1186 samples (sequences of accident data) for various

periods of analysis (between 5 to 12 years) were constructed, which are listed in Table 4.1.

Table 4.1 Number of samples for checking the probability distribution of accident count data

Year of analysis	92-03	93-03	94-03	95-03	96-03	97-03	98-03	99-03
Analysis period (in years)	12	11	10	9	8	7	6	5
Number of samples*	625	2	2	16	5	6	5	525

* not including junctions with zero accident for the whole analysis period

4.5 Testing the validity of Poisson assumption

The sequences of accident count data are characterized by small sample sizes, so the commonly used χ^2 test may not be appropriate for testing the goodness-of-fit of the accident count data. For similar reason, the dispersion index test may also not be appropriate. Among the earlier reviewed goodness-of-fit tests for discrete distribution, the Nicholson and Wong’s exact test, Rao and Chakravarti’s likelihood test, Campbell and Oprian’s Kolomogorov-Smirnov (KS) test and Cressie and Read’s (CR) test may be suitable for testing the validity of Poisson assumption of accident count data. These four tests have different ranges of capability in testing the goodness-of-fit. All the tests, with the exception of CR test, have been designed for testing the goodness-of-fit of Poisson distribution.

Although the exact test is more powerful than the standard chi-square test for testing the goodness-of-fit of the Poisson distribution, this test cannot be performed on very small number of total accident occurrences as there are no applicable lower and/or upper bounds of variance-to-mean ratio. The lowest number of accidents able to be tested by the exact test depends on the sequence duration (analysis period). For example, if the sequence duration is equal to 12, the exact test can only be performed if the total number of accidents for a junction is larger than 9 (for 95% confidence level). If the total number of accidents for the 12-year period falls between 3 and 8, then only the upper bound is available. If the total number of accidents is 1 or 2 during the 12-year period, then no lower and upper bounds are available. However, some sequences that have only an upper bound limit of

variance-to-mean ratio can still be tested with confidence if the observed variance-to-mean ratios are larger than the upper bound, which means that the null hypothesis that the sequences follow Poisson distribution should be rejected. If the total number of accidents is greater than 34 then the dispersion index test is used since it approximates the exact test for large number of accidents. Overall, a total of 915 (77.15%) samples of various analysis periods could not be tested using the exact test.

The likelihood test seems to have a slightly wider range than the exact test. It can be used for sequence period and total number of accidents that are equal to or greater than 3, as long as both the sequence period and the total number of accidents are not equal to 3. When this test was applied to the samples then only 467 (39.38%) samples failed to be tested. The KS test may have advantages over other goodness-of-fit tests for Poisson distribution because, unlike the exact test and likelihood test, it can be used to test all the samples. However, only Campbell and Oprian's KS test can appropriately be used since the parameter of the Poison's distribution was estimated from the samples. Another test that may be appropriate is the CR test. Yet, the test is not applicable in testing 674 (56.83%) samples.

The results of evaluating the goodness-of-fit to Poisson distribution using the four tests at the 95% confidence level are presented in Appendix D. Tables 4.2 and 4.3 give the summary results. Table 4.2 shows that the KS test can test and reject more sequences compared to the other tests. The exact test can only test less than 23% of the samples, but it succeeded in rejecting 14% out of it. Table 4.3 shows that only 102 samples (8.6%) out of 1186 samples of various periods were rejected by at least one of the four tests to be Poisson-distributed at 95% confidence level.

Table 4.2 Results of goodness-of-fit tests to Poisson distribution

	Exact test	Likelihood test	KS test	CR test
Number of samples	1186	1186	1186	1186
Number of untestable samples	915	467	0	674
%-age of untestable samples	77.15%	39.38%	0.00%	56.83%
Number of samples rejected to be PD	38	44	68	41
%-age of rejected samples (from total number of samples)	3.20%	3.71%	5.73%	3.46%
%-age of rejected samples (from number of testable samples)	14.02%	6.12%	5.73%	8.01%

Table 4.3 Pattern of rejection to be Poisson-distributed on rejected samples

Test	Pattern of rejection												Total rejected
Exact test	x	-	-	-	x	x	-	-	x	-	x	x	38
LH test	-	x	-	-	x	-	x	-	x	x	-	x	44
KS test	-	-	x	-	-	x	x	x	x	x	x	x	68
CR test	-	-	-	x	-	-	-	x	-	x	x	x	41
Total rejected	11	5	20	16	2	4	8	6	11	9	1	9	102

Notes:

- x = the test can reject the samples to be Poisson at 95% significance level
- = the test cannot reject the samples to be Poisson at 95% significance level

As many as 59 of the rejected samples were over-dispersed and 39 samples were under-dispersed. It is interesting that 4 samples had dispersion index (variance to mean ratio) equal to one, but they were still rejected to be Poisson-distributed, and the only test that can reject this kind of observation is the Campbell and Oprian’s KS test. This is because the KS test examines the goodness of fit by comparing the largest deviation between the theoretical and observed frequency with a critical value, so it is possible for a sequence with dispersion index equal to one (which usually implies that it is Poisson-distributed) to be rejected.

When the sequence duration was shortened then the number of samples that can be tested to follow Poisson distribution became smaller due to the limitation of available tests as discussed previously. As a result, fewer samples were rejected to be Poisson at 95% significance level. For example, when the sequence duration was 4 years only, the number of samples rejected to be Poisson distributed was 10 samples out of total 1032 samples (including samples with zero accidents during the period). When 3-year and 2-year sequence durations were used, the number of samples rejected to be Poisson distributed became only 3 samples and 1 sample,

respectively. When the sequence duration was only one year then no single test can be performed.

4.6 Stationary Poisson process vs non-stationary Poisson process

As stated previously, the accident occurrence may be governed by a stationary Poisson process or by a non-stationary Poisson process. The result as given in Table 4.2 was obtained by assuming a stationary Poisson process as the underlying accident rate was assumed to be constant over time. However, Figures 3.6 to 3.13 have indicated that the accident rate for the whole study sites was not constant over the years. Therefore, the existence of a non-stationary Poisson process in the accident count data at particular sites was also examined.

The examination was carried out on the 1186 samples (sequences of accident data) from various periods between 1992-2003. Trend analysis was carried out to detect possible trends in the accident count data sequences. On a sequence by sequence basis, it was found that most of the samples were fitted better by a linear trend than by any other trend forms. The analysis revealed that 50% of the samples had downward trends while 41% had upward trends and the rest showed no trends. Furthermore, these trends were not statistically significant (at 5% level) for 95% of the samples. This means that it is acceptable to assume that the accident occurrences at individual sites follow a stationary distribution. Table 4.4 gives the complete results of the trends examination.

Table 4.4 Trends examination

Year of analysis	92-03	93-03	94-03	95-03	96-03	97-03	98-03	99-03
Analysis period	12	11	10	9	8	7	6	5
Number of samples with (at 5% level):								
- significant down-trend	23	1						8
- significant up-trend	15		2	2	1	1		3
- insignificant down-trend	314	1		5	1	2	3	233
- insignificant up-trend	263			7	3	3	2	185
- no trend	10			2				96
Total	625	2	2	16	5	6	5	525

Allowing for trend in the accident count sequence will influence the data dispersion. If no trend is considered, the variance of the accident count sequence is computed from the sum of squares of deviations (total sum of squares) about a horizontal line through the mean. If a trend exists in the data, the variance is computed from the sum of squares of deviations (residual sum of squares) about the trend line.

If the variance-to-mean ratio for the ‘no trend’ case is equal to $(s^2/\bar{x})$ and the variance-to-mean ratio for the ‘with trend’ case is equal to $(s^2/\bar{x})^*$, Nicholson (1985) showed that:

$$\left(\frac{s^2}{\bar{x}}\right)^* = (1 - R^2) \left(\frac{s^2}{\bar{x}}\right) \quad (4.31)$$

Therefore, it is clear that assuming a constant underlying true accident rate (allowing no trend) gives a larger variance-to-mean ratio while assuming a changing underlying true accident rate (allowing for trend) gives a smaller variance-to-mean ratio. It means that when a sequence has an overly small dispersion index and is rejected as following a stationary Poisson process for being under-dispersed then the sequence must also be rejected as following a non-stationary Poisson process as the variance-to-mean ratio for allowing trend is even lower than if no trend is assumed. However, when a sequence has an overly large dispersion index and is rejected as following a stationary Poisson process then no conclusion can be made on the suitability of the non-stationary Poisson process. Nevertheless, one can expect a lower rejection percentage given that over-dispersion was the main reason for rejection.

It is noted that no tests were performed to examine the goodness-of-fit of non-stationary Poisson process as an even more generalised distribution was used subsequently.

4.7 Poisson distribution or generalised Poisson (GP) distribution?

It has been shown (in Tables 4.2 and 4.3) that about 91% of the accident count data at the junctions were not rejected as following the standard Poisson distribution. The next issue is whether the GP distribution can fit the data better than the Poisson distribution. The following paragraphs review the issue more thoroughly while making some conclusions regarding the issue.

4.7.1 Application of generalised Poisson (GP) distribution to accident data

The GP distribution may experience truncation for certain conditions when the data are under-dispersed ($\lambda < 0$ or $\alpha < 0$). One way to overcome this truncation problem is by multiplying each of $f(x|\mu, \alpha)$ with cumulative sum of GP probability up to the point of truncation as suggested by some researchers that were cited by Consul (1989). However, this truncation problem can be avoided if the GP distribution parameters are properly estimated.

As noted earlier, the GP distribution parameters can be estimated through maximum-likelihood (ML) estimator, moment estimator, and mean and zero frequency estimator. With the knowledge gained from the review of various related studies, the following rules are proposed in estimating the GP distribution parameters for this research:

- When all value of x_i 's are equal:
 - Estimate parameters using moment estimator, and mean and zero-frequency estimator.
 - Check whether the value of λ or α estimated by both estimators meet the following requirement (depending on the GP distribution model used) :
 - $\max(-1, -\theta/w) < \lambda \leq 1$ for GP model of Equation (4.1), or
 - $\max(-1/\theta, -1/w) < \alpha < 1/\theta$ for GP model of Equation (4.4)
 - If parameters from both estimators fulfill the requirement then:

- Use moment estimator if $|\lambda| < 0.5$ and $\theta > 2$
- Use mean and zero frequency estimator for other conditions
- If parameter from only one estimator fulfills the requirement then use the estimated parameters from that estimator.
- If parameters from both estimators do not meet the requirement then change the value of λ or α to the minimum value that fulfills the requirement.
- Others: use ML estimator or moment estimator:
 - Estimate parameters using ML estimator and moment estimator.
 - Check whether the value of λ or α estimated by both estimators meet the following requirement (depending on the GP distribution model used) :
 - $\max(-1, -\theta/w) < \lambda \leq 1$ for GP model of Equation (4.1), or
 - $\max(-1/\theta, -1/w) < \alpha < 1/\theta$ for GP model of Equation (4.4)
 - If parameters from both estimators fulfill the requirement then choose estimator that gives the smallest sum of squares of error and make sure that no truncation error exists.
 - If parameters from both estimators do not meet the requirement, or if truncation error exists when using parameters from both estimators then change the value of λ or α to the minimum value that fulfills the requirement and make sure that no truncation error exists.

For subsequent analysis purpose, this research adopted the GP distribution model as defined in Equation (4.1).

4.7.2 Goodness-of-fit tests

The current state of development has not pointed to any single appropriate test that can be used to test the goodness-of-fit of GP distribution. Due to the small sample size, the standard χ^2 test and the score statistic test cannot be used. The KS test by Conover (1972, 1999) is also not appropriate since it requires that the GP parameters must not be estimated from the samples. The only possible test that can

be used is the CR test. However, the number of samples that can be tested by this test was very small such that 1001 samples out of 1186 samples cannot be tested.

Given that Poisson distribution is a member in the family of GP distribution, it follows that samples not rejected as following the Poisson distribution will also not be rejected as following the GP distribution. For these samples, the focus is on whether they are better fitted by the GP distribution than the Poisson distribution.

For samples that have been rejected to be Poisson-distributed, the concern is whether they can be fitted by the GP distribution. Sequences that failed both the Poisson and GP distributions may have not come from a Poisson family. The CR test at 95% confidence level was performed on the 102 samples that did not follow Poisson distribution. The result is given in the following table.

Table 4.5 Result of goodness-of-fit test to GP distribution

	CR test	
Number of samples	102	100%
Number of untestable samples	62	60.78%
Number of samples rejected to be GPD	12	11.76%
Number of samples not rejected to be GPD	28	27.46%

Table 4.5 shows that 60.8% of the samples cannot be tested by the CR test while 11.8% were rejected as being GP distributed and 27.46% were not rejected as being GP distributed (or about 2 in 3 testable samples that failed Poisson distribution did not fail the GP distribution). Overall, the GP distribution was able to fit a larger proportion of the accident sequences than the Poisson distribution.

4.7.3 Sums of squares of error (SSE) and χ^2 statistics

Table 4.6 shows the comparison of χ^2 statistics and sums of squares of errors (SSE) between the standard Poisson distribution and the GP distribution for the overall samples (inclusive of samples that were rejected to be Poisson and GP distributed). The comparison is aimed at examining whether the GP distribution can better describe accident data than the standard Poisson distribution. This kind of

comparison technique has been used by Nicholson (1987) and Schluter (1992) for a similar purpose.

Table 4.6 shows that there were great improvements in adopting GP distribution over Poisson distribution, in terms of the *SSE* and χ^2 statistics. The improvements in *SSE* were in the range of 27.4% to 1983.6%, with 52.4% reduction in *SSE* for samples with a 5-year analysis period. The improvements in the χ^2 statistics fluctuated from 32.9% to 866.5% with 62.8% reduction in χ^2 value for samples with a 5-year analysis period while noting that this statistic is not fully comparable given the (slightly) different degrees of freedom between the Poisson and GP models.

Table 4.6 Comparison of overall *SSE* and χ^2 statistics of Poisson and GP distributions

Analysis period	Number of samples		Sum in <i>SSE</i> *		Sum of χ^2 statistic*		% Improvement of	
	Total	With zero accident	Poisson	GP	Poisson	GP	<i>SSE</i>	χ^2 statistic
5 years	1032	507	1011.890	664.061	1683.490	1033.947	52.4%	62.8%
6 years	13	8	11.915	8.192	10.429	7.847	45.5%	32.9%
7 years	10	4	4.039	0.498	2.633	0.679	710.5%	287.6%
8 years	7	2	23.527	18.469	13.563	9.836	27.4%	37.9%
9 years	20	4	25.394	9.795	26.012	9.848	159.3%	164.1%
10 years	2	0	4.211	0.266	1.662	0.195	1480.9%	751.4%
11 years	2	0	4.705	0.226	2.054	0.212	1983.6%	866.5%
12 years	960	335	2556.688	1601.963	5861.156	1250.729	59.6%	368.6%

* sums for sites which have at least 1 accident for the period of analysis

Table 4.7 shows the differences of *SSE* and χ^2 statistics between the standard Poisson and GP distributions for individual sequences. If the differences are greater than zero, it means that GP distribution fits the accident data better than Poisson distribution. For equal differences, it means that GP distribution is as good as Poisson distribution in modelling accident data. Equal differences are obtained if the accident data sequences have index of dispersion equal to 1. If the differences are smaller than zero, it means that Poisson distribution fits better the accident data. The result of the analysis showed that GP distribution fitted better in about 78% of the samples, and was as good as Poisson distribution in about 16% of the samples.

Table 4.7 Comparison of individual *SSE* and χ^2 statistics of Poisson and GP distributions

Analysis period	Number of comparable samples	Differences of <i>SSE</i> between Poisson and GP distributions*			Differences in χ^2 statistic between Poisson and GP distributions*		
		< 0	= 0	> 0	< 0	= 0	> 0
5 years	525	14 (2.7%)	161 (30.7%)	350 (66.7%)	54 (10.3%)	162 (30.9%)	309 (58.9%)
6 years	5	0 (0%)	0 (0%)	5 (100%)	1 (20%)	0 (0%)	4 (80%)
7 years	6	1 (16.7%)	2 (33.3%)	3 (50%)	0 (0%)	2 (33.3%)	4 (66.7%)
8 years	5	1 (20%)	1 (20%)	3 (60%)	1 (20%)	1 (20%)	3 (60%)
9 years	16	5 (31.3%)	2 (12.5%)	9 (56.3%)	4 (25%)	2 (12.5%)	10 (62.5%)
10 years	2	0 (0%)	0 (0%)	2 (100%)	0 (0%)	0 (0%)	2 (100%)
11 years	2	0 (0%)	0 (0%)	2 (100%)	0 (0%)	0 (0%)	2 (100%)
12 years	625	51 (8.2%)	18 (2.9%)	556 (89%)	23 (3.7%)	18 (2.9%)	584 (93.4%)
Total	1186	72 (6.1%)	184 (15.5%)	930 (78.4%)	83 (7%)	185 (15.6%)	918 (77.4%)

* numbers in bracket denotes percentage of samples of total number of comparable samples

4.8 Summary

From the analysis of the goodness-of-fit, a total of 102 samples (8.6%) out of 1186 samples for various periods were rejected by at least one of four tests as being Poisson-distributed, at 95% confidence level. As many as 59 of the rejected samples were over-dispersed, 39 samples were under-dispersed and 4 samples had index of dispersion (variance-to-mean ratio) equal to one.

Out of the 102 samples that were rejected as being Poisson-distributed, 28 samples (27.46%) were found to follow the GP distribution at 95% confidence level while 12 samples (11.76%) were rejected to be GP distributed; the remaining 62 samples (60.78%) were unable to be tested as to whether they follow a GP distribution or not. It appears that about 2 out of 3 (testable) samples that failed the Poisson distribution can be modelled by the GP distribution though.

Overall, the GP distribution appeared to fit the accident data better than the standard Poisson distribution that is, in general, the goodness-of-fit statistics (*SSE* and χ^2) were significantly improved for the GP distribution. This provided a strong basis in

adopting the GP distribution over the standard Poisson distribution in developing the empirical Bayes (EB) method for estimating road safety, as described in Chapter 7.

It is noted that the GP distribution is not the only alternative to the Poisson distribution in describing the probability distribution of accident occurrences. Other distributions such as binomial or negative binomial (NB) may also be appropriate. However, the binomial and NB probability distributions are not general enough to cover the accident counts that are either under-dispersed, equi-dispersed or over-dispersed. The GP distribution has an advantage in that it can cater to the data that follow binomial, Poisson or NB distributions, and model them all with one single probability distribution.

CHAPTER 5

METHODOLOGY FOR DEVELOPMENT OF ACCIDENT PREDICTION MODELS

5.1 Introduction

This chapter describes the background theory on the development of accident prediction models involving multivariate regression technique using nonlinear models. The methodology for building the models, which involves model formulation, modelling strategy, selection of model explanatory variables, and modelling procedure is explained in respective sections.

5.2 Background theory

There are a number of modelling approaches to deal with discrete count data. The more familiar approach is Poisson regression model. However, it is sometimes the case that the Poisson regression model is not appropriate to model accident count data whenever there is large variability within the data. In this case, a negative binomial (NB) regression model or a generalised Poisson (GP) regression model may be used as they can cater to the over-dispersion in the data. NB model assumes that the distribution in the number of accidents at a site is Poisson and the distribution of expected accident counts in the population is Gamma. Sometimes the accident count data experience excess zeros so a zero inflated Poisson (ZIP) regression model or a zero inflated negative binomial (ZINB) regression model may be appropriate. The background theory regarding the above-mentioned regression models are given in the following subsections.

5.2.1 Poisson regression model

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

$$f(Y_i = y_i) = \frac{\mu_i^{y_i} \exp(-\mu_i)}{y_i!}; \quad i = 1, \dots, n \quad (5.1)$$

where

$$E(Y_i) = \mu_i = \exp(X_i' \beta) = \exp\left(\sum_1^p x_{ij} \beta_j\right) \quad (5.2)$$

and

$$\text{Var}(Y_i) = \mu_i \quad (5.3)$$

where x_{ij} is the value of the j th covariate for observation i . In other words, $\mu(X_i, \beta)$ denotes the function that relates the mean response $\mu_i (n \times 1)$ to $X_i (n \times p)$, the values of the explanatory variables for case i , and to $\beta (p \times 1)$, the values of the regression coefficients.

The Poisson regression model does have its limitations, especially when the variance of the accident data is constrained to be equal to its mean, $\text{Var}(Y_i) = E(Y_i) = \mu$. In many cases, accident data were shown to have variance greater than the mean (a case of over-dispersion), which can result in biased model coefficients and erroneous standard errors. To overcome this problem of over-dispersion, the NB regression model or the GP regression model is recommended.

5.2.2 Negative binomial (NB) regression model

The NB distribution is suitable for describing discrete, non-negative events like frequency of traffic accidents. It has advantage over the Poisson distribution in that it does not require the mean to be equal to the variance. The appropriateness of using NB model over Poisson regression model can be determined by the statistical significance of an estimated coefficient α which is a parameter related to the degree of over-dispersion. When α is significantly larger than zero (as measured by the t-statistic), then the suitability of the NB regression model is confirmed. Otherwise, the Poisson model would prevail (Poch and Mannering, 1996).

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

$$f(Y_i = y_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} ; i = 1, \dots, n \quad (5.4)$$

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

Here,

$$E(Y_i) = \mu_i = \exp(X_i' \beta) = \exp\left(\sum_1^p x_{ij} \beta_j\right) \quad (5.5)$$

and

$$Var(Y_i) = \mu_i + \alpha\mu_i^2 \quad (5.6)$$

The Poisson regression model can be regarded as a limiting model of the NB regression model as α approaches zero. However, it is noted that the NB model is unable to cater to under-dispersed data.

Miaou (1994) stated that although the NB regression model is more general than the Poisson regression model, it requires more extensive computations to estimate model parameters and to generate inferential statistics than the Poisson regression model.

5.2.3 Generalised Poisson (GP) regression model

If Y_i is a count variable that follows a GP distribution, then the probability function of Y_i is given by:

$$f(Y_i = y_i) = \left(\frac{\mu_i}{1 + \alpha\mu_i}\right) \frac{(1 + \alpha y_i)^{y_i-1}}{y_i!} \exp\left(\frac{-\mu_i(1 + \alpha y_i)}{1 + \alpha\mu_i}\right) ; i = 1, \dots, n \quad (5.7)$$

with α being the dispersion parameter.

where

$$E(Y_i) = \mu_i = \exp(X_i' \beta) = \exp\left(\sum_1^p x_{ij} \beta_j\right) \quad (5.8)$$

and

$$Var(Y_i) = \mu_i (1 + \alpha \mu_i)^2 \quad (5.9)$$

The GP regression model is a generalisation of the standard Poisson regression model such that when $\alpha = 0$ the GP probability function reduces to the Poisson regression model. When $\alpha > 0$ the GP regression model represents count data with over-dispersion, and when $\alpha < 0$, the GP regression model represents count data with under-dispersion. The GP regression model should be used carefully when $\alpha < 0$ as the sum of the GP probability function may not be equal to unity. However this problem does not exist when $\alpha > 0$ in that the sum of the GP probability function always equal to unity (Famoye *et al.*, 2004).

Famoye *et al.* (2004) suggested that the GP regression model should be used when the data dispersion is unknown since it is more flexible, but when the data are not under-dispersed, then the NB regression model may be used. Famoye *et al.* (2004) also found similarity of the results of NB regression model with those of the GP regression model when the data are over-dispersed.

5.2.4 Zero inflated Poisson (ZIP) regression model

The ZIP regression model has the following form:

$$\begin{aligned} P(Y_i = y_i) &= q_i + (1 - q_i) e^{-\lambda_i} \quad ; \quad y_i = 0 \\ P(Y_i = y_i) &= (1 - q_i) \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad ; \quad y_i = 1, \dots, n \end{aligned} \quad (5.10)$$

and

$$\log\left(\frac{q_i}{1 - q_i}\right) = \tau \sum_1^p x_{ij} \beta_j \quad (5.11)$$

$$\lambda_i = \exp(X_i' \beta) = \exp\left(\sum_1^p x_{ij} \beta_j\right) ; \quad i = 1, \dots, n \quad (5.12)$$

where τ is a scalar parameter.

The mean and variance of Y_i are:

$$E(Y_i) = \mu_i = (1 - q_i) \lambda_i \quad (5.13)$$

$$Var(Y_i) = \mu_i + \left(\frac{q_i}{1 - q_i}\right) \mu_i^2 \quad (5.14)$$

When $q_i = 0$, the ZIP regression model is identical to the Poisson regression model. Also, when $0 < q_i < 1$, the variance of Y_i will exceed its mean. Thus, the ZIP regression model allows over-dispersion in the data due to excess zeros when compared to the standard Poisson regression model.

When the data have excess zeros, then the ZIP regression model is preferable to NB regression model as it allows for the over-dispersion in the data due to excess zeros. Miaou (1994) concluded in his study that the interpretation of the ZIP model could be difficult as the accident frequency is related to other accident-related variables in a much more complicated way in the ZIP model than in the Poisson regression model. But he also noted that the ZIP model does ensure the non-negativity of accident counts.

5.2.5 Zero inflated negative binomial (ZINB) regression model

The ZINB regression model has the following form:

$$P(Y_i = y_i) = q_i + (1 - q_i) \left[\frac{1}{1 + \alpha \mu_i} \right]^{1/\alpha} ; \quad y_i = 0$$

$$P(Y_i = y_i) = (1 - q_i) \left[\frac{\Gamma\left(y_i + \frac{1}{\alpha}\right) \left(\frac{1}{1 + \alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i}}{\Gamma\left(\frac{1}{\alpha}\right) y_i!} \right] ; y_i = 1, \dots, n \quad (5.15)$$

with α being the dispersion parameter, and

$$\log\left(\frac{q_i}{1 - q_i}\right) = \tau \sum_1^p x_{ij} \beta_j \quad (5.16)$$

where τ is a scalar parameter.

The mean and variance of Y_i are:

$$E(Y_i) = \mu_i = (1 - q_i) \lambda_i \quad (5.17)$$

$$Var(Y_i) = \mu_i + \left(\frac{q_i + \alpha}{1 - q_i}\right) \mu_i^2 \quad (5.18)$$

where

$$\lambda_i = \exp(X_i' \beta) = \exp\left(\sum_1^p x_{ij} \beta_j\right) ; i = 1, \dots, n \quad (5.19)$$

Since the variance of NB regression model has the form $Var(Y_i) = \mu_i + \alpha\mu_i^2$, the term $q_i/(1 - q_i)$ could erroneously be interpreted as α . When applying a zero inflated model, the problem then becomes one of distinguishing the underlying NB or Poisson distribution from the zero inflated distribution. A statistic that can be used to assess the appropriateness of zero inflated distribution is Vuong statistic.

Vuong statistic is used to assess the suitability of using the zero inflated models (ZIP or ZINB), which is computed as:

$$V = \frac{\overline{M}\sqrt{n}}{S_M} \quad (5.20)$$

where $\bar{M}$ is the mean with $M = \log[f_1(.) / f_2(.)]$, $f_1(.)$ being the density function to ZIP or ZINB distribution and $f_2(.)$ is the density function of the standard Poisson or NB distribution, S_M and n are the standard deviation and sample size, respectively. A value of $V > 1.96$ (95% confidence level for the t-test) favours the ZIP or ZINB while a value of $V < -1.96$ favours the standard Poisson or NB distribution, and values between -1.96 and 1.96 indicate the test is indecisive (Shankar *et al.*, 1997).

5.2.6 Assessment of model goodness-of-fit

Evaluation of the goodness of the fit between the observed values y_i and the fitted values $\hat{\mu}_i$ for the generalised linear models family can be assessed by a number of statistics. Two well known ones are the Pearson X^2 and Scaled Deviance G^2 statistics. Here the models are assumed to be nested, with the larger model having the greater number of parameters. These statistics follow χ^2 distribution with $(n-p)$ degrees of freedom where n is the number of data points (observations) and p is the number of estimated parameters. If the value of expected value of G^2 ($E(G^2)$) is close to $(n-p)$, it is taken as an indication that the fitted model is adequate (Pickering *et al.*, 1986).

The Pearson X^2 statistic is defined as sum of squares of standardised observations, while the Scaled Deviance G^2 can be defined as twice the logarithm of the ratio of the likelihood of the data under the larger model, to that under the smaller model, or mathematically,

$$\text{Pearson's } X^2 = \sum_{i=1}^n \frac{(y_i - E[y_i])^2}{\text{VAR}[y_i]} \quad (5.21)$$

and

$$\text{Scaled Deviance } G^2 = 2(l(y|y) - l(\mu|y)) \quad (5.22)$$

where

- $l(y|y)$ is the log-likelihood for the maximal model, i.e. that model which exactly fits all the data points
- $l(\mu|y)$ is the log-likelihood for the reduced model

The Pearson X^2 and Scaled Deviance G^2 statistics for the Poisson model are:

$$X^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \quad (5.23)$$

$$G^2 = 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right] \quad (5.24)$$

For the NB model:

$$X^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i + \alpha \hat{\mu}_i^2} \quad (5.25)$$

$$G^2 = 2 \sum_{i=1}^n \left(\ln \left(\frac{y_i}{\hat{\mu}_i} \right)^{y_i} - (y_i + 1/\alpha) \ln \left(\frac{1 + \alpha y_i}{1 + \alpha \hat{\mu}_i} \right) \right) \quad (5.26)$$

Other measures that can be used, but not as popular as the Pearson X^2 and Scaled Deviance G^2 , are log-likelihood value and Akaike Information Criterion (*AIC*). A model with a high log-likelihood value and low *AIC* value is preferred (Miaou, 1994). *AIC* can be computed through the following equation:

$$AIC = -2l(\mu|y) + 2p \quad (5.27)$$

where $l(\mu|y)$ and p are as defined previously.

5.2.7 Identification of outliers

Outliers are observations that are unusual or extreme which are not typical of the rest of the data in a dataset. They occur in a set of data either because they are genuinely different from the rest of the data or because errors took place during data collection and/or recording. Outliers may be extreme in their values for the response

variable, the explanatory variables, or both which deserve special investigation in order to decide whether they should be removed from the dataset or not as they may strongly affect the values of the model parameter estimates.

A well-known measure to detect influential outliers on the fitted value μ_i is Cook's distance (D_i), which is defined as:

$$D_i = \frac{r_{p_i}^2 h_{ii}}{p(1 - h_{ii})^2} \quad (5.28)$$

where h_{ii} is leverage statistic of the model (the diagonal elements of the $n \times n$ hat matrix, which is the matrix that multiplies the vector of observations to yield the vector of predicted values), p is the number of parameters estimated in the model and r_{p_i} is Pearson Residual of the model, which is:

$$r_{p_i} = \frac{y_i - \mu_i}{\sqrt{\text{Var}(\mu_i)}} \quad (5.29)$$

Cook's distance measures the standardized change in the fitted value μ_i when the i th observation is deleted. Meanwhile, leverage statistic h_{ii} is useful for identifying outliers on the X -space (matrix of explanatory variables).

Data points having a high Cook's distance are influential outliers that play a significant role in determining the model parameter estimates. Unfortunately, no rules are currently available to judge what constitutes a high Cook's distance for nonlinear models. The judgement whether an observation is outlying or overly influential is typically made on the basis of a subjective visual assessment on a plot of Cook's distance value versus the observation number i .

The Numerical Algorithms Group (1994) has proposed the following procedure for identifying influential outliers:

- The data are sorted in descending order according to the Cook's distance value.

- Remove the observations one by one, starting with the observation having the largest Cook's distance values, and then compute the drop in scaled deviance after the removal of each observation.
- Observations causing a significant drop in scaled deviance are influential outliers, i.e. if the deletion of an observation causes a drop in scaled deviance exceeding $\chi^2_{0.05,1} = 3.84$.

It should be noted that unlike the Poisson regression model, the scaled deviance of two NB regression models with different values of dispersion parameter cannot be directly compared. The value of dispersion parameter for the model with n observations should be imposed upon the model with $n-1$ observations before the scaled deviance can be compared.

Besides Cook's distance, Kutner *et al.* (2004) have also suggested methods to detect observations that are influential to Pearson X^2 statistic and Scaled Deviance G^2 . Similar to the Cook's distance, a case-deletion diagnostics should also be used to assess the effect of individual observations on the result of the analysis.

The i th ΔX_i^2 statistic is defined as the change in the Pearson statistic when the i th observation is deleted:

$$\Delta X_i^2 = X^2 - X_i^2 \quad (5.30)$$

The i th ΔG_i^2 statistic is defined as the change in the deviance statistic when the i th observation is deleted:

$$\Delta G_i^2 = G^2 - G_i^2 \quad (5.31)$$

For faster computation, the following approximations have been developed:

$$\Delta X_i^2 = r_{SP_i}^2 \quad (5.32)$$

$$\Delta G_i^2 = h_{ii} r_{SP_i}^2 + G_i^2$$

where $r_{SP_i}^2$ is the studentized Pearson residuals, which are defined as:

$$r_{SP_i} = \frac{r_{P_i}}{\sqrt{1 - h_{ii}}} \quad (5.33)$$

Therefore, the ΔX_i^2 and ΔG_i^2 statistics provide measures of the influence of the i th observation on the Pearson X^2 and Scaled Deviance G^2 statistics.

5.3 Model formulation

It is common in accident analysis to assume that accident rate (accident per unit of exposure) is constant over time. However, this assumption does not usually represent the reality. While exposure is increasing over the years sometimes it turns out that the number of accidents is decreasing, which indicates that the assumption of a constant accident rate is not valid. In general, accident rate over the years could be declining although the exposure is increasing due to, for example, effective road safety programmes.

The assumption of constant accident rate is usually sufficient when the accident prediction model is developed based on short-term period, which is usually taken between 3 to 5 years period. If a longer period is used, it is important to allow for possibility of a change in accident risk. Although the accident data analysis presented in Chapter 3 could not reveal that such trend exists in the accident data at study junctions during 1992-2005 period, it is still important to examine the existence of the trend in the modelling stage since the obtained trend had not been corrected for exposures.

If there is a declining trend of accidents during year 1 to year t , the exclusion of a trend factor from an accident prediction model (assuming a constant accident rate) would underestimate the frequency in year 1 and overestimate the frequency in year t and also the predicted accident frequency in year $(t+1)$, and so on. Therefore, trend factor should be included to obtain a more reliable accident prediction model if the accident occurrences have a declining trend over the years.

Basically, two types of accident prediction models were developed in this research. The first type of model used longer (1992-2003) accident data series and this so-called long-term model incorporated a trend factor and it was developed only in cases where the trend factor was found to be statistically significant. Such long-term model should also be useful in making accident predictions over a longer future period. The second type of model was developed using shorter (between 1 to 5 years' period, starting from 2003 backwards) accident data, and such short-term model was formulated based on constant accident rate over the period; before model development, the assumption of constant accident rate in the short-term model was appropriately verified. However, the verification was only carried out for models that were developed based on 5-year period only since 5-year period may be too long a period to assume that the accident rates had not changed. After model developments, the long-term and short-term models were compared and the best model form was selected based on the performance of the respective model in predicting short-term (2004, 2005 and 2006) accident frequencies.

Long-term models were developed using disaggregated data, so that accidents occurring in each year at each junction represent a unit of data. Mountain *et al.* (1998) stated that using disaggregated data has the benefit of obtaining a minimum variance estimate of the average effect. However, by using disaggregated data the error structure would be a mixture of between-site error which is random and within-site error (the latter being highly correlated). This may result in a variable to be significant when data are disaggregated but becoming insignificant upon data aggregation such that the standard errors of the model's parameters for disaggregated case are underestimated. For this reason, Maher and Summersgill (1996) suggested it is better to use aggregated form for a plausible interpretation of the actual error structure, but at the expense of not being able to incorporate the year-to-year variation in the accident data. The approach of using disaggregated data for long-term models is consistent with the findings of Shankar *et al.* (1998) who found that the NB model for disaggregated data is statistically adequate when the temporal correlation is explicitly included in the form of trend variable.

Let μ_t equal to expected accident number at year t (where $t = 1, 2, 3, \dots$), G_t equals to junction geometric/operational variables and $k, \gamma, \gamma_1, \gamma_2, \beta_i, \omega$ are the model parameters to be estimated. The general form of the accident prediction model is:

$$\mu_t = k(t)^\omega (f(Q)_t) e^{\sum(\beta_i G_i)} \quad (5.34)$$

and $f(Q)_t$ is the functional form of traffic flow at year t that can be of the forms:

$$f(Q) = (Q)^\gamma \quad (5.35)$$

$$f(Q) = (Q_p Q_s)^\gamma \quad (5.36)$$

$$f(Q) = (Q_p)^{\gamma_1} (Q_s)^{\gamma_2} \quad (5.37)$$

where

Q = total daily junction inflow

Q_p = total daily junction inflow in the primary direction

Q_s = total daily junction inflow in the secondary direction

For X-junctions, primary direction is the direction which has a larger traffic inflow while primary direction for T-junctions is the direction in which two legs are facing each other (the head of T). Equation (5.35) assumes that accident is related to the sum of traffic inflows from all legs while Equations (5.36) and (5.37) assume that accident is related to the cross-product of traffic inflows from the primary direction and the secondary direction.

The constant k may be regarded as a measure of risk. The risk changes from year to year by a trend factor $(t)^\omega$. A value of $\omega < 1$ means the risk is declining while $\omega > 1$ means that the risk is increasing over time, and $\omega = 0$ means that the risk is not changing. Consequently, if for year 1

$$\mu_1 = k(f(Q)_1)^\gamma e^{\sum(\beta_i G_i)} \quad (5.38)$$

then for subsequent years, the constant k will be corrected by the trend factor $(t)^\omega$.

The total expected accident number over the whole study period is then:

$$\sum \mu_t = k \sum (t)^\omega (f(Q)_t)^\gamma e^{\Sigma(\beta_i G_i)} \quad (5.39)$$

It should be noted that the model form described by Equation (5.34) assumes that the effect of road safety programmes is not changing from year to year while in reality the effect of road safety programmes for a certain year may be larger than for other years due to, for example, a substantial road safety programme that was applied in that year.

The trend factor could be in a functional form different from the power function used in the above model, for example a geometric function as used by Maher and Summersgill (1996) and Mountain *et al.* (1998). However, the use of geometric trend requires some complicated transformations before the model can be solved using the standard method of solving non-linear model form. The power function is selected firstly, because it is simple so that no single transformation is needed in solving the model parameters and secondly, because it is in accordance with the accident trend specifically at the study junctions and also generally nationwide. The third reason is that it has properties whereby the risk in the first year is equal to model constant k and the risk in the subsequent years is changing by a trend factor $(t)^\omega$ relative to the first year risk.

When the trend factor is excluded as for developing the short-term models, then the model form in Equation (5.34) changes to:

$$\mu = k(f(Q)) e^{\Sigma(\beta_i G_i)} \quad (5.40)$$

The accident data for developing short-term models were in aggregated form, where accidents occurring during each modelled year at each junction represented one unit of data and Q represented average daily junction flow.

5.4 Modelling strategy

Both the long-term and short-term models are to be developed at two different stages. Stage 1 models are simple accident-exposure models, where accident counts

are linked with traffic flows. The functional forms of the traffic flows are varied according to the three forms described previously. The developed models at this stage are disaggregated into X-signalised model and T-signalised model. No Stage 1 model can be developed for unsignalised junctions since no traffic exposure data were available. The intention of developing Stage 1 model is to select the best functional forms of traffic flows that can best describe the interaction of exposure with accident occurrences, which are used later in Stage 2 modelling. The Stage 1 (long-term) model is of the following forms:

$$\mu_t = k(t)^\omega (Q_t)^\gamma \quad (5.41)$$

$$\mu_t = k(t)^\omega (Q_{p_t} Q_{s_t})^\gamma \quad (5.42)$$

$$\mu_t = k(t)^\omega (Q_{p_t})^{\gamma_1} (Q_{s_t})^{\gamma_2} \quad (5.43)$$

The Stage 1 (short-term) model is of the following forms:

$$\mu = k(Q)^\gamma \quad (5.44)$$

$$\mu = k(Q_p Q_s)^\gamma \quad (5.45)$$

$$\mu = k(Q_p)^{\gamma_1} (Q_s)^{\gamma_2} \quad (5.46)$$

At Stage 2, a full model is built for which accident counts are linked with traffic flow and junction geometric/operational variables. The Stage 2 (long-term) models for X-signalised and T-signalised junctions shall take any of the following forms (depending on the result of the Stage 1 modelling):

$$\mu_t = k(t)^\omega (Q_t)^\gamma e^{\Sigma(\beta_i G_i)} \quad (5.43)$$

$$\mu_t = k(t)^\omega (Q_{p_t} Q_{s_t})^\gamma e^{\Sigma(\beta_i G_i)} \quad (5.44)$$

$$\mu_t = k(t)^\omega (Q_{p_t})^{\gamma_1} (Q_{s_t})^{\gamma_2} e^{\Sigma(\beta_i G_i)} \quad (5.45)$$

The Stage 2 (short-term) models for X-signalised and T-signalised junctions shall take any of the following forms (depending on the result of the Stage 1 modelling):

$$\mu = k(Q)^\gamma e^{\Sigma(\beta_i G_i)} \quad (5.46)$$

$$\mu = k(Q_p Q_s)^\gamma e^{\Sigma(\beta_i G_i)} \quad (5.47)$$

$$\mu = k(Q_p)^{\gamma_1}(Q_s)^{\gamma_2}e^{\Sigma(\beta_i G_i)} \tag{5.48}$$

The Stage 2 (long-term) model for X-unsignalised and T-unsignalised junctions is of the form:

$$\mu_i = (t)^{\omega} e^{\Sigma(\beta_i G_i)} \tag{5.49}$$

while the Stage 2 (short-term) model for X-unsignalised and T-unsignalised junctions is of the form:

$$\mu_i = (t)^{\omega} e^{\Sigma(\beta_i G_i)} \tag{5.50}$$

It is to be noted that accident prediction models for X-unsignalised and T-unsignalised junctions can be developed using only the geometric data of the junctions since no traffic count data were readily available for unsignalised junctions in Singapore. It is undesirable to make use of rough estimates in the flow which would introduce further error into the models. On the other hand, collection of data by field survey would be overly resource-intensive. As the flow of traffic entering a junction could be proportional to the number of approach lanes, the variable of number of approach lanes was evaluated as a possible proxy indicator of traffic flow (see Figures 5.1 and 5.2).

Figure 5.1 Relationship between X-signalised junctions’ flow with number of approach lanes

Figure 5.2 Relationship between T-signalised junctions’ flow with number of approach lanes

Figures 5.1 and 5.2 show that the total junction inflow entering a junction was somewhat related to the number of approach lanes. However, the relationship was rather weak. Therefore, the number of approach lanes should not be used as a proxy, even though no traffic flow data were available at unsignalised junctions.

5.5 Model explanatory variables

In general, there are two types of explanatory variables to be used in a regression model, which are quantitative and qualitative explanatory variables. A qualitative variable should first be transformed into quantitative indicators before it can be used in a regression model. There are two common ways of doing it, firstly by using indicator variables and secondly by using allocated codes. An indicator variable takes on the value 0 or 1 to quantitatively identify classes of a qualitative variable. This approach is easy to use and is widely employed. If only two classes are available, then only one indicator variable is needed. If more than two classes are available, then the number of indicator variables needed is equal to the number of classes minus one, each taking on the value 0 or 1. Indicator variables are also called dummy variables.

In the allocated codes approach, a single variable is employed to represent a qualitative variable, and values are assigned for classes of the variable. For example, if there are three classes (class A, B, C) of a qualitative variable, then values 1, 2 and 3 are assigned for each class. The coding 1, 2, 3 implies that the mean response changes by the same amount when going from class A to class C. This is the main disadvantages of using allocated codes in that such coding assigns a predefined distance between the classes, which may not be in accordance with reality. On the other hand, no assumption regarding the spacing of the classes is needed for indicator variables. The effect on mean response relies on the data itself.

It is unavoidable in accident modelling to use both quantitative and qualitative explanatory variables. The example of quantitative variables is traffic flow, while the geometric/operational characteristics may be quantitative (e.g. number of approach lanes) or qualitative (e.g. land use). Tables 5.1 and 5.2 present the explanatory variables that were considered in developing the accident prediction models in this research. Indicator variables were used to do the transformation of the qualitative variables. It is to be noted that not all of the geometric/ operational characteristics collected (as presented in Table 3.5) were eventually used. Some geometric/operational variables were omitted due to lack of data variation (e.g. most sites had no traffic island or had one single speed limit of 50 km/h) or some were linear combinations of other variables (e.g. number of shared lanes vs number of shared left-thru lanes, number of shared right-thru lanes, number of shared left-right lanes and number of shared left-thru-right lanes; sliproad vs sliproad with storage lane and sliproad without storage lane; number of exclusive right turn lanes vs number of exclusive right turn lanes with storage lane and without storage lane; number of lanes vs number of approach lanes and number of exit lanes).

Table 5.1 Explanatory variables for signalised junctions

Explanatory variables	Codes	Types	Coding, as applicable	Remarks
Traffic flow	Q, Q _p , Q _s	Quantitative		
Number of approach lanes	NAPLN	Quantitative		
Number of exit lanes	NEXLN	Quantitative		
Total leg width (m)	TOTLW	Quantitative		
Total approach width (m)	TOTAPW	Quantitative		
Total exit width (m)	TOTEXW	Quantitative		
Number of exclusive right-turn lanes	RTLANE	Quantitative		
Number of shared lanes	SHLANE	Quantitative		
Number of signal phases	NPHASE	Quantitative		
Classification of junction	CLASS	Qualitative	1 : if major junction 0 : otherwise	JTYPE 1 0
Sliproad	SLR	Qualitative	1 : if sliproad present on all junctions legs 0 : otherwise	SLR NOSLR 0 1 0 0 1 0
	NOSLR	Qualitative	1 : if sliproad not present on all junctions legs 0 : otherwise	
Median	MED	Qualitative	1 : if median present on all junctions legs 0 : otherwise	MED NOMED 0 1 0 0 1 0
	NOMED	Qualitative	1 : if median not present on all junctions legs 0 : otherwise	
Pedestrian crossing	PEDCROSS	Qualitative	1 : if pedestrian crossing present on all junctions legs 0 : otherwise	PEDCROSS 1 0
Downstream merger	MERG	Qualitative	1 : if downstream merger present on at least one junction leg 0 : otherwise	MERG 1 0
U-turn	UTURN	Qualitative	1 : if U-turn present on at least one junction leg 0 : otherwise	UTURN 1 0
Red-light camera	RLCAM	Qualitative	1 : if red-light camera presents on at least one junction leg 0 : otherwise	RLCAM 1 0

Table 5.1 Explanatory variables for signalised junctions

Explanatory variables	Codes	Types	Coding, as applicable	Remarks
Yellow box	YBOX	Qualitative	1 : if yellow box presents 0 : otherwise	Yellow box presents Yellow box not presents YBOX 1 0
Landuse	CITY	Qualitative	1 : if junction is located in city area 0 : otherwise	Junction is located in city area 0
	TOWN	Qualitative	1 : if junction is located in town area 0 : otherwise	Junction is located in town area Junction is located in other area CITY 0 1 TOWN 0 0

Table 5.2 Explanatory variables for unsignalised junctions

Explanatory variables	Codes	Types	Coding, as applicable	Remarks
Number of approach lanes	NAPLN	Quantitative		
Number of exit lanes	NEXLN	Quantitative		
Total leg width (m)	TOTLW	Quantitative		
Total approach width (m)	TOTAPW	Quantitative		
Total exit width (m)	TOTEXW	Quantitative		
Number of exclusive right-turn lanes	RTLANE	Quantitative		
Number of shared lanes	SHLANE	Quantitative		
Median	MED	Qualitative	1 : if median present on all junctions legs 0 : otherwise	No median for all legs MED 0
	NOMED	Qualitative	1 : if median not present on all junctions legs 0 : otherwise	Median for some legs Median for all legs NOMED 0 1 0
Yellow box	YBOX	Qualitative	1 : if yellow box presents 0 : otherwise	Yellow box presents Yellow box not presents YBOX 1 0
Landuse	CITY	Qualitative	1 : if junction is located in city area 0 : otherwise	Junction is located in city area CITY 1
	TOWN	Qualitative	1 : if junction is located in town area 0 : otherwise	Junction is located in town area Junction is located in other area TOWN 0 1 0

5.6 Modelling procedure

Figure 5.3 describes the general framework in developing the accident prediction model. Firstly, the database was cleaned by removing junctions with non-Poisson accident data sequence during 1999-2003 to stay in line with the basic assumption in building the regression model that within-site error structure is Poisson. After that, the database was ready to be used to develop the Stage 1 models. It has to be noted that the accidents at the non-Poisson junctions were predicted using the same accident prediction models developed in this research though these junctions were not involved in the development of the models.

Four types of nonlinear regression models (error structures) were considered: Poisson, NB, ZIP and ZINB. The GP regression model was excluded from analysis based on Famoye's recommendation stated previously (see Section 5.2.3). Figure 5.4 describes the modelling procedure to determine the most appropriate regression model (error structure) for each form of model being examined. LIMDEP software was used for estimating the parameters of the models.

Stage 1 models were developed according to the procedure illustrated in Figure 5.4. Three model forms were examined and the resulting model was considered appropriate when:

- The estimated parameters of all model explanatory variables were statistically significant at 95% confidence level, as shown by the value of the t-value of the estimated parameters that should be greater than or equal to 1.96, or by P-values of the estimated coefficients that should be less than or equal to 0.05.
- The model passed goodness-of-fit assessment, in the form of Scaled Deviance G^2 and Pearson X^2 statistics.

The best model form was obtained when a model passed the above criteria and had the highest log-likelihood value and lowest *AIC* value among the selected models. Other criteria, such as simplicity of the model, were also considered in selecting the best model form.

Figure 5.3 Modelling framework

Figure 5.4 Modelling procedure

Stage 2 modelling involved the pre-selection of geometric/operational variables to be included in the model. Unlike Stage 1 modelling, only selected junction geometric/operational variables as given in Tables 5.1 and 5.2 were used. Any selected geometric/operational explanatory variable should not have strong correlation with other explanatory variable(s). Correlation analysis was done to minimise the problem of multicollinearity that may bias the standard error of the estimated parameter. For the present study, explanatory variables that were highly correlated (coefficient of correlation $\rho > 0.6$) were compared and the less significant variables were dropped.

Modelling was then carried out for signalised-junction model using regression model (error structure) that was established in Stage 1 modelling. For unsignalised junctions, modelling was carried out using the procedure shown in Figure 5.4. In both cases, variable selection was carried out using a backward elimination procedure whereby the least significant variable at 95% confidence level was progressively eliminated one by one from the model. This process was continued until all remaining variables were statistically significant to be retained in the model.

Afterwards, goodness-of-fit assessment was performed on the fitted model, in the form of Scaled Deviance G^2 and Pearson X^2 statistics. Sometimes, a model can fail a goodness-of-fit test even when all the parameter estimates in the model are significant. When this condition occurs, one should attempt to identify the cause. Causes may include inappropriate formulation of model, inappropriate regression model or existence of outliers. If the cause is due to inappropriate formulation of model or inappropriate regression model, then one should try to use an alternative formulation or regression model. However, if the cause is due to existence of outliers then the outliers should be removed from the model.

The identification of the possible causes of poorly-fitted model was carried out using a trial and error procedure. The first step was to check the existence of outliers by detecting observations that have a strong influence on Pearson X^2 or

Scaled Deviance G^2 statistic (using Cook's distance D_i and leverage statistic h_{ii}). SAS application software was used to examine the existence of outliers in the models. If no outliers are detected then use a different regression model. If all the model's parameters are significant and the goodness-of-fit statistics improve but still does not pass the required criteria then re-check for the existence of outliers. If the new regression model cannot improve the goodness-of-fit statistics then use a different model formulation. However, this last case did not occur in the present research since various model formulations were already assessed in previous stages. Even when a model passes the goodness-of-fit assessment, it is still important to investigate the existence of outliers in the model by using Cook's distance D_i and leverage statistic h_{ii} as outliers can influence the estimated model parameters. A final model is obtained when all the model's parameters are significant, no influential outliers exist and the model passes the goodness-of-fit criteria.

5.7 Summary

The main focus of this chapter is to gather together a set of sound procedures in developing well-fitting accident prediction models which were applied in building accident prediction models for Singapore road junctions. Basically, there were two types of model developed in this research, which were long-term models and short-term models. The long-term models were models that were formulated by including the trend in accident occurrences using long (1992-2003) accident data series. The short-term model was formulated by assuming constant accident rate over several periods within 1999-2003. The long-term and short-term models were to be compared and the best model selected based on the performance of the respective model in predicting short-term (2004, 2005 and 2006) accident frequencies.

The accident prediction models shall be developed in two stages; Stage 1 models are exposure only models while Stage 2 models consider both exposure and geometric/operational variables. For signalised-junction model, three accident-exposure model forms were considered. Poisson, NB, ZIP and ZINB regression models (error structures) were assessed. The model's explanatory variables were

selected through correlation analysis and backward elimination procedure. Goodness-of-fit tests were carried out using Scaled Deviance G^2 and Pearson X^2 statistics while outliers analysis were carried out using Cook's distance D_i and leverage statistic h_{ii} . The best model form was obtained when a model passed the goodness-of-fit criteria and had the highest log-likelihood value and lowest AIC value among the selected models. The developed models are described further in Chapter 6.

CHAPTER 6

ACCIDENT PREDICTION MODELS

6.1 Introduction

This chapter covers the model development which entails detailed analysis into the factors that are associated with accident occurrences at the road junctions. A number of accident prediction models are developed according to the strategies and procedures presented in Chapter 5. The notation and definition of the explanatory variables are as presented in Tables 5.1 and 5.2. The resulting models are to be used in estimating safety by the empirical Bayes (EB) method in the next stage of research. A comparison is also made between the accident prediction models which incorporate long-term accident trend (long-term models) and ones which do not incorporate trend (short-term models).

6.2 Long-term models

The long-term accident prediction models were developed based on the maximum available period of accident data from years 1992-2003 in which 93% of the junctions had the maximum period of 12 years while the period for the rest of junctions varied between 5 to 11 years. A trend variable was included in the development of the long-term models to examine whether there has been a decrease in the accident rate over 1992-2003 period.

Each of the long term models for signalised junctions was developed in two stages, which were accident-exposure models (stage 1) and models that contained junction geometric/operational variables on top of exposure variable (stage 2). Basically, there were two types of models developed in stage 2, which were named as full long-term accident prediction model and simple long-term accident prediction model. The full long-term accident prediction model contained exposure variables and junction geometric/operational variables that were selected using stepwise

regression method as the model's independent variables on top of a trend variable. Here, the full long-term accident prediction model was developed based on the assumption that all the junction geometric/operational variables were not changed during 1992-2003 period. On the other hand, the simple long-term accident prediction model contained exposure variable and junction geometric/operational variables that can be confirmed to remain unchanged from 1992-2003 period on top of a trend variable. In this case, only variables representing junction classification (*CLASS*) and landuse (*CITY* and *TOWN*) variables were used in the development of the simple long-term model.

Long-term models for unsignalised junctions could not be developed because no traffic exposure was available during 1992-2003 period. Moreover, assuming that junction geometric/operational variables remained unchanged since 1992 did not result in models (for X-unsignalised and T-unsignalised junctions) with significant trend variables.

The resulting long-term models for X-signalised and T-signalised junctions are presented in **Sections 6.2.1 and 6.2.2**, respectively.

6.2.1 Long-term models for X-signalised junctions

There were 32 junctions which had non-Poisson accident data sequences during the modelled period (1992-2003). These junctions were therefore removed from the database thus leaving 171 junctions. The data for those junctions were disaggregated, so the number of data that could be formed for one junction in the database equal to the number of available period for the junction. There were thus a total of 1947 annual counts for modelling purposes.

At stage 1, three model forms were examined according to the procedures described in Figure 5.4. It was found that NB was the most appropriate regression model. The results are presented in Table 6.1.

Table 6.1 Long term accident-exposure models for X-signalised junctions

Model 1: $\mu_t = e^{(-9.289-0.133\ln(t)+0.931\ln(Q_t))}$ or $\mu_t = 9.247 \times 10^{-5} (t)^{-0.133} (Q_t)^{0.931}$			
Degrees of freedom	1944		
Pearson X^2	1991.022	$\chi^2_{0.05,1944} = 2048$	
Scaled Deviance G^2	2071.618	$\chi^2_{0.05,1944} = 2048$	
Log-likelihood	-2971.784		
AIC	5949.568		
Variable	Coefficient	t-value	P-value
Constant c	-9.289	-19.360	0.000
$\ln(t)$	-0.133	-3.703	0.000
$\ln(Q_t)$	0.931	20.487	0.000
Dispersion parameter α	0.504	10.777	0.000
Model 2: $\mu_t = e^{(-8.557-0.133\ln(t)+0.466\ln(Q_{pt}Q_{st}))}$ or $\mu_t = 1.921 \times 10^{-4} (t)^{-0.133} (Q_{pt}Q_{st})^{0.466}$			
Degrees of freedom	1944		
Pearson X^2	1978.651	$\chi^2_{0.05,1944} = 2048$	
Scaled Deviance G^2	2061.286	$\chi^2_{0.05,1944} = 2048$	
Log-likelihood	-2942.472		
AIC	5890.944		
Variable	Coefficient	t-value	P-value
Constant c	-8.557	-20.142	0.000
$\ln(t)$	-0.133	-3.744	0.000
$\ln(Q_{pt}Q_{st})$	0.466	21.466	0.000
Dispersion parameter α	0.466	10.396	0.000
Model 3: $\mu_t = e^{(-7.669-0.127\ln(t)+0.193\ln(Q_{pt})+0.662\ln(Q_{st}))}$ or $\mu_t = 4.695 \times 10^{-4} (t)^{-0.127} (Q_{pt})^{0.193} (Q_{st})^{0.662}$			
Degrees of freedom	1943		
Pearson X^2	1962.002	$\chi^2_{0.05,1943} = 2047$	
Scaled Deviance G^2	2053.462	$\chi^2_{0.05,1943} = 2047$	
Log-likelihood	-2932.494		
AIC	5872.988		
Variable	Coefficient	t-value	P-value
Constant c	-7.669	-16.923	0.000
$\ln(t)$	-0.127	-3.576	0.000
$\ln(Q_{pt})$	0.193	3.187	0.001
$\ln(Q_{st})$	0.662	13.379	0.000
Dispersion parameter α	0.456	10.283	0.000

The results show that the parameter estimates for all explanatory variables were all significant at 95% confidence level, as indicated by the $|t\text{-values}|$ of all the parameters being larger than 1.96 or P-values being smaller than 0.05. All models passed the Pearson X^2 goodness-of-fit criterion but not the Scaled Deviance G^2 goodness-of-fit criterion which means that outliers might be present in the database. However, no outliers' analysis was performed for these models since modelling in stage 1 was mainly intended to select the best functional forms of traffic flows that could best describe the interaction of exposure with accident occurrences, with the chosen functional forms being used in stage 2 modelling.

The loglikelihood and AIC values show that model 2 had a relatively similar performance with model 3. Both the models are superior to model 1. Therefore, among the models, both model 2 and model 3 may be considered as the better accident-exposure model for X-signalised junctions. However, model 2 was taken for the next stage of modelling since it had fewer explanatory variables than model 3, which means it is a simpler model.

The modelling was then continued to stage 2. In this stage, two types of long-term accident prediction models were developed based on model 2 with the addition of junction geometric/operational variables as explained previously. Beforehand, the geometric variables to be considered in the models were pre-selected in order to avoid strong correlation among explanatory variables based on correlation analyses which are presented in Appendix E. Multivariate modelling was then carried out using the NB regression model for each model. The junction geometric/operational variables that were found to be significant in the fitted full (long-term) accident prediction model were *CLASS*, *NPHASE*, *SHLANE*, *MED*, *RLCAM*, *YBOX* and *CITY* whereas those that were found to be significant in the fitted simple (long-term) accident prediction model were *CLASS* and *CITY* variables.

Outlier analysis using Cook's distance and leverage statistics were then performed on the fitted models. The graphs of Cook's distance and leverage statistics plotted against the observation number for the full long-term model are given in Figures 6.1

and 6.2, respectively. The plots indicate several potential outliers in terms of Cook's distance and leverage values. Further analysis revealed that 9 observations with the highest Cook's distance were influential outliers which were thus removed from the model.

Figure 6.1 Cook's distances in full long-term X-signalised junctions model

Figure 6.2 Leverage values in full long-term X-signalised junctions model

The graphs of Cook's distance and leverage statistics plotted against the observation number for the simple long-term model are given in Figures 6.3 and 6.4,

respectively. The plots indicate several potential outliers in terms of Cook's distance and leverage values. Further analysis revealed that 17 observations with the highest Cook's distance were influential outliers which were thus removed from the model.

Figure 6.3 Cook's distances in simple long-term X-signalised junctions model

Figure 6.4 Leverage values in simple long-term X-signalised junctions model

The final full long-term model is presented in Table 6.2 while the final simple long-term model is presented in Table 6.3.

Table 6.2 Full long-term accident prediction model for X-signalised junctions

$\mu_t = e^{\left(\begin{matrix} -8.501 - 0.121 \ln(t) + 0.456 \ln(Q_{pt} Q_{st}) + 0.251 CLASS - 0.091 NPHASE + 0.039 SHLANE + 0.305 MED \\ + 0.273 RLCAM + 0.124 YBOX - 0.413 CITY \end{matrix} \right)}$ $\mu_t = 2.032 \times 10^{-4} (Q_{pt} Q_{st})^{0.456} (t)^{-0.121} e^{\left(\begin{matrix} 0.251 CLASS - 0.091 NPHASE + 0.039 SHLANE + 0.305 MED + \\ 0.273 RLCAM + 0.124 YBOX - 0.413 CITY \end{matrix} \right)}$ $Var(\mu)_t = \mu_t + \frac{\mu_t^2}{3.172}$			
Degree of freedom	1928		
Pearson X^2	1797.411	$\chi^2_{0.05,1928} = 2031$	
Scaled Deviance G^2	2028.784	$\chi^2_{0.05,1928} = 2031$	
Log-likelihood	-2810.448		
AIC	5640.896		
Variable	Coefficient	t-value	P-value
Constant c	-8.501	-15.163	0.000
$\ln(t)$	-0.121	-3.395	0.001
$\ln(Q_{pt} Q_{st})$	0.456	14.430	0.000
CLASS	0.251	4.064	0.000
NPHASE	-0.091	-2.162	0.031
SHLANE	0.039	2.520	0.012
MED	0.305	4.275	0.019
RLCAM	0.273	3.958	0.000
YBOX	0.124	2.107	0.035
CITY	-0.413	-5.218	0.000
Dispersion parameter α	0.315	7.920	0.000

Table 6.3 Simple long-term accident prediction model for X-signalised junctions

$\mu_t = e^{\left(\begin{matrix} -8.895 - 0.117 \ln(t) + 0.479 \ln(Q_{pt} Q_{st}) + 0.324 CLASS - 0.555 CITY \end{matrix} \right)}$ $\mu_t = 1.371 \times 10^{-4} (Q_{pt} Q_{st})^{0.479} (t)^{-0.117} e^{\left(\begin{matrix} 0.324 CLASS - 0.555 CITY \end{matrix} \right)}$ $Var(\mu)_t = \mu_t + \frac{\mu_t^2}{3.050}$			
Degree of freedom	1925		
Pearson X^2	1780.681	$\chi^2_{0.05,1925} = 2028$	
Scaled Deviance G^2	2026.095	$\chi^2_{0.05,1925} = 2028$	
Log-likelihood	-2796.203		
AIC	5602.406		
Variable	Coefficient	t-value	P-value
Constant c	-8.895	-16.995	0.000
$\ln(t)$	-0.117	-3.259	0.001
$\ln(Q_{pt} Q_{st})$	0.479	17.142	0.000
CLASS	0.324	5.173	0.000
CITY	-0.555	-9.476	0.000
Dispersion parameter α	0.328	7.953	0.000

The fit of the variance function of the model to the sequenced averages of model squared residuals for the full long-term model and simple long-term model are shown in Figures 6.5 and 6.6, respectively. The averages of squared residuals were clustered around the variance function which indicates a good fitting model.

Figure 6.5 Variance function of full long-term X-signalised junctions model

Figure 6.6 Variance function of simple long-term X-signalised junctions model

6.2.2 Long-term models for T-signalised junctions

Similar models were developed for T-signalised junctions using 1898 annual counts from 168 out of 201 T-signalised junctions. It is noted 33 sites were not included for modelling for the following reasons: 15 junctions did not have traffic data due to faulty GLIDE detectors or were simply not covered by the GLIDE system, which is an intelligent traffic control system for Singapore's road network; 4 junctions did not have entry traffic data on the one-way minor road and 14 junctions were excluded due to their (1992-2003) accident counts not being Poisson-distributed.

At stage 1, three model forms were examined according to the procedure described in Figure 5.4. The NB was found to be the most appropriate regression model. The results of modelling using the NB model are presented in Table 6.4.

The results show that the parameter estimates for all explanatory variables were all significant at 95% confidence level, as indicated by the $|t\text{-values}|$ of all the parameters being larger than 1.96 or P-values being smaller than 0.05. All the models passed the Pearson X^2 and Scaled Deviance G^2 goodness-of-fit criteria. Model 1 had the lowest log-likelihood and the highest AIC value, which suggests that it is not as good as models 2 and 3.

The log-likelihood and AIC values for model 2 and model 3 were relatively similar, which means that choosing either model 2 or model 3 should give relatively similar results. Therefore, both model 2 and model 3 may be considered as the better accident-exposure model for T-signalised junctions. Model 2 and model 3 did have limitation in that when the minor road has zero traffic inflow (one-way minor road), then the accident number cannot be predicted. It is suggested that when such condition happens, change the number of zero flow into one so the problem will disappear. However, model 2 was taken for the next stage of modelling since it had smaller number of explanatory variables than model 3, which means it is a simpler model.

Table 6.4 Long term accident-exposure models for T-signalised junctions

Model 1: $\mu_t = e^{(-9.289-0.133\ln(t)+0.931\ln(Q_t))}$ or $\mu_t = 9.247 \times 10^{-5} (t)^{-0.133} (Q_t)^{0.931}$			
Degrees of freedom	1895	$\chi^2_{0.05,1895} = 1997$ $\chi^2_{0.05,1895} = 1997$	
Pearson X^2	1930.857		
Scaled Deviance G^2	1728.108		
Log-likelihood	-2014.227		
AIC	4034.454		
Variable	Coefficient	t-value	P-value
Constant c	-9.410	-14.928	0.000
$\ln(t)$	-0.210	-4.044	0.000
$\ln(Q_t)$	0.916	14.609	0.000
Dispersion parameter α	0.641	7.158	0.000
Model 2: $\mu_t = e^{(-8.557-0.133\ln(t)+0.466\ln(Q_{pt}Q_{st}))}$ or $\mu_t = 1.921 \times 10^{-4} (t)^{-0.133} (Q_{pt}Q_{st})^{0.466}$			
Degrees of freedom	1895	$\chi^2_{0.05,1895} = 1997$ $\chi^2_{0.05,1895} = 1997$	
Pearson X^2	1912.515		
Scaled Deviance G^2	1702.963		
Log-likelihood	-2001.469		
AIC	4008.939		
Variable	Coefficient	t-value	P-value
Constant c	-8.923	-15.790	0.000
$\ln(t)$	-0.205	-4.161	0.000
$\ln(Q_{pt}Q_{st})$	0.480	15.433	0.000
Dispersion parameter α	0.590	6.729	0.000
Model 3: $\mu_t = e^{(-7.669-0.127\ln(t)+0.193\ln(Q_{pt})+0.662\ln(Q_{st}))}$ or $\mu_t = 4.695 \times 10^{-4} (t)^{-0.127} (Q_{pt})^{0.193} (Q_{st})^{0.662}$			
Degrees of freedom	1894	$\chi^2_{0.05,1894} = 1996$ $\chi^2_{0.05,1894} = 1996$	
Pearson X^2	1918.843		
Scaled Deviance G^2	1701.959		
Log-likelihood	-2000.963		
AIC	4009.927		
Variable	Coefficient	t-value	P-value
Constant c	-9.069	-15.518	0.000
$\ln(t)$	-0.205	-4.170	0.000
$\ln(Q_{pt})$	0.535	8.458	0.001
$\ln(Q_{st})$	0.433	7.700	0.000
Dispersion parameter α	0.590	6.693	0.000

The modelling was then continued to stage 2. In this stage, two types of long-term accident prediction model were developed based on model 2 with the addition of junction geometric/operational variables as explained previously. Multivariate modelling was then carried out using the NB regression model for each model. The junction geometric/operational variables that were found to be significant in the fitted full long-term accident prediction model were *CLASS*, *SHLANE*, *UTURN*, *RLCAM*, *YBOX* and *CITY* whereas those that were found to be significant in the fitted simple long-term accident prediction model were *CLASS* and *CITY* variables.

Outlier analysis using Cook’s distance and leverage statistics were then performed on the fitted models. The graphs of Cook’s distance and leverage statistics plotted against the observation number for the full long-term model are given in Figures 6.7 and 6.8, respectively. The plots indicate several potential outliers in terms of Cook’s distance and leverage values. Further analysis revealed that 8 observations with the highest Cook’s distance were influential outliers which were thus removed from the model.

Figure 6.7 Cook’s distances in full long-term T-signalised junctions model

Figure 6.8 Leverage values in simple long-term T-signalised junctions model

The graphs of Cook’s distance and leverage statistics plotted against the observation number for the simple long-term model are given in Figures 6.9 and 6.10, respectively. The plots indicate several potential outliers in terms of Cook’s distance and leverage values. Further analysis revealed that 6 observations with the highest Cook’s distance were influential outliers which were thus removed from the model.

Figure 6.9 Cook’s distances in simple long-term T-signalised junctions model

Figure 6.10 Leverage values in simple long-term T-signalised junctions model

The final full long-term model and simple long-term model are presented in Tables 6.5 and 6.6, respectively.

Table 6.5 Full long-term accident prediction model for T-signalised junctions

$\mu_t = e^{\left(\begin{matrix} -6.670 - 0.158 \ln(t) + 0.328 \ln(Q_{pt} Q_{st}) + 0.564 CLASS + 0.125 SHLANE + 0.317 UTURN + 0.472 RLCAM \\ + 0.227 YBOX - 2.213 CITY \end{matrix} \right)}$ $\mu_t = 1.268 \times 10^{-3} (Q_{pt} Q_{st})^{0.328} (t)^{-0.158} e^{\left(\begin{matrix} 0.564 CLASS + 0.125 SHLANE + 0.317 UTURN + 0.472 RLCAM \\ + 0.227 YBOX - 2.213 CITY \end{matrix} \right)}$ $Var(\mu)_t = \mu_t + \frac{\mu_t^2}{4.413}$			
Degree of freedom	1881		
Pearson X^2	1964.955	$\chi^2_{0.05, 1881} = 1983$	
Scaled Deviance G^2	1778.326	$\chi^2_{0.05, 1881} = 1983$	
Log-likelihood	-1867.407		
AIC	3752.814		
Variable	Coefficient	t-value	P-value
Constant c	-6.670	-10.483	0.000
$\ln(t)$	-0.158	-3.436	0.001
$\ln(Q_{pt} Q_{st})$	0.328	9.338	0.000
CLASS	0.564	6.606	0.000
SHLANE	0.125	3.413	0.000
UTURN	0.317	3.767	0.000
RLCAM	0.472	5.439	0.000
YBOX	0.227	3.301	0.000
CITY	-2.213	-6.234	0.000
Dispersion parameter α	0.227	3.361	0.000

Table 6.6 Simple long-term accident prediction model for T-signalised junctions

$\mu_t = e^{(-7.325 - 0.190 \ln(t) + 0.387 \ln(Q_{pt}Q_{st}) + 0.597CLASS - 2.656CITY)}$ $\mu_t = 6.587 \times 10^{-4} (Q_{pt}Q_{st})^{0.387} (t)^{-0.190} e^{(0.597CLASS - 2.656CITY)}$ $Var(\mu)_t = \mu_t + \frac{\mu_t^2}{2.605}$			
Degree of freedom	1893		
Pearson X^2	1880.389	$\chi^2_{0.05,1893} = 1995$	
Scaled Deviance G^2	1727.746	$\chi^2_{0.05,1893} = 1995$	
Log-likelihood	-1915.686		
AIC	3841.372		
Variable	Coefficient	t-value	P-value
Constant c	-7.325	-11.990	0.000
$\ln(t)$	-0.190	-4.068	0.001
$\ln(Q_{pt}Q_{st})$	0.387	11.396	0.000
CLASS	0.597	7.132	0.000
CITY	-2.656	-6.576	0.000
Dispersion parameter α	0.384	5.159	0.000

The fit of the variance function of the model to the sequenced averages of model squared residuals for the full long-term model is shown in Figure 6.11 while the fit for the simple long-term model is shown in Figure 6.12.

Figure 6.11 Variance function of full long-term T-signalised junctions model

Figure 6.12 Variance function of simple long-term T-signalised junctions model

The fit for the full long-term model was good at lower accident frequency but as predicted accident frequency increases, the sequenced averages of model squared residuals were lower than the values of the variance functions. The fit for the simple long-term model was also good at lower accident frequency but as predicted accident frequency increases, the sequenced averages of model squared residuals were first higher and then lower than the values of the variance functions. This may be due to limited datapoints that were available for higher predicted accident frequency.

6.3 Short-term models

The short-term accident prediction models were developed based on time-aggregated accident counts data. Several short-term models were developed based on the duration of modelled period, which were between 1 year and 5 years. They were named as the short-term (5-year) models (developed based on 1999-2003 period), short-term (4-year) models (developed based on 2000-2003 period), short-term (3-year) models (developed based on 2001-2003 period), short-term (2-year) models (developed based on 2002-2003 period), and short-term (1-year) models (developed based on 2003 modelled year).

Similar to the long-term models, each of the short-term models for signalised junctions was developed in two stages, which were accident-exposure models (stage 1) and full models (stage 2). The full models were the accident prediction models which also contained junction geometric/operational variables on top of exposure variable. For unsignalised junctions, the models were developed in one stage only due to the unavailability of traffic flow data.

At stage 1, the assumption of constant accident rate for the short-term models were firstly verified. No statistically significant trend variable was found for X-signalised and T-signalised models, which confirmed constant accident rate condition during 1999-2003 period. Likewise, the trend variable for X-unsignalised model was also found to be not significant. However, the assumption of constant accident rate could not be verified for T-unsignalised model during 1999-2003 period since significant trend variable at 95% confidence level was found for the model. Therefore, the short-term model which incorporated trend based on 5-year data was also developed for T-unsignalised junction besides its short-term models without trend.

The geometric variables to be included in the models were pre-selected in order to avoid strong correlation between explanatory variables based on correlation analyses which are presented in Appendix E. Multivariate modelling was then carried out on the traffic flows and the chosen geometric/operational variables. The resulting short-term models for X-signalised, T-signalised, X-unsignalised and T-unsignalised junctions are presented in **Sections 6.3.1 – 6.3.4**, respectively.

6.3.1 Short-term models for X-signalised junctions

The development of the short-term models for X-signalised junctions was initiated by removing junctions with non-Poisson accident count sequence during the modelled period from the initial database for each kind of model. In this case, for the short-term (5-year) model for X-signalised junctions, there were 32 junctions which had non-Poisson accident data sequences during the modelled period (1999-2003). These junctions were therefore removed from the database thus leaving 171

junctions. At stage 1, three model forms were examined according to the procedures described in Figure 5.4. It was found that NB was the most appropriate regression model. The results are presented in Table 6.7.

Table 6.7 Short-term (5-year) accident-exposure models for X-signalised junctions

Model 1: $\mu = e^{(-8.765-1.005\ln(Q_t))}$ or $\mu = 1.561 \times 10^{-4} (Q)^{1.005}$			
Degrees of freedom	169		
Pearson X^2	184.399	$\chi^2_{0.05,169} = 200.3$	
Scaled Deviance G^2	196.805	$\chi^2_{0.05,169} = 200.3$	
Log-likelihood	-491.176		
AIC	986.351		
Variable	Coefficient	t-value	P-value
Constant c	-8.765	-7.411	0.000
$\ln(Q)$	1.005	8.962	0.000
Dispersion parameter α	0.636	7.721	0.000
Model 2: $\mu = e^{(-8.200+0.517\ln(Q_pQ_s))}$ or $\mu = 2.746 \times 10^{-4} (Q_p, Q_{s1})^{0.517}$			
Degrees of freedom	169		
Pearson X^2	184.099	$\chi^2_{0.05,169} = 200.3$	
Scaled Deviance G^2	196.134	$\chi^2_{0.05,169} = 200.3$	
Log-likelihood	-484.516		
AIC	973.833		
Variable	Coefficient	t-value	P-value
Constant c	-8.200	-7.919	0.000
$\ln(Q_pQ_s)$	0.517	9.700	0.000
Dispersion parameter α	0.578	7.583	0.000
Model 3: $\mu = e^{(-7.082+0.169\ln(Q_p)+0.770\ln(Q_s))}$ or $\mu = 8.728 \times 10^{-4} (Q_p)^{0.162} (Q_s)^{0.774}$			
Degrees of freedom	168		
Pearson X^2	181.633	$\chi^2_{0.05,168} = 199.2$	
Scaled Deviance G^2	195.524	$\chi^2_{0.05,168} = 199.2$	
Log-likelihood	-481.942		
AIC	969.884		
Variable	Coefficient	t-value	P-value
Constant c	-7.044	-6.283	0.000
$\ln(Q_p)$	0.162	1.076	0.282
$\ln(Q_s)$	0.774	6.152	0.000
Dispersion parameter α	0.553	7.585	0.000

The results show that model 1 and model 2 had all their parameter estimates significant at 95% confidence level, as indicated by the |t-values| of all the parameters being larger than 1.96 or P-values being smaller than 0.05 whereas the

parameter estimate for $\ln(Q_p)$ in model 3 was not significant at 95% confidence level. Therefore model 3 was not considered further. The Pearson X^2 and Scaled Deviance G^2 tests show that both model 1 and model 2 passed the goodness-of-fit criterion which means that model 1 is as good as model 2, although the goodness-of-fit criterion for model 2 is slightly better. The loglikelihood and AIC values show that model 2 is superior to model 1. Therefore, among the models, model 2 may be considered as the best accident-exposure model for X-signalised junctions.

The modelling was then continued to stage 2, where the accident prediction model was developed based on model 2 with addition of junction geometric/operational variables. Multivariate modelling was carried out using the NB regression model. It was found that *CLASS* and *CITY* were the only significant junction geometric/operational variables. Outlier analysis using Cook's distance and leverage statistics were then performed on the fitted model. The graphs of Cook's distance and leverage statistics plotted against the observation number are given in Figures 6.13 and 6.14, respectively. The plots indicate several potential outliers in terms of Cook's distance and leverage values. Further analysis revealed that 2 observations with the highest Cook's distance were influential outliers which were thus removed from the model.

Figure 6.13 Cook's distances in short-term (5-year) X-signalised junctions model

Figure 6.14 Leverage values in short-term (5-year) X-signalised junctions model

The final short-term (5-year) model is presented in Table 6.8 and the fit of the variance function of the model to the sequenced averages of model squared residuals is shown in Figure 6.15. The averages of squared residuals were clustered around the variance function which indicates a good fitting model.

Table 6.8 Short-term (5-year) accident prediction model for X-signalised junctions

$\mu = e^{(-6.779 + 0.434 \ln(Q_p Q_s) + 0.624 CLASS - 0.649 CITY)}$ $\mu = 1.137 \times 10^{-3} (Q_p Q_s)^{0.434} e^{(0.624 CLASS - 0.649 CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{2.404}$			
Degree of freedom	165		
Pearson χ^2	148.424	$\chi^2_{0.05, 165} = 196$	
Scaled Deviance G^2	193.253	$\chi^2_{0.05, 165} = 196$	
Log-likelihood	-470.679		
AIC	949.357		
Variable	Coefficient	t-value	P-value
Constant c	-6.779	-5.290	0.000
$\ln(Q_p Q_s)$	0.434	6.346	0.000
CLASS	0.624	4.059	0.000
CITY	-0.649	-3.902	0.000
Dispersion parameter α	0.416	6.646	0.000

Figure 6.15 Variance function of short-term (5-year) X-signalised junctions model

The other short-term models for X-signalised junctions were subsequently developed using similar procedure used to develop the short-term (5-year) model. The resulting final short-term (4-year), short-term (3-year), short-term (2-year) and short-term (1-year) models for X-signalised junctions are presented in Tables 6.9 to 6.12, respectively.

Table 6.9 Short-term (4-year) accident prediction model for X-signalised junctions

$\mu = e^{(-8.359 + 0.507 \ln(Q_p Q_s) + 0.441 CLASS - 0.645 CITY)}$ $\mu = 2.342 \times 10^{-4} (Q_p Q_s)^{0.507} e^{(0.441 CLASS - 0.645 CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{3.806}$			
Degree of freedom	189		
Pearson χ^2	163.346	$\chi^2_{0.05, 189} = 222.1$	
Scaled Deviance G^2	215.858	$\chi^2_{0.05, 189} = 222.1$	
Log-likelihood	-476.195		
AIC	960.39		
Variable	Coefficient	t-value	P-value
Constant c	-8.359	-7.323	0.000
$\ln(Q_p Q_s)$	0.507	8.414	0.000
CLASS	0.441	3.564	0.000
CITY	-0.645	-4.706	0.000
Dispersion parameter α	0.263	5.298	0.000

Table 6.10 Short-term (3-year) accident prediction model for X-signalised junctions

$\mu = e^{(-7.695 + 0.455 \ln(Q_p Q_s) + 0.503 CLASS - 0.616 CITY)}$ $\mu = 4.552 \times 10^{-4} (Q_p Q_s)^{0.455} e^{(0.503 CLASS - 0.616 CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{4.278}$			
Degree of freedom	193		
Pearson X^2	173.136	$\chi^2_{0.05,193} = 226.4$	
Scaled Deviance G^2	223.237	$\chi^2_{0.05,193} = 226.4$	
Log-likelihood	-437.907		
AIC	883.815		
Variable	Coefficient	t-value	P-value
Constant c	-7.695	-6.612	0.000
$\ln(Q_p Q_s)$	0.455	7.455	0.000
CLASS	0.503	3.795	0.000
CITY	-0.616	-4.514	0.000
Dispersion parameter α	0.234	4.183	0.000

Table 6.11 Short-term (2-year) accident prediction model for X-signalised junctions

$\mu = e^{(-8.457 + 0.475 \ln(Q_p Q_s) + 0.450 CLASS - 0.650 CITY)}$ $\mu = 2.125 \times 10^{-4} (Q_p Q_s)^{0.475} e^{(0.450 CLASS - 0.650 CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{4.775}$			
Degree of freedom	195		
Pearson X^2	185.493	$\chi^2_{0.05,195} = 228.6$	
Scaled Deviance G^2	224.258	$\chi^2_{0.05,195} = 228.6$	
Log-likelihood	-393.189		
AIC	794.378		
Variable	Coefficient	t-value	P-value
Constant c	-8.457	-6.57712	0.000
$\ln(Q_p Q_s)$	0.475	7.11373	0.000
CLASS	0.450	3.137	0.002
CITY	-0.650	-4.526	0.000
Dispersion parameter α	0.209	3.300	0.000

Table 6.12 Short-term (1-year) accident prediction model for X-signalised junctions

$\mu = e^{(-9.111+0.475\ln(Q_pQ_s)+0.572CLASS-0.683CITY)}$ $\mu = 1.104 \times 10^{-4} (Q_pQ_s)^{0.475} e^{(0.572CLASS-0.683CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{4.064}$			
Degree of freedom	197		
Pearson χ^2	179.774	$\chi^2_{0.05,197} = 230.7$	
Scaled Deviance G^2	214.438	$\chi^2_{0.05,197} = 230.7$	
Log-likelihood	-303.518		
AIC	611.037		
Variable	Coefficient	t-value	P-value
Constant c	-9.111	-5.707	0.000
$\ln(Q_pQ_s)$	0.475	5.729	0.000
CLASS	0.572	3.007	0.003
CITY	-0.683	-4.006	0.000
Dispersion parameter α	0.246	2.342	0.019

The fit of the variance function of the model to the sequenced averages of model squared residuals for those models are shown in Figures 6.16 to 6.19. The figures show that the short-term (4-year), short-term (3-year), short-term (2-year) and short-term (1-year) models had good fits since the averages of squared residuals for each kind of model were clustered around the variance function

6.3.2 Short-term models for T-signalised junctions

The initial T-signalised junctions’ database contained dataset from 201 junctions. However, 15 junctions were excluded since they did not have traffic data due to faulty GLIDE detector or were simply not covered by the GLIDE system and 4 more junctions were also excluded since they did not have entry traffic data on the one-way minor road. The development of the short-term models for T-signalised junctions was then followed by removing junctions with non-Poisson accident count sequence during the modelled period from the initial database for each kind of model. In this case, for the short-term (5-year) model for T-signalised junctions, there were 10 junctions with non-Poisson accident count sequence during the modelled period (1999-2003).

Figure 6.16 Variance function of short-term (4-year) X-signalised junctions model

Figure 6.18 Variance function of short-term (2-year) X-signalised junctions model

Figure 6.17 Variance function of short-term (3-year) X-signalised junctions model

Figure 6.19 Variance function of short-term (1-year) X-signalised junctions model

At stage 1, three model forms were examined according to the procedure described in Figure 5.4. The NB was found to be the most appropriate regression model. The results of modelling using the NB model are presented in Table 6.13.

Table 6.13 Short-term (5-year) accident-exposure models for T-signalised junctions

Model 1: $\mu = e^{(-8.845 - 0.973 \ln(Q_t))}$ or $\mu = 1.441 \times 10^{-4} (Q)^{0.973}$			
Degrees of freedom	169	$\chi^2_{0.05,169} = 200.3$ $\chi^2_{0.05,169} = 200.3$	
Pearson X^2	215.474		
Scaled Deviance G^2	189.974		
Log-likelihood	-366.963		
AIC	737.925		
Variable	Coefficient	t-value	P-value
Constant c	-8.845	-8.165	0.000
$\ln(Q)$	0.973	9.065	0.000
Dispersion parameter α	0.865	6.276	0.000
Model 2: $\mu = e^{(-8.920 + 0.543 \ln(Q_p Q_s))}$ or $\mu = 1.336 \times 10^{-4} (Q_p Q_s)^{0.543}$			
Degrees of freedom	169	$\chi^2_{0.05,169} = 200.3$ $\chi^2_{0.05,169} = 200.3$	
Pearson X^2	199.666		
Scaled Deviance G^2	189.005		
Log-likelihood	-363.588		
AIC	731.176		
Variable	Coefficient	t-value	P-value
Constant c	-8.920	-8.431	0.000
$\ln(Q_p Q_s)$	0.543	9.377	0.000
Dispersion parameter α	0.805	6.012	0.000
Model 3: $\mu = e^{(-8.967 + 0.574 \ln(Q_p) + 0.511 \ln(Q_s))}$ or $\mu = 1.276 \times 10^{-4} (Q_p)^{0.574} (Q_s)^{0.511}$			
Degrees of freedom	168	$\chi^2_{0.05,168} = 199.2$ $\chi^2_{0.05,168} = 199.2$	
Pearson X^2	202.524		
Scaled Deviance G^2	190.891		
Log-likelihood	-363.555		
AIC	733.111		
Variable	Coefficient	t-value	P-value
Constant c	-8.967	-8.467	0.000
$\ln(Q_p)$	0.574	4.693	0.000
$\ln(Q_s)$	0.511	3.402	0.000
Dispersion parameter α	0.804	5.871	0.000

The results show that the parameter estimates for all explanatory variables were all significant at 95% confidence level, as indicated by the |t-values| of all the parameters being larger than 1.96. Only model 2 passed the Pearson X^2 and the

Scaled Deviance G^2 goodness-of-fit criteria. Model 1 and model 3 did not pass the Pearson X^2 goodness-of-fit-criterion. Therefore, model 2 may be considered as the best accident-exposure model for T-signalised junctions. The modelling was then continued to stage 2, where the accident prediction model was developed based on model 2 with addition of junction geometric/operational variables. Multivariate modelling was then carried out using the NB regression model. It was found that *CLASS* and *CITY* were the only significant junction geometric/operational variables.

Outlier analysis using Cook's distance and leverage statistics were then performed on the fitted model. The graphs of Cook's distance and leverage statistics plotted against the observation number are given in Figures 6.20 and 6.21, respectively. The plots indicate several potential outliers in terms of Cook's distance and leverage values. Further analysis revealed that 1 observation with the highest Cook's distance was an influential outlier which was thus removed from the model.

Figure 6.20 Cook's distances in short-term (5-year) T-signalised junctions model

Figure 6.21 Leverage values in short-term (5-year) T-signalised junctions model

The final model is presented in Table 6.14 and the fit of the variance function of the model to the sequenced averages of model squared residuals is shown in Figure 6.22. The averages of squared residuals were clustered around the variance function which indicates a good fitting model.

Table 6.14 Short-term (5-year) accident prediction model for T-signalised junctions

$\mu = e^{(-8.676 + 0.523 \ln(Q_p Q_s) + 0.601 CLASS - 1.496 CITY)}$ $\mu = 1.706 \times 10^{-4} (Q_p Q_s)^{0.523} e^{(0.601 CLASS - 1.496 CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{1.624}$			
Degree of freedom	167		
Pearson χ^2	167.513	$\chi^2_{0.05, 167} = 198.2$	
Scaled Deviance G^2	178.399	$\chi^2_{0.05, 167} = 198.2$	
Log-likelihood	-347.638		
AIC	703.276		
Variable	Coefficient	t-value	P-value
Constant c	-8.676	-5.311	0.000
$\ln(Q_p Q_s)$	0.523	5.869	0.000
CLASS	0.601	2.809	0.005
CITY	-1.496	-3.150	0.002
Dispersion parameter α	0.616	5.357	0.000

Figure 6.22 Variance function of short-term (5-year) T-signalised junctions model

The other short-term models for T-signalised junctions were subsequently developed using similar procedure used to develop the short-term (5-year) model. The resulting final short-term (4-year), short-term (3-year), short-term (2-year) and short-term (1-year) models for T-signalised junctions are presented in Tables 6.15 to 6.18, respectively.

Table 6.15 Short-term (4-year) accident prediction model for T-signalised junctions

$\mu = e^{(-9.077 + 0.534 \ln(Q_p Q_s) + 0.604 CLASS - 2.096 CITY)}$ $\mu = 1.142 \times 10^{-4} (Q_p Q_s)^{0.534} e^{(0.604 CLASS - 2.096 CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{1.846}$			
Degree of freedom	176		
Pearson X^2	171.955	$\chi^2_{0.05, 176} = 208$	
Scaled Deviance G^2	194.714	$\chi^2_{0.05, 176} = 208$	
Log-likelihood	-339.606		
AIC	687.213		
Variable	Coefficient	t-value	P-value
Constant c	-9.077	-5.904	0.000
$\ln(Q_p Q_s)$	0.534	6.390	0.000
CLASS	0.604	2.863	0.004
CITY	-2.096	-3.116	0.002
Dispersion parameter α	0.542	5.034	0.000

Table 6.16 Short-term (3-year) accident prediction model for T-signalised junctions

$\mu = e^{(-8.693+0.495\ln(Q_pQ_s)+0.674CLASS-2.715CITY)}$ $\mu = 1.678 \times 10^{-4} (Q_pQ_s)^{0.495} e^{(0.674CLASS-2.715CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{2.153}$			
Degree of freedom	178		
Pearson X^2	177.547	$\chi^2_{0.05,178} = 210.1$	
Scaled Deviance G^2	186.529	$\chi^2_{0.05,178} = 210.1$	
Log-likelihood	-296.092		
AIC	600.183		
Variable	Coefficient	t-value	P-value
Constant c	-8.693	-5.351	0.000
$\ln(Q_pQ_s)$	0.495	5.609	0.000
CLASS	0.674	3.465	0.000
CITY	-2.715	-2.973	0.003
Dispersion parameter α	0.464	4.042	0.000

Table 6.17 Short-term (2-year) accident prediction model for T-signalised junctions

$\mu = e^{(-9.281+0.501\ln(Q_pQ_s)+0.713CLASS-2.263CITY)}$ $\mu = 9.314 \times 10^{-5} (Q_pQ_s)^{0.501} e^{(0.713CLASS-2.263CITY)}$ $Var(\mu) = \mu + \frac{\mu^2}{2.562}$			
Degree of freedom	178		
Pearson X^2	185.169	$\chi^2_{0.05,178} = 210.1$	
Scaled Deviance G^2	179.760	$\chi^2_{0.05,178} = 210.1$	
Log-likelihood	-239.055		
AIC	486.110		
Variable	Coefficient	t-value	P-value
Constant c	-9.281	-4.810	0.000
$\ln(Q_pQ_s)$	0.501	4.768	0.000
CLASS	0.713	3.122	0.002
CITY	-2.263	-2.502	0.012
Dispersion parameter α	0.390	2.497	0.013

Table 6.18 Short-term (1-year) accident prediction model for T-signalised junctions

$\mu = e^{(-10.772+0.550\ln(Q_pQ_s))}$ $\mu = 2.099 \times 10^{-5} (Q_pQ_s)^{0.550}$ $Var(\mu) = \mu + \frac{\mu^2}{1.082}$			
Degree of freedom	180		
Pearson χ^2	206.417	$\chi^2_{0.05,180} = 212.3$	
Scaled Deviance G^2	149.710	$\chi^2_{0.05,180} = 212.3$	
Log-likelihood	-176.016		
AIC	360.031		
Variable	Coefficient	t-value	P-value
Constant c	-10.772	-5.708	0.000
$\ln(Q_pQ_s)$	0.550	5.407	0.000
Dispersion parameter α	0.924	2.307	0.021

The fit of the variance function of the model to the sequenced averages of model squared residuals for these models are shown in Figures 6.23 to 6.26. It can be seen in Figures 6.23 to 6.26 that the fits of the sequenced averages of model squared residuals to the variance functions for the short-term (4-year), short-term (3-year), short-term (2-year) and short-term (1-year) models were good at lower accident frequency. However, at higher predicted accident frequency, the sequenced averages of model squared residuals were lower than the values of the variance functions. This may be due to limited datapoints that were available for higher predicted accident frequency. Out of the short-term models developed for T-signalised junctions, the fits for the short-term (3-year) and short-term (5-year) models seemed to be better than the fits of other models.

Figure 6.23 Variance function of short-term (4-year) T-signalised junctions model

Figure 6.25 Variance function of short-term (2-year) T-signalised junctions model

Figure 6.24 Variance function of short-term (3-year) T-signalised junctions model

Figure 6.26 Variance function of short-term (1-year) T-signalised junctions model

6.3.3 Short-term model for X-unsignalised junctions

Since no traffic exposure was available for unsignalised junctions, the model was developed based on geometric/operational variables only. The procedures illustrated in Figure 5.4 were used to develop the accident prediction model combined with the backward elimination procedure to select the geometric variables for the model. The models for X-unsignalised junctions were developed using the geometric/operational variables using 71 samples from 71 X-unsignalised junctions. Prior to this, junctions with non-Poisson accident count sequence during the modelled period were removed. In this case, 2 junctions were removed from the database for the short-term (5-year) model. It was found that the best short-term (5-year) model that can be obtained is an NB regression model containing *NAPLN* and *RTLANE* only.

Outlier analysis using Cook’s distance and leverage statistics were performed on the fitted model. The graphs of Cook’s distance and leverage statistics plotted against the observation number are given in Figures 6.27 and 6.28, respectively. The plots indicate several potential outliers in terms of Cook’s distance. Further analysis did not reveal any influential outlier in the sample.

Figure 6.27 Cook’s distances in short-term (5-year) X-unsignalised junctions model

Figure 6.28 Leverage values in short-term (5-year) X-unsignalised junctions model

The final short-term (5-year) model for X-unsignalised junctions is presented in Table 6.19. The Scaled Deviance G^2 and Pearson X^2 values of the model were both significant at the 95% confidence level, which indicate an acceptable fit in NB regression model.

Table 6.19 Short-term (5-year) accident prediction model for X-unsignalised junctions

$\mu = e^{(-2.809 + 0.572 \text{ NAPLN} - 1.272 \text{ RTLANE})}$ $\mu = 0.060 \times e^{(0.572 \text{ NAPLN} - 1.272 \text{ RTLANE})}$ $\text{Var}(\mu) = \mu + \frac{\mu^2}{1.100}$			
Degree of freedom	67	$\chi^2_{0.05,67} = 87.108$	
Pearson X^2	67.531		
Scaled Deviance G^2	62.616		
Log-likelihood	-89.829	$\chi^2_{0.05,67} = 87.108$	
AIC	185.658		
Variable	Coefficient	t-value	P-value
Constant c	-2.809	-3.392	0.000
NAPLN	0.572	3.506	0.000
RTLANE	-1.272	-2.258	0.024
Dispersion parameter α	0.909	2.186	0.029

Figure 6.29 shows the fit of the variance function of the model to the sequenced averages of model squared residuals. The fit of averages of squared residuals to the variance function was good at lower predicted accident frequency but as predicted accident frequency increases, the sequenced averages of model squared residuals were first higher and then lower than the values of the variance function. This may be due to limited samples that were available for higher predicted accident counts.

Figure 6.29 Variance function of short-term (5-year) X-unsignalised junctions model

The other short-term models for X-unsignalised junctions were subsequently developed using similar procedure used to develop the short-term (5-year) model. However, no well-fitting short-term (1-year) model and short-term (2-year) model were found. The short-term (3-year) and short-term (4-year) models were developed using Poisson regression model. The resulting short-term (4-year) and short-term (3-year) models are presented in Tables 6.20 to 6.21, respectively. The fit of the variance function of the model to the sequenced averages of model squared residuals for those models are shown in Figures 6.30 to 6.31. The fits were rather erratic, which may be due to limited samples available for X-unsignalised junctions model.

Table 6.20 Short-term (4-year) accident prediction model
for X-unsignalised junctions

$\mu = e^{(-3.5439+0.622\text{ NAPLN}-1.408\text{ RTLANE})}$ or $\mu = 0.029 \times e^{(0.622\text{ NAPLN}-1.408\text{ RTLANE})}$ $Var(\mu) = \mu$			
Degree of freedom	63		
Pearson X^2	71.541	$\chi^2_{0.05,63} = 82.53$	
Scaled Deviance G^2	72.980	$\chi^2_{0.05,63} = 82.53$	
Log-likelihood	-68.102		
AIC	142.204		
Variable	Coefficient	t-value	P-value
Constant c	-3.543	-6.789	0.000
NAPLN	0.622	8.189	0.000
RTLANE	-1.408	-3.928	0.000

Table 6.21 Short-term (3-year) accident prediction model
for X-unsignalised junctions

$\mu = e^{(-3.201+0.546\text{ NAPLN}-1.226\text{ RTLANE})}$ or $\mu = 0.041 \times e^{(0.546\text{ NAPLN}-1.226\text{ RTLANE})}$ $Var(\mu) = \mu$			
Degree of freedom	66		
Pearson X^2	83.631	$\chi^2_{0.05,66} = 85.96$	
Scaled Deviance G^2	83.330	$\chi^2_{0.05,66} = 85.96$	
Log-likelihood	-71.550		
AIC	149.100		
Variable	Coefficient	t-value	P-value
Constant c	-3.201	-6.250	0.000
NAPLN	0.546	7.045	0.000
RTLANE	-1.226	-3.378	0.000

Figure 6.30 Variance function of short-term (4-year)
X-unsignalised junctions model

Figure 6.31 Variance function of short-term (3-year) X-unsignalised junctions model

6.3.4 Short-term model for T-unsignalised junctions

The short-term models without trend

Similar to the X-unsignalised junctions model, the models for T-unsignalised junctions were developed based on geometric/operational variables only due to unavailability of traffic exposure data. The procedures illustrated in Figure 5.4 were used to develop the accident prediction model combined with the backward elimination procedure to select the geometric variables for the model. The model was developed using the geometric/operational variables for 557 samples from 557 T-unsignalised junctions. Prior to this, junctions that had non-Poisson accident count sequence over the modelled period were removed from the database for each kind of model. In this case, 4 junctions were removed from the database of the short-term (5-year) model. It was found that the best short-term (5-year) model that can be obtained is an NB regression model containing *NAPLN*, *SHLANE*, *NOMED*, *YBOX* and *TOWN*.

Outlier analysis using Cook’s distance and leverage statistics was performed on the fitted model. The graphs of Cook’s distance and leverage statistics plotted against the observation number are shown in Figures 6.32 and 6.33, respectively. The plots indicate several potential outliers in terms of Cook’s distance. Further analysis

revealed that 7 influential outliers with the highest Cook’s distance existed in the sample which were thus removed from the database.

Figure 6.32 Cook’s distances in T-unsignalised junctions model

Figure 6.33 Leverage values in T-unsignalised junctions model

The final short-term (5-year) model for T-unsignalised junctions is presented in Table 6.22 and the fit of the variance function of the model to the sequenced averages of model squared residuals is shown in Figure 6.34.

Table 6.22 Short-term (5-year) accident prediction model for T-unsignalised junctions

$\mu = e^{(-3.032+0.306NAPLN+0.337SHLANE-0.512LCON+0.721YBOX+0.547TOWN)}$ $\mu = 0.048 \times e^{(0.306NAPLN+0.337SHLANE-0.512LCON+0.721YBOX+0.547TOWN)}$ $Var(\mu) = \mu + \frac{\mu^2}{0.933}$			
Degree of freedom	541		
Pearson X^2	519.681	$\chi^2_{0.05,541} = 596.2$	
Scaled Deviance G^2	411.079	$\chi^2_{0.05,541} = 596.2$	
Log-likelihood	-486.854		
AIC	985.710		
Variable	Coefficient	t-value	P-value
Constant c	-3.032	-8.106	0.000
$NAPLN$	0.306	4.797	0.000
$SHLANE$	0.337	3.479	0.000
$LCON$	-0.512	-2.279	0.023
$YBOX$	0.721	4.215	0.000
$TOWN$	0.547	3.1691	0.002
Dispersion parameter α	1.072	4.081	0.000

Figure 6.34 Variance function of short-term (5-year) T-unsignalised junctions model

The Scaled Deviance G^2 and Pearson X^2 values of the model were both significant at the 95% confidence level, which show an acceptable fit in NB regression model. The averages of squared residuals were generally clustered around the variance function which indicates a good fitting model although for

higher predicted accident frequency the averages of squared residuals were lower than the variance function. This may be due to limited samples that were available for higher predicted accident counts.

The other short-term models for T-unsignalised junctions were subsequently developed using similar procedure used to develop the short-term (5-year) model. The resulting short-term (4-year), short-term (3-year), short-term (2-year) and short-term (1-year) models for T-unsignalised junctions are presented in Tables 6.23 to 6.26, respectively.

Table 6.23 Short-term (4-year) accident prediction model for T-unsignalised junctions

$\mu = e^{(-2.999+0.298\,NAPLN+0.323\,SHLANE-0.726\,LCON+0.559\,YBOX+0.586\,TOWN)}$ $\mu = 0.050 \times e^{(0.298\,NAPLN+0.323\,SHLANE-0.726\,LCON+0.559\,YBOX+0.586\,TOWN)}$ $Var(\mu) = \mu + \frac{\mu^2}{0.760}$			
Degree of freedom	550	$\chi^2_{0.05,550} = 605.7$ $\chi^2_{0.05,550} = 605.7$	
Pearson X^2	580.054		
Scaled Deviance G^2	446.198		
Log-likelihood	-457.875		
AIC	925.750		
Variable	Coefficient	t-value	P-value
Constant c	-2.999	-8.036	0.000
$NAPLN$	0.298	4.466	0.000
$SHLANE$	0.323	3.448	0.000
$LCON$	-0.726	-3.221	0.001
$YBOX$	0.559	2.875	0.004
$TOWN$	0.586	3.457	0.000
Dispersion parameter α	1.315	4.633	0.000

Table 6.24 Short-term (3-year) accident prediction model for
T-unsignalised junctions

$\mu = e^{(-3.294+0.231NAPLN+0.334SHLANE-0.906LCON+0.676YBOX+0.754TOWN)}$ $\mu = 0.037 \times e^{(0.231NAPLN+0.334SHLANE-0.906LCON+0.676YBOX+0.754TOWN)}$ $Var(\mu) = \mu + \frac{\mu^2}{1.433}$			
Degree of freedom	543		
Pearson X^2	508.148	$\chi^2_{0.05,543} = 598.3$	
Scaled Deviance G^2	342.342	$\chi^2_{0.05,543} = 598.3$	
Log-likelihood	-337.044		
AIC	686.088		
Variable	Coefficient	t-value	P-value
Constant c	-3.294	-7.070	0.000
$NAPLN$	0.231	2.980	0.003
$SHLANE$	0.334	3.292	0.000
$LCON$	-0.906	-3.441	0.000
$YBOX$	0.676	3.565	0.000
$TOWN$	0.754	3.984	0.000
Dispersion parameter α	0.698	2.304	0.021

Table 6.25 Short-term (2-year) accident prediction model for
T-unsignalised junctions

$\mu = e^{(-3.723+0.240NAPLN+0.329SHLANE-0.903LCON+0.786YBOX+0.704TOWN)}$ $\mu = 0.024 \times e^{(0.240NAPLN+0.329SHLANE-0.903LCON+0.786YBOX+0.704TOWN)}$ $Var(\mu) = \mu + \frac{\mu^2}{0.867}$			
Degree of freedom	551		
Pearson X^2	572.910	$\chi^2_{0.05,551} = 606.7$	
Scaled Deviance G^2	278.004	$\chi^2_{0.05,551} = 606.7$	
Log-likelihood	-270.322		
AIC	552.643		
Variable	Coefficient	t-value	P-value
Constant c	-3.723	-7.860	0.000
$NAPLN$	0.240	2.773	0.006
$SHLANE$	0.329	2.637	0.008
$LCON$	-0.903	-3.035	0.002
$YBOX$	0.786	3.206	0.001
$TOWN$	0.704	3.061	0.002
Dispersion parameter α	1.154	2.166	0.003

Table 6.26 Short-term (1-year) accident prediction model for T-unsignalised junctions

$\mu = e^{(-4.095+0.237\text{ NAPLN}+0.655\text{ YBOX}+0.909\text{ TOWN})}$ $\mu = 0.017 \times e^{(0.237\text{ NAPLN}+0.655\text{ YBOX}+0.909\text{ TOWN})}$ $Var(\mu) = \mu$			
Degree of freedom	553		
Pearson χ^2	602.792	$\chi^2_{0.05,553} = 608.8$	
Scaled Deviance G^2	251.559	$\chi^2_{0.05,553} = 608.8$	
Log-likelihood	-569.305		
AIC	1146.610		
Variable	Coefficient	t-value	P-value
Constant c	-4.095	-8.724	0.000
NAPLN	0.237	2.710	0.007
YBOX	0.655	2.308	0.021
TOWN	0.909	3.263	0.001

The fit of the variance function of the model to the sequenced averages of model squared residuals for these models are shown in Figures 6.35 to 6.38. Generally the fits seemed to be rather poor compared to the short-term (5-year) model which suggested that the short-term (5-year) model was the best model for the T-unsignalised junctions.

Figure 6.35 Variance function of short-term (4-year) T-unsignalised junctions model

Figure 6.36 Variance function of short-term (3-year) T-unsignalised junctions model

Figure 6.37 Variance function of short-term (2-year) T-unsignalised junctions model

Figure 6.38 Variance function of short-term (1-year) T-unsignalised junctions model

The short-term model with trend

As previously stated, the short-term model with trend was also developed for T-unsignalised junction since the trend variable was found to be significant at 95% confidence level, which indicated that assumption of constant accident rate during 1999-2003 period may not be appropriate. By disaggregating the accident data during the 1999-2003 period and using similar procedure for long-term model with trend, the short-term model with trend for T-unsignalised junction was obtained, as presented in Table 6.27. The model was also developed using NB regression.

Table 6.27 Short-term accident prediction model (with trend)
for T-unsignalised junctions

$\mu = e^{(-3.385 - 0.233 \ln(t) + 0.278 NAPLN - 0.610 RTLANE + 0.216 SHLANE - 0.430 LCON - 0.507 NOMED + 0.485 BOX - 0.504 CITY + 0.442 TOWN)}$ $\mu = 0.034 \times t^{-0.233} \times e^{(0.278 NAPLN - 0.610 RTLANE + 0.216 SHLANE - 0.430 LCON - 0.507 NOMED + 0.485 BOX - 0.504 CITY + 0.442 TOWN)}$ $Var(\mu) = \mu + \frac{\mu^2}{0.975}$			
Degree of freedom	2755	$\chi^2_{0.05, 2755} = 2878$ $\chi^2_{0.05, 2755} = 2878$	
Pearson X^2	2681.744		
Scaled Deviance G^2	1094.151		
Log-likelihood	-947.143		
AIC	1914.287		
Variable	Coefficient	t-value	P-value
Constant c	-3.385	-8.134	0.000
$\ln(t)$	-0.233	-2.097	0.036
$NAPLN$	0.278	4.280	0.000
$RTLANE$	-0.610	-2.414	0.016
$SHLANE$	0.216	2.959	0.003
$LCON$	-0.430	-2.481	0.013
$NOMED$	-0.507	-3.101	0.002
$YBOX$	0.485	3.760	0.000
$CITY$	-0.504	-2.106	0.035
$TOWN$	0.442	3.265	0.001
Dispersion parameter α	1.026	2.774	0.006

It can be seen that short-term model with trend had significant variables almost twice as many of short-term model without trend. This is related to the issue of using disaggregated data versus aggregated data as presented previously in Chapter 5 in that a variable may become significant when data are disaggregated but become

insignificant upon data aggregation such that the standard errors of the model’s parameters (for disaggregated case) are underestimated.

The fit of the variance function of the model to the sequenced averages of model squared residuals is shown in Figure 6.39. The figure shows that the T-unsignalised junction short-term (5-year) model with trend has a good fit since the sequence averages of models residual were clustered around its variance function. The comparison of performance of the model with trend with the other models for T-unsignalised junctions is given in next section.

Figure 6.39 Variance function of short-term (5-year) T-unsignalised junctions model with trend

6.4 Comparison of models’ performance

Comparisons of the developed accident prediction models were made to examine the performance of the models, which were assessed in the form of root-mean square error (*RMSE*) which is equal to the root of sum-squares of error (*SSE*) between the estimated/predicted accident counts and observed accident counts divided by the number of observations, when the models were used to estimate and predict accident counts for the study area. The performances of the models for each type of junctions are presented in Tables 6.28 to 6.31.

Table 6.28 Performance of X-signalised junctions models

Year		RMSE values							
		Full long-term model with trend (1992-2003)	Simple long-term model with trend (1992-2003)	Short-term (5-year) model (1999-2003)	Short-term (4-year) model (2000-2003)	Short-term (3-year) model (2001-2003)	Short-term (2-year) model (2002-2003)	Short-term (1-year) model (2003)	
Estimation period	1999	1.859	1.858	1.822	-	-	-	-	
	2000	1.916	1.877	1.839	1.842	-	-	-	
	2001	1.604	1.637	1.612	1.629	1.621	-	-	
	2002	1.829	1.829	1.825	1.829	1.822	1.824	-	
	2003	1.653	1.683	1.641	1.648	1.657	1.659	1.633	
Prediction period	2004	1.834	1.816	1.800	1.798	1.803	1.803	1.800	
	2005	1.572	1.578	1.625	1.623	1.600	1.601	1.669	
	2006	1.632	1.622	1.649	1.657	1.633	1.635	1.690	

Table 6.29 Performance of T-signalised junctions models

Year		RMSE values						
		Full long-term model with trend (1992-2003)	Simple long-term model with trend (1992-2003)	Short-term (5-year) model (1999-2003)	Short-term (4-year) model (2000-2003)	Short-term (3-year) model (2001-2003)	Short-term (2-year) model (2002-2003)	Short-term (1-year) model (2003)
Estimation period	1999	0.975	1.029	1.037	-	-	-	-
	2000	0.987	1.016	1.000	1.022	-	-	-
	2001	1.058	1.102	1.199	1.232	1.276	-	-
	2002	0.961	1.008	1.000	1.009	1.060	1.142	-
	2003	0.909	0.913	0.929	0.916	0.940	1.004	1.114
Prediction period	2004	0.908	0.889	0.900	0.908	0.896	0.892	0.944
	2005	0.814	0.838	0.893	0.912	0.888	0.875	0.871
	2006	1.092	1.085	1.112	1.124	1.108	1.106	1.125

Table 6.30 Performance of X-unsignalised junctions models

Year		RMSE values							
		Short-term (5-year) model (1999-2003)	Short-term (4-year) model (2000-2003)	Short-term (3-year) model (2001-2003)	Short-term (2-year) model (2002-2003)	Short-term (1-year) model* (2003)			
Estimation period	1999	0.869	-	-	-	-			
	2000	0.455	0.453	-	-	-			
	2001	0.597	0.607	0.657	-	-			
	2002	0.721	0.716	0.708	-	-			
	2003	1.098	1.101	1.110	-	-			
Prediction period	2004	0.845	0.847	0.848	-	-			
	2005	0.479	0.464	0.451	-	-			
	2006	0.692	0.696	0.715	-	-			

* no model could be developed since no variables were found to be significant in the final models

Table 6.31 Performance of T-unsignalised junctions models

Year		RMSE values									
		Short-term (5-year) model with trend (1999-2003)	Short-term (5-year) model (1999-2003)	Short-term (4-year) model (2000-2003)	Short-term (3-year) model (2001-2003)	Short-term (2-year) model (2002-2003)	Short-term (1-year) model (2003)				
Estimation period	1999	0.413	0.422	-	-	-	-				
	2000	0.434	0.435	0.435	-	-	-				
	2001	0.378	0.383	0.382	0.381	-	-				
	2002	0.338	0.339	0.340	0.339	0.339	-				
	2003	0.345	0.351	0.351	0.348	0.348	0.349				
Prediction period	2004	0.404	0.400	0.400	0.400	0.402	0.403				
	2005	0.397	0.392	0.393	0.392	0.394	0.399				
	2006	0.421	0.417	0.416	0.419	0.422	0.420				

In the tables, the observed and estimated accident counts from the models for each year of estimation were compared (in the form of their *RMSE* values). The observed and predicted accident counts from the models for years 2004, 2005 and 2006 were also compared. Smaller *RMSE* indicates a better model relative to the other model(s). The comparisons of the performance are given in sections 6.4.1 to 6.4.4 for X-signalised, T-signalised, X-unsignalised and T-unsignalised junctions models, respectively.

6.4.1 Comparison of X-signalised junctions models' performance

Comparisons of the X-signalised junctions models' performance were carried out in two steps as Table 6.28 shows that no single model distinctly outperformed the other models which makes it less straight forward to compare the performance of the models. The first step was to compare the performance of the full long-term model with trend to the simple long-term model with trend and to compare the performance of the short-term models. This first step would indicate the better long-term model with trend and the better short-term model(s). The performance of these resulting models was then further compared in the second step.

Figure 6.40 gives performance of the short-term models for X-signalised junctions in estimating accidents at the years of estimation while Figure 6.41 gives performance of the short-term models for X-signalised junctions in predicting accidents at the years of prediction.

It can be seen that out of the models for X-signalised junctions, generally the 5-year model was able to estimate accidents better than the 2-year, 3-year, and 4-year models in 2000, 2001, 2003 while the 3-year model estimated accidents better than the 2-year, 4-year, and 5-year models in 2002. On the other hand, the models developed from 2-3 years period were able to predict accidents better than the models developed from 4-5 years period in that the 2-3 years models had the smaller *RMSE* when predicting accidents in 2005 and 2006 whereas the 4-5 years models had the smaller *RMSE* when predicting accidents in 2004.

Figure 6.40 Performance of the short-term X-signalised junctions models at estimation years

Figure 6.41 Performance of the short-term X-signalised junctions models at prediction years

Although the 1-year model estimated accidents better in its estimation period (2003) and also had quite similar performance with the 4-5 years models when predicting accidents in 2004, the 1-year model had the worst performance when predicting accidents in 2005 and 2006, with the *RMSE* values for those periods higher than those for the other models. The 2-year and 3-year models performed quite similarly at every estimation and prediction years with slightly better performance for the 3-

year model. Therefore, the 1-year model was less favoured than the other four short-term models while the 3-year and 5-year models were more favoured than the other short-term models.

The performance of the full long-term model and the simple long-term model for X-signalised junctions is given in Figure 6.42.

Figure 6.42 Performance of the long-term X-signalised junctions models

The figure shows that the full long-term model and the simple long-term model both had quite similar *RMSE* values which means that the performance of both models for estimating and predicting accident counts were quite similar. Thus, the simple long-term model was more favoured than the full long-term model as the former was a simpler model that it had fewer explanatory variables but still can estimate and predict accident counts as well as the more complicated model.

Therefore, the better models for X-signalised junctions were the simple long-term model, short-term (5-year) model and short-term (3-year) model. The performance of these models is given in Figure 6.43. The result shows that it seems the 5-year model was the best model for X-signalised junctions as it gave more years with better estimation/prediction of accidents during 1999-2006 compared to the other models and the simple long-term model with trend was the second best model.

Figure 6.43 Performance of the simple long-term, short-term (5-year) and short-term (3-year) models for X-signalised junctions

6.4.2 Comparison of T-signalised junctions models’ performance

Table 6.29 shows that the full long-term model for T-signalised junctions was the best model for T-signalised junctions as the model gave more years with better estimation/prediction of accidents during 1999-2006 compared to the other models. The short-term 1-year model was the worst-performing model since it had the largest *RMSE* for its estimation/prediction year of 2003, 2004 and 2006 although among the short-term models it best predicted accidents in 2005.

The performance of the short-term models for T-signalised junctions is given in Figures 6.44 and 6.45.

Figure 6.44 Performance of the short-term T-signalised junctions models at estimation years

Figure 6.45 Performance of the short-term T-signalised junctions models at prediction years

It can be seen that among the short-term models for T-signalised junctions, generally the models developed from 2-3 years period were able to predict accidents better than the models developed from 4-5 years period in that the 2-3 years models had the smaller *RMSE* when predicting accidents during 2004-2006. The 2-year model had quite similar performance with the 3-year model for predicting accidents during 2004-2006 but the 2-year model could not estimate accidents during 2002-

2003 as well as the 3-year model. On the other hand, the 5-year model was able to estimate accidents better than the other models in 2000, 2001, 2002 while the 4-year model estimated accidents better than the other models in 2003. In addition, the performance of the 5-year model in predicting accident counts was actually not much different with that of the 3-year model. Therefore, the 5-year model may be considered as the better short-term model for T-signalised junctions.

The performance of the short-term (5-year) model, the full long-term model and the simple long-term model for T-signalised junctions was then compared as shown in Figure 6.46.

Figure 6.46 Performance of the full long-term, simple long-term and short-term (5-year) models for T-signalised junctions

The results indicate that the full long-term model was the better model for T-signalised junctions in that it had generally the lowest *RMSE* among the models (except at years 2004 and 2006, where the simple long-term models performed better). Although the 5-year model was the best among the short-term models, it turned out that its performance was generally not as good as the long-term models.

6.4.3 Comparison of X-unsignalised junctions models’ performance

Models for X-unsignalised junctions consist of only short-term 3-year, 4-year, and 5-year models. The 1-year and 2-year models could not be developed as no significant parameters for the models’ explanatory variables were found in the modelling process. Figure 6.47 shows the performance of the models for X-unsignalised junctions in the form of their *RMSE* values. The figure indicates that the 5-year model was the better model for X-unsignalised junctions since it had more estimation/prediction years with the smallest *RMSE* value among the models.

Figure 6.47 Performance of the short-term models for X-unsignalised junctions

6.4.4 Comparison of T-unsignalised junctions models’ performance

Besides the short-term 1-year, 2-year, 3-year, 4-year, and 5-year models, models for T-unsignalised junctions also consisted of 5-year model that was developed by including a short-term trend. Figure 6.48 shows that all the short-term models without trend performed about equally well during their estimation years. However, as shown in Figure 6.49, the 1-year and 2-year models without trend could not perform as well as the other models without trend during the prediction years although the differences in *RMSE* value were not large. Among these short-term

models without trend, the 5-year model may be considered as the better model since it had generally the lowest *RMSE* values during the prediction years.

Figure 6.48 Performance of the short-term T-unsignalised junctions models without trend at estimation years

Figure 6.49 Performance of the short-term T-unsignalised junctions models without trend at prediction years

The 5-year model with trend generally estimated accidents better than the models without trend during the estimation period, as shown in Figure 6.50 although in 2000 and 2002 the 5-year model with trend and without trend estimated accidents

about equally well. However, the 5-year model with trend could not predict accidents during the prediction years as well as the 5-year model without trend. Therefore, the 5-year model without trend may be considered as the better model for T-unsignalised junctions.

Figure 6.50 Performance of the short-term (5-year) T-unsignalised junctions models with trend and without trend

6.5 Sensitivity analysis

A simple sensitivity analysis was carried out to examine the impact of inaccuracy of the estimated traffic flow data on the developed signalised junctions models. This kind of analysis is needed because traffic flow data used to develop the accident prediction models were estimated using several assumptions, as described earlier in Chapter 3. The sensitivity analysis was carried out by varying the traffic flow data by between -30% to +30% of the initial value of traffic. The result of sensitivity analysis for the short-term (5-year) X-signalised junctions model, the short-term (5-year) T-signalised junctions model, the simple long-term X-signalised junctions model and the simple long-term T-signalised junctions model are presented in Tables 6.32 to 6.35, respectively.

Table 6.32 Sensitivity analysis of short-term X-signalised junctions model

Changes in	Changes in traffic flow (of modelling value)						
	-30%	-20%	-10%	0	+10%	+20%	+30%
μ (of initial value)	0.734	0.824	0.913	1	1.086	1.171	1.255

Table 6.33 Sensitivity analysis of short-term T-signalised junctions model

Changes in	Changes in traffic flow (of modelling value)						
	-30%	-20%	-10%	0	+10%	+20%	+30%
μ (of initial value)	0.689	0.792	0.896	1	1.105	1.210	1.316

Table 6.34 Sensitivity analysis of long-term X-signalised junctions model

Changes in	Changes in traffic flow (of modelling value)						
	-30%	-20%	-10%	0	+10%	+20%	+30%
μ (of initial value)	0.711	0.808	0.904	1	1.096	1.191	1.286

Table 6.35 Sensitivity analysis of long-term T-signalised junctions model

Changes in	Changes in traffic flow (of modelling value)						
	-30%	-20%	-10%	0	+10%	+20%	+30%
μ (of initial value)	0.759	0.841	0.922	1	1.077	1.152	1.225

It can be observed from Table 6.32 that varying traffic flow for X-signalised junctions by -30% to +30% of modelling value based on short-term (5-year) model would result in expected accident number to change by between 0.734 to 1.255 of its initial value whereas varying traffic flow for T-signalised junctions by -30% to +30% of modelling value based on short-term (5-year) model would result in expected accident number to change by between 0.689 to 1.316 of its initial value. If the analyses were based on long-term models as shown in Tables 6.34 and 6.35 for X-signalised junctions and T-signalised junctions, respectively, then varying traffic flow by -30% to +30% of modelling value would result in expected accident number to change by between 0.711 to 1.286 of its initial value (for X-signalised junctions model) and between 0.759 up to 1.225 of its initial value (for T-signalised junctions model).

The above analysis indicates that the short-term (5-year) model for X-signalised junctions was relatively less sensitive to changes in traffic flow than the simple

long-term model for X-signalised junctions. On the other hand, the short-term (5-year) model for T signalised junctions was relatively more sensitive to changes in traffic flow than the simple long-term model for T-signalised junctions. In general, a change in traffic flow by a certain percentage would result in about the same percentage of change in expected accident number μ .

It should be noted that the expected accident numbers obtained from the accident prediction model served to give an estimate of safety of the reference population. The empirical Bayes (EB) and modified empirical Bayes (MEB) methods for safety estimation (as presented in Chapter 7) make use of the safety of the reference population that is combined with historical accident counts of a site in order to obtain the safety of a site. It was found that the change of traffic flow value had some but lesser effect on the outputs of EB and MEB methods. For example, varying the traffic flow in the short-term model for X-signalised junctions by +30% of the initial value of traffic would result in between 3% to 19% change in the expected accident number μ among the sites, with an average of 7%. If the simple long-term model for X-signalised junctions were used, the changes in the expected accident number μ would be between 1% to 18%. For the short-term T-signalised junctions model, the changes were between 4% to 26%, and for the simple long-term T-signalised junctions model the changes were between 1% up to 17%.

Therefore, although the developed accident prediction models were sensitive to changes in traffic flow, their effects were moderated when their outputs were incorporated into the EB and MEB methods.

6.6 Factors associated with accident occurrences

There are two main applications of accident prediction model. Firstly, it is used to predict the accident frequency and safety of an entity, and secondly it can give associative relationships with accident occurrences. In any case it must be cautioned that one should not interpret accident prediction model as giving causative relationships.

The following subsections examine the factors that are associated with accident occurrences at signalised and unsignalised road junctions in Singapore.

6.6.1 Factors associated with accident occurrences at signalised junctions

The modelling results show that accident occurrences at the X-signalised and T-signalised junctions are related to the product of total traffic inflows at primary direction and secondary direction (cross-flow product). The finding is consistent with the findings of other researchers such as Leong (1973) and Pickering *et al.* (1986). Two other forms of traffic interactions were examined in this research with significant relationships but the cross-flow products performed slightly better than the other forms.

The short-term models for X-signalised and T-signalised junctions failed to produce geometric/operational characteristics that were significantly associated with accident occurrences at the X-signalised and T-signalised junctions except for two features which were location of the junction (in terms of junction environment or landuse type) and junction classification. However, several geometric/operational characteristics were found to be significantly associated with accident occurrences in the full long-term models for X-signalised and T-signalised junctions. For X-signalised junctions, the significant geometric/operational characteristics were junction classification, number of signal phases, number of shared lanes, presence of median, presence of red-light camera, presence of yellow box and junction location. For T-signalised junctions, the significant geometric/operational characteristics were junction classification, number of shared lanes, presence of U-turn, presence of red-light camera, presence of yellow box and junction location.

Junction classification can influence accident occurrences at X-signalised and T-signalised junctions in the manner that a major junction has generally more accidents than a minor junction due to more complex features that a major junction has than a minor junction. This research defines a major junction as a junction

where primary roads intersect with each other or where primary road(s) intersect with secondary road(s), otherwise it is defined as a minor junction.

Having more number of signal phase at X-signalised junctions can result in fewer accidents because the number of conflict points at the junctions is reduced. The more number of shared lanes at X-signalised and T-signalised junctions is associated with a decrease in safety in general as more conflicting movements may be created during the start of the green signal phase.

The presence of a central median at X-signalised junctions is associated with generally higher accident number. However, one should not interpret the presence of medians at junction approaches as being hazardous; it could be that the medians are generally installed at large junctions with heavy traffic flows to better segregate/channelise certain traffic movements in order to lessen the risk of accidents. A similar explanation can be given to the installation of red-light cameras at X-signalised and T-signalised junctions which are also associated with higher accident occurrences. The finding on red-light camera is similar to the finding of a study on safety at Singapore's signalised T-junctions by Kumara and Chin (2003).

Another factor associated with generally higher accident occurrences at X-signalised and T-signalised junctions is the presence of the yellow box in the intersection area. The yellow box might have been installed at junctions to improve safety. The presence of a yellow box can lead to vehicles reducing their speed when approaching junctions in order not to encroach onto the yellow box, especially when close to the signal change, which can thus reduce red-light running. Without the yellow box, the accident rate could well be higher.

The presence of U-turn at T-signalised junctions is also associated with higher accident occurrences. The presence of U-turn creates conflicting movements between vehicles that make a U-turn and straight-through vehicles.

Junction environment (landuse) variable was originally defined by two indicator variables to distinguish city, town and other locations. Later it was found that there was no significant difference between X-signalised and T-signalised junctions located in town area and other non-city area. The junction environment was thus reduced to a city versus non-city binary variable. The junction environment influences accident occurrences at signalised junctions in the manner that junctions located in the city areas have generally fewer accidents compared to junctions with similar characteristics located in the non-city area. This may be due to the relatively lower operating speed in the city area while noting that the distances between junctions are relatively short and there are greater traffic activities within the city area. Moreover, the traffic composition in the city area is also different to that of non-city areas. According to a study by Naing (2004), there was a much higher proportion of motorcycles outside of the city area (14% of the total traffic) compared to in the city area (7% of the total traffic); the higher motorcycle proportion would also contribute to higher accident occurrences in non-city areas.

6.6.2 Factors associated with accident occurrences at unsignalised junctions

The model for X-unsignalised junctions showed that accident occurrences are associated with the number of approach lanes and number of exclusive right-turn lanes while the model for T-unsignalised junctions showed that accident occurrences are associated with the number of approach lanes, number of shared lanes, presence of median and yellow box as well as landuse type.

The number of approach lanes influences accident occurrences at unsignalised junctions in that the greater the total number of approach lanes, the higher the likelihood of accident occurrences. This is reasonable as more approach lane means greater exposure to traffic flow.

The more number of exclusive right-turn lanes in X-unsignalised junction is associated with an increase in safety. It is clear that the presence of exclusive right-

turn lane reduces the likelihood of head-to-rear accidents that is caused by diverging conflict in shared right-through lanes or in shared left-through-right lanes.

The more number of shared lanes (i.e. shared left-through lanes, shared left-right lanes, shared right-through lanes) is associated with increases in accident occurrences. The presence of a shared right-through lane in the top of a T-unsignalised junction increases the likelihood of head-to-rear accidents that is caused by a stopped right-turning vehicle that is waiting for a gap as through vehicles approach in the same lanes at prevailing speed. Similar situation can occur for shared left-through lanes in the middle of a T and shared left-through lanes in the minor of a T.

Other geometric factors found to be significant at T-unsignalised junctions are the lane configuration and the presence of yellow box. The lane configuration influences the accident occurrences in that accidents at a junction with similar lane configuration on all its legs are generally lower than those occur at a junction with different lane configuration. This is related to more difficult situations encountered by drivers at junctions with different lane configuration. A yellow box is often installed at T-unsignalised junctions with short links and high traffic flows that potentially had higher accident occurrences to force drivers to be more cautious when entering the junctions. Without the yellow box, the accident could well be higher.

Junction environment or landuse influences accident occurrences in T-unsignalised junctions in the manner that junctions located in town area generally have more accidents compared to junctions with similar characteristics located in city area and other (outside city and town) area. In this regard, town area is defined as a cluster of commercial-residential area that is located outside the city but with heavy traffic/pedestrian activities. In Singapore, for example, such area is located around the Mass Rapid Transit (MRT) station. The traffic flows in the town area are not as high as in the city area thus lesser traffic density and higher speed. Vehicle-pedestrian accidents may increase when there is greater approach width and higher

speed. On the other hand, T-unsignalised junctions that are located outside the city and town areas (e.g. residential suburbs) have generally fewer accidents as compared to junctions with similar characteristics in town area. This may be due to less pedestrian activities as compared to town area.

6.6.3 Factors associated with severity of accidents

Certain combination of geometric/operational characteristics of the junctions may influence the severity level of accidents. Tables 6.36 and 6.37 show the average severity levels of accidents occurring at signalised junctions and unsignalised junctions, respectively.

It can be seen in Table 6.36 that X-signalised junctions with sliproads, yellow box and median on all its legs, a red-light camera and located outside city area had the largest rate of fatal accidents, which was 0.89 fatal accidents per site in 5 years (1999-2003). This group of junctions also had the highest rates of serious injury accidents and slight injury accidents during the same period, which was 1.22 serious injury accidents per site and 20.56 slight injury accidents per site.

For unsignalised junctions as seen in Table 6.37, a T-unsignalised junction with no right-turn lane, but has a yellow box and median on at least one leg, and is located outside the city and town area was found to have the highest rate of fatal accidents during 1999-2003 (0.1 fatal accidents per site). Similar T-unsignalised junction but located in the city area had the highest rate of serious injury accidents during the same period, which were 0.17 serious injury accidents per site. X-unsignalised junction with no right-turn lane, no yellow box but had median on at least one leg had the highest rate of slight injury accidents during 1999-2003 period, which was (3.43 slight injury accidents per site).

Table 6.36 Accident severities at signalised junctions

JTYPE	SLROAD	MED	RLCAM	YBOX	CITY	N	Average of fatal accidents	Average of serious injury accidents	Average of slight injury accidents
0	0	0	0	0	0	56	0	0.09	0.91
					1	9	0	0.11	1.33
				1	0	21	0.05	0.1	2.1
			1		1	12	0	0.08	0.92
				0	0	4	0	0.25	4
		1	0	1	0	3	0.33	0.33	10
				0	0	21	0.1	0.14	2.9
				1	0	11	0.36	0.09	3.91
					1	1	0	0	1
			1	0	0	2	0	0	3.5
				1	0	3	0	0.33	5.67
	1	0	0	0	0	8	0.13	0.13	4
				1	0	5	0	0.6	4.8
			1	0	0	1	0	0	2
				0	0	27	0.11	0.22	3.74
		1	0	0	1	0	0.1	0.2	4.9
				1	0	4	0	0	9.5
					1	0	0	0.33	13.33
1	0	0	0	0	0	28	0	0.11	2.96
					1	2	0	0	0
				1	0	17	0.12	0.06	4.88
					1	54	0.11	0.22	6.02
			1	0	0	3	0	0	2.67
				1	0	1	0	0	7
					1	5	0.6	0	9.8
		1	0	0	0	13	0.23	0.23	5.77
					1	1	0	0	15
				1	0	8	0.25	0.38	8.38
			1		1	3	0.33	0	8
				0	0	1	1	1	9
				1	0	4	0.25	0.25	8.5
	1	0	0	0	0	4	0	0.25	4.25
				1	0	3	0.33	0	6
			1	0	0	1	0	1	2
		1	0	0	0	26	0.15	0.23	9.54
				1	0	15	0.33	0.6	13
					1	1	0	0	9
			1	0	0	4	0	1	17.75
				1	0	9	0.89	1.22	20.56

Table 6.37 Accident severities at unsignalised junctions

JTYPE	RTLANE	NOMED	YBOX	CITY	TOWN	N	Average of fatal accidents	Average of serious injury accidents	Average of slight injury accidents
0	0	0	0	0	0	22	0	0.09	1.14
					1	22	0	0.05	0.77
				1	0	7	0	0	0.14
			1	0	0	10	0.1	0.1	1.9
					1	21	0.05	0.05	2.24
				1	0	12	0	0.17	0.42
		1	0	0	0	177	0.03	0.01	0.15
					1	94	0	0.01	0.57
				1	0	46	0	0.02	0.28
			1	0	0	58	0.02	0.02	0.67
					1	42	0	0.02	0.69
				1	0	17	0	0.06	0.53
	1	0	0	0	0	2	0	0	2.5
					1	5	0	0.2	2.6
			1	0	0	7	0	0.14	1
					1	5	0	0	1.6
		1	0	0	0	1	0	0	0
					1	0	0	0	0.67
			1	1	0	1	0	0	0
					1	0	0	0	0
	2	0	0	0	0	1	0	0	0
					1	0	0	0	0
		1	0	1	0	1	0	0	0
1	0	0	0	0	0	7	0	0	3.43
					1	1	0	0	0
				1	0	2	0	0	0.5
		1	0	0	0	42	0.02	0.1	1.4
					1	8	0	0.13	1.63
				1	0	3	0	0	0
	1	0	0	0	0	1	0	0	0
					1	1	0	0	0
				1	0	1	0	0	3
		1	0	1	0	2	0	0	0
	2	0	0	0	0	1	0	0	2
					1	1	0	0	1
				1	0	1	0	0	0

6.7 Summary

Several accident prediction models were developed for X-signalised, T-signalised, X-unsignalised and T-unsignalised junctions according to procedure established in Chapter 5. The models for X-signalised and T-signalised junctions comprised of long-term models and short-term models while no long-term models were developed for X-unsignalised and T-unsignalised junctions due to unavailability of

data. The long-term models, which consisted of simple long-term model and full long-term model, were formulated by including a long-term trend in accident occurrences during 1992-2003 period. The short-term models were formulated by assuming constant accident rate during the modelled periods, which were varied between 1 year and 5 years' period in which all the modelled periods started from year 2003 backwards. The performance of the models was compared and the better model(s) was/were selected based on the models' performance when used to estimate accident frequencies during their estimation years and to predict accident frequencies at prediction years of 2004, 2005 and 2006.

Both short-term models and long-term models have their own advantages and disadvantages. The short-term models have an advantage in terms of less effort to assemble calibration data. However, there is a loss in degrees of freedom as data-points are combined and also less certainty that the modelled data are fully tested as to whether they follow Poisson distribution. On the contrary, the long-term models require a more complicated database as each observation in a given year represents one data-point. However, the long-term models can be used to project accidents for longer future period than the short-term models. Since the models developed in this research are to be used for predicting short-term (2004, 2005 and 2006) accidents only, the use of short-term models are more favoured than the long-term models whenever both types of models performed similarly well.

The modelling results have shown the superiority of the NB regression model over Poisson, ZIP and ZINB models in the development of accident prediction model. Almost all the developed models were found to be best-fitted by an NB model. The exceptions were for the short-term (4-year) X-unsignalised model, short-term (3-year) X-unsignalised model and short-term (1-year) T-unsignalised model, which were best-fitted by a Poisson model. In most cases, the ZINB model failed to perform well even when the data had excess zeroes. However, in some cases, the ZIP model may be a suitable alternative compared to the NB model as the Vuong statistics of the ZIP models were found to be significant. The NB model is also more favourable than the ZIP model in that an NB model can give a simpler and

more direct interpretation. Therefore, in cases where the goodness-of-fit criteria of NB and ZIP models are comparable, the NB model should be selected.

Overall, the short-term (5-year) models were found to have better performance than the other models, with exception for T-signalised junction models in which the long-term models performed better. Among the short-term models for T-signalised junctions, the short-term (5-year) model was the better model. Overall, the use of the short-term (5-year) models was recommended. The recommended models were applied for estimating/predicting safety at road junctions by empirical Bayes (EB) and modified empirical Bayes (MEB) method in the next stage of the research. Since the long-term models for T-signalised junctions performed better than the other models and the simple long-term model for X-signalised junctions was also a relatively well-fitting model for X-signalised junctions, the simple long-term models for X-signalised and T-signalised junctions were also included in the EB and MEB analysis. The recommended models are shown in Table 6.38.

The short-term models for X-signalised and T-signalised junctions failed to produce geometric/operational characteristics that were significantly associated with accident occurrences at the X-signalised and T-signalised junctions except for two features which were location of the junction (in terms of junction environment or landuse type) and junction classification. Several geometric/operational characteristics were found to be significantly associated with accident occurrences in the full long-term models for X-signalised and T-signalised junctions. For X-signalised junctions, the significant geometric/operational characteristics were junction classification, number of signal phases, number of shared lanes, presence of median, presence of red-light camera, presence of yellow box and junction location. For T-signalised junctions, the significant geometric/operational characteristics were junction classification, number of shared lanes, presence of U-turn, presence of red-light camera, presence of yellow box and junction location.

The model for X-unsignalised junctions showed that accident occurrences are associated with the number of approach lanes and number of exclusive right-turn

lanes while the model for T-unsignalised junctions showed that accident occurrences are associated with the number of approach lanes, number of shared lanes, presence of median and yellow box as well as landuse type.

Table 6.38 Summary of accident prediction models

<p>Simple long-term X-signalised junctions model:</p> $\mu_t = 1.371 \times 10^{-4} (Q_{p_t} Q_{s_t})^{0.479} (t)^{-0.117} e^{(0.324CLASS - 0.555CITY)} \text{ (in accidents/year)}$ $Var(\mu)_t = \mu_t + \frac{\mu_t^2}{3.050}$
<p>Simple long-term T-signalised junctions model:</p> $\mu_t = 6.587 \times 10^{-4} (Q_{p_t} Q_{s_t})^{0.387} (t)^{-0.190} e^{(0.597CLASS - 2.656CITY)} \text{ (in accidents/year)}$ $Var(\mu)_t = \mu_t + \frac{\mu_t^2}{2.605}$
<p>Short-term (5-year) X-signalised junctions model:</p> $\mu = 1.137 \times 10^{-3} (Q_p Q_s)^{0.434} e^{(0.624CLASS - 0.649CITY)} \text{ (in accidents/5 years)}$ $Var(\mu) = \mu + \frac{\mu^2}{2.404}$
<p>Short-term (5-year) T-signalised junctions model:</p> $\mu = 1.706 \times 10^{-4} (Q_p Q_s)^{0.523} e^{(0.601CLASS - 1.496CITY)} \text{ (in accidents/5 years)}$ $Var(\mu) = \mu + \frac{\mu^2}{1.624}$
<p>Short-term (5-year) X-unsignalised junctions model:</p> $\mu = 0.060 \times e^{(0.572NAPLN - 1.272RTLANE)} \text{ (in accidents/5 years)}$ $Var(\mu) = \mu + \frac{\mu^2}{1.100}$
<p>Short-term (5-year) T-unsignalised junctions model:</p> $\mu = 0.117 \times e^{(0.229NAPLN + 0.174SHLANE - 0.573NOMED + 0.738YBOX + 0.493TOWN)} \text{ (in accidents/5 years)}$ $Var(\mu) = \mu + \frac{\mu^2}{1.042}$

CHAPTER 7

ESTIMATION OF SAFETY BY EMPIRICAL BAYES METHOD

7.1 Introduction

This chapter presents estimation of safety at the study sites by empirical Bayes (EB) method. In the first part, the theory of EB method is presented followed by modification of the EB method and discussion on the properties of both methods. Procedures for estimation and prediction of safety are given afterwards with examples of calculations. The analysis of result is then presented followed by a conclusion.

7.2 Empirical Bayes (EB) method

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

To derive the EB model, Hauer (1997) started with introducing two variables namely m (the expected number of accidents of an entity occurring within a group of similar entities during a specified time period) and K (historical accident counts of the entity in a specified period). Let the mean and variance of m in the reference population (group of entities) be $E(m)$ and $Var(m)$. Here, $E(m)$ is taken as the safety of the population. Where information on the historical accident counts of an

entity is not available, $E(m)$ is also the best estimate of safety m of the entity. However, if such information is available then the best estimate of safety m of the entity that has K accidents is given by $E(m|K)$. In this regard, $E(m|K)$ and $Var(m|K)$ respectively denote the mean and the variance of m in the subpopulation that recorded K accidents.

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

$$g(m) = \frac{a^b m^{b-1} e^{-am}}{\Gamma(b)} \quad (7.1)$$

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

$$E(m) = \frac{b}{a}, \quad Var(m) = \frac{b}{a^2} \quad (7.2)$$

then,

$$a = \frac{E(m)}{Var(m)}, \quad b = \frac{E(m)^2}{Var(m)} \quad (7.3)$$

$E(m)$ is estimated by using the result of accident prediction model where the expected accident count K is given by μ , which is similar to $E(m)$, with variance equal to $Var(\mu)$. Hauer (1992) showed that $Var(\mu) = E(m) + Var(m)$. For the case of accident prediction model that is built using NB regression model then $Var(m) = \alpha E(m)^2$ because the variance of the NB regression model is given by $Var(\mu) = \mu + \alpha \mu^2$ and also $E(K) = E(m)$. Here, α is the dispersion parameter of the NB model and $b = 1/\alpha$.

According to Bayes' theorem, the probability distribution for $f(m|K)$ is given by the following relationship:

$$f(m|K) = (\text{constant})P(K|m)g(m) \quad (7.4)$$

Hence, because the accident count K is assumed to be Poisson-distributed,

$$P(K|m) = \frac{m^K e^{-m}}{K!} \quad (7.5)$$

Using the expression for $P(K|m)$ and $g(m)$, then

$$f(m|K) = (\text{constant}_1) \frac{m^K e^{-m}}{K!} \frac{a^b m^{b-1} e^{-am}}{\Gamma(b)} = (\text{constant}_2) m^{K+b-1} e^{-m(1+a)} \quad (7.6)$$

After integration over all $m \geq 0$, the normalizing constant_2 turns out to be

$$\frac{(1+a)^{K+b}}{\Gamma(K+b)} \quad (7.7)$$

Thus, we obtain

$$f(m|K) = \frac{(1+a)^{K+b} m^{K+b-1} e^{-m(1+a)}}{\Gamma(K+b)} \quad (7.8)$$

with mean and variance

$$E(m|K) = \frac{K+b}{1+a} \quad (7.9)$$

$$\text{Var}(m|K) = \frac{K+b}{(1+a)^2} \quad (7.10)$$

7.3 Modified empirical Bayes (MEB) method

The estimation of safety by the EB method described earlier rests on two basic assumptions, which are:

- a. the probability distribution of m in the entities of the reference population is described by a Gamma probability density function $g(m)$,
and
- b. the accident count K is Poisson-distributed.

Previous analysis has shown that the GP distribution may be more appropriate for describing the probability distribution of accident count K of an entity. In this regard, the original EB method was modified by considering the accident count K as being GP-distributed instead of Poisson-distributed in order to allow using actual dispersion of accident count data in the estimation. Hence, the index of dispersion of accident count data is not assumed to be unity as required by the Poisson distribution. The probability distribution of m in the entities of the reference population shall continue to be described by a Gamma probability density function - $g(m)$.

The use of the GP distribution in the estimation of safety does not imply that K is not Poisson-distributed but is aimed at improving the precision of safety estimation. Moreover, using a Poisson assumption when the actual variation of the historical accident count data is greater or less than required by the Poisson distribution (index of dispersion is greater or less than unity) may have implication on making the incorrect decision when determining the hazard level of a location which in turn affects the determination of the amount of countermeasure actions needed for the site. Nicholson (1985) noted that the use of Poisson assumption when historical accident counts are fairly regular (index of dispersion less than one) increases the probability of taking no action when it is warranted as the change in underlying true accident rate will not be detected sufficiently. On the other hand, the use of Poisson assumption when historical accident counts are fairly irregular (index of dispersion greater than one) causes the probability of taking some actions more than it is warranted because of the greater likelihood in assuming an accident count fluctuation as being a change in the underlying true accident rate.

The derivation of the modified Empirical Bayes (MEB) method to estimate safety is presented in the following.

As similar to the EB method,

$$f(m|K) = (\text{constant})P(K|m)g(m) \quad (7.4)$$

The probability distribution of m in the entities of the reference population is similarly described by a Gamma probability density function given by:

$$g(m) = \frac{a^b m^{b-1} e^{-am}}{\Gamma(b)} \quad (7.1)$$

where

$$a = \frac{E(m)}{VAR(m)}, \quad b = \frac{E(m)^2}{VAR(m)} \quad (7.3)$$

Now the historical accident count K is GP-distributed, so

$$P(K|m) = \frac{m(1-\lambda)(m(1-\lambda) + \lambda K)^{K-1} e^{-(m(1-\lambda) + \lambda K)}}{K!} \quad (7.11)$$

Substituting Equations (7.1) and (7.11) into Equation (7.4):

$$f(m|K) = (constant_1) \frac{m(1-\lambda)(m(1-\lambda) + \lambda K)^{K-1} e^{-(m(1-\lambda) + \lambda K)}}{K!} \frac{a^b m^{b-1} e^{-am}}{\Gamma(b)} \quad (7.12)$$

$$f(m|K) = (constant_2) m(m(1-\lambda) + \lambda K)^{K-1} e^{-m(1-\lambda)} m^{b-1} e^{-am} \quad (7.13)$$

$$f(m|K) = (constant_2) e^{-m(1-\lambda+a)} m^b (m(1-\lambda) + \lambda K)^{K-1} \quad (7.14)$$

Integrate Equation (7.14) over all $m \geq 0$:

$$\int_0^\infty f(m|K) dm = (constant_2) \int_0^\infty m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm \quad (7.15)$$

$$constant_2 = \frac{\int_0^\infty f(m|K) dm}{\int_0^\infty m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} \quad (7.16)$$

$$constant_2 = \frac{1}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} \quad (7.17)$$

Because the analytical solution of the integration form in Equation (7.17) is not easily obtained, it shall be solved numerically in order to obtain the value of $constant_2$. The MATHCAD software package version 12 was used for the numerical estimation task in this research.

The value of $constant_2$ in Equation (7.17) is substituted into Equation (7.15), such that:

$$f(m|K) = \frac{m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1}}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} \quad (7.18)$$

The mean and variance of the $f(m|K)$ can be obtained analytically by applying:

$$E(m|K) = \frac{\int_0^{\infty} m f(m|K) dm}{\int_0^{\infty} f(m|K) dm} = \int_0^{\infty} m f(m|K) dm \quad (7.19)$$

and

$$Var(m|K) = \int_0^{\infty} (m - E(m|K))^2 f(m|K) dm \quad (7.20)$$

Therefore, the MEB safety estimate of an entity is:

$$E(m|K) = \frac{\int_0^{\infty} m^{b+1} e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} \quad (7.21)$$

with variance:

$$Var(m|K) = \frac{\int_0^{\infty} (m - E(m|K))^2 m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} \quad (7.22)$$

7.4 Properties of EB and MEB estimates

It is clear that no accident prediction model can ever represent all the variables that influence safety of an entity, and a model cannot be expected to account for every site-specific feature of all the entities. This has motivated the development of the EB/MEB method for safety estimation. In the EB/MEB method, the accident prediction model only provides a portion of the information about an entity's safety in terms of how safety varies in a reference population of similar entities. In the absence of any other knowledge about the entity, then the safety of the reference population is taken as the estimated safety of the entity. However, when historical accident data of the entity are available, the EB/MEB method then refines the safety estimation. Under the EB/MEB method, safety varies in a subpopulation of entities having not only similar traffic and geometric characteristics but also similar historical accident data. Thus the EB/MEB safety estimate of an entity can be viewed as the result of correcting the accident prediction model to yield a site-specific safety estimate.

In the EB estimation, the accident count sequence of an entity is assumed to be Poisson distributed, which means that λ is restricted to be equal to zero whereas in the MEB estimation λ can be smaller or larger than zero depending on the variation observed in the historical accident data over the long run. This is the main advantage of the MEB method over the EB method as no restriction is implied in the variation of the data. Moreover, using the observed variation is important so as to avoid problems in identifying/ranking hazardous location as noted earlier in Section 7.2.

7.4.1 Properties of EB estimate

The EB estimate as in Equation (7.9) can be rewritten as:

$$E(m|K) = \eta E(m) + (1 - \eta)K \quad (7.23)$$

where

$$\eta = \frac{E(m)}{E(m) + Var(m)} = \frac{1}{1 + \frac{Var(m)}{E(m)}} \quad (7.24)$$

Equation (7.23) shows that $E(m|K)$ is a weighted combination of $E(m)$ and K with the value of $E(m|K)$ that is always between $E(m)$ and K . Hence, $E(m|K)$ is the result of correcting $E(m)$ in the direction of K . When $E(m)$ is smaller than K then $E(m)$ will be corrected upward to the direction of K . When $E(m)$ is greater than K then $E(m)$ will be corrected downward to the direction of K . Furthermore, $Var(m|K)$ will always be smaller than $E(m|K)$ as shown by Equation (7.10) which indicates that an EB estimate is a refinement of the accident prediction model whereas $Var(m)$ of the accident prediction model may be larger than $E(m)$.

Equations (7.23) and (7.24) show that when $Var(m)$ is very small (approaching zero) then the EB estimate of an entity's safety $E(m|K)$ is close to the safety of reference population given by $E(m)$ as η is close to unity. Since all accident prediction models in this research were developed using the NB regression model, it should be noted that the magnitude of $Var(m)$ is very much dependent on the magnitude of the dispersion parameter α . Therefore, the larger is the value of α , the smaller will be the value of $Var(m)$, which means that $E(m)$ is a very good approximation of the entity's safety.

On the other hand, when the value of α is very small (near zero) then $Var(m)$ is very large (near infinity) which makes η close to zero. In this case, $E(m)$ would be

a poor indicator of the safety given by $E(m|K)$, and the value of safety would be very close to the accident count K .

7.4.2 Properties of MEB estimate

Unlike the EB estimate, the properties of the MEB estimate are not easy to be explained mathematically. Thus, graphical analysis was carried out to study such properties. Graphs were generated using outputs of the accident prediction model in the form of $E(m)$ and $Var(m)$ pairs (see sample procedure and calculation in Section 7.5 later). MEB estimates were calculated using these $E(m)$ and $Var(m)$ pairs combined with λ values between -0.99 and $+0.99$ that were varied according to selected values of accident count K .

Six example pairs of $E(m)$ and $Var(m)$ were selected that represented very low, low, lower-medium, upper-medium, high and very high values of $E(m)$ which form the MEB estimate graphs as presented respectively in Figures 7.1 – 7.6. The calculations also revealed that for very low to lower-medium value of $E(m)$, $Var(m)$ is always smaller than $E(m)$. For upper-medium value of $E(m)$, $Var(m)$ is generally very close to the value of $E(m)$. For high and very high values of $E(m)$, $Var(m)$ is always greater than $E(m)$.

The figures indicate that generally the estimation of $E(m|K)$ by the MEB method does not pose any major problem for positive values of λ . For λ in the positive range, as λ increases, the $E(m|K)$ value for $K > 0$ converges to a value slightly greater than $E(m)$ while the $E(m|K)$ value for $K = 0$ gradually tends to the value of $E(m)$. This means that MEB method can be used without any limitation for any positive value of λ .

Figure 7.1a MEB estimates, $E(m)=0.233$ and $Var(m)=0.012$

Figure 7.1b MEB estimates, $E(m)=0.233$ and $Var(m)=0.012$

Figure 7.2a MEB estimates, $E(m)=2.047$ and $Var(m)=0.905$

Figure 7.2b MEB estimates, $E(m)=2.047$ and $Var(m)=0.905$

Figure 7.3a MEB estimates, $E(m)=4.949$ and $Var(m)=2.370$

Figure 7.3b MEB estimates, $E(m)=4.949$ and $Var(m)=2.370$

Figure 7.4a MEB estimates, $E(m) = 9.097$ and $Var(m) = 8.006$

Figure 7.4b MEB estimates, $E(m) = 9.097$ and $Var(m) = 8.006$

Figure 7.5 MEB estimates, $E(m) = 14.191$ and $Var(m) = 19.484$

Figure 7.6 MEB estimates, $E(m) = 24.071$ and $Var(m) = 56.058$

When λ takes on a negative value, there may be a problem of an overly high or low value of $E(m|K)$, especially when $K > E(m)$. This kind of problem is indicated by the existence of a fluctuating zone in the estimation, where the value of $E(m|K)$ changes considerably over a small change in λ value. It is usually followed by a $Var(m|K)$ that is greater than $E(m|K)$ or even a negative $Var(m|K)$. The $E(m|K)$ value may also become negative, for example, the $E(m|K)$ value for $K = 10$, $\lambda = -0.85$ and $\lambda = -0.9$ in Figure 7.4. Such a problem may be serious and should thus be avoided wherever possible. Should this problem occur (which was found to be quite rare for sites in this study – see Table 7.3 later), it is recommended to revert to the Poisson assumption for the affected data.

There may also be a possibility that the probability of $E(m|K) > E(m)$ exceeds unity. When this condition occurs in conjunction with an overly high or low $E(m|K)$ value or a negative variance, one should revert to the Poisson assumption. However, if the condition occurs by itself and the resulting probability is not much greater than one, a value of one is assumed.

Another important property of the MEB method that makes it different from the EB method is that for high values of $E(m)$, which generally means that

$Var(m) > E(m)$, the estimate of $Var(m|K)$ may be greater than the estimate of $E(m|K)$. However, when such condition occurs, the estimate of $Var(m|K)$ is usually still smaller than $Var(m)$ so it is not a serious matter. The problem may be serious when λ is negative and the value of $E(m|K)$ lies in the fluctuating zone.

It is also observed from the graphs of MEB estimates that the value of $E(m|K)$ does not necessarily lie between K and $E(m)$, especially for overly large positive or negative λ value as unlike the case of EB estimates. As long as the $E(m|K)$ estimates do not differ greatly from the $E(m)$ estimates, it is not a serious problem. However, when the estimates differ greatly, one should examine more on the problem through the plot of λ with $E(m|K)$. A serious problem may exist when λ is negative and the value of $E(m|K)$ lies in the fluctuating zone. Thus, MEB estimates should be used with great care while always being cognizant of the various pitfalls.

7.5 Procedures for safety estimation and prediction

The EB and MEB safety estimation methods shall be applied to estimate the safety of an entity for a given observation period as well as to predict the safety of an entity in the future period. The MEB safety estimation method shall later be applied to determine and rank hazardous locations.

7.5.1 A calculation procedure for estimation of safety

The procedure to estimate safety for a given observation period, using both the EB and MEB methods, is as follows:

1. Develop the accident prediction model, using multivariate (NB) regression (see Chapters 5 and 6).

2. If the accident prediction model is developed through aggregating accident data during the modelled period (no trend is included in the model), then:

$$E(m) = \mu \quad (7.25)$$

$$Var(m) = \alpha E(m)^2 \quad (7.26)$$

If the accident prediction model is developed through disaggregating accident data during the modelled period (a trend may or may not be included in the model), then:

- a. Calculate $E(m)_t$ and $Var(m)_t$ for the observation period, using the fitted accident prediction model:

$$E(m)_t = \mu_t \quad (7.27)$$

$$Var(m)_t = \alpha E(m)_t^2 \quad (7.28)$$

- b. Calculate total $E(m)$ for the observation period:

$$E(m) = \sum_{t=1}^n E(m)_t \quad (7.29)$$

If m_t are statistically independent (when trend is not included in the model), calculate total $Var(m)$ for the observation period:

$$Var(m) = \sum_{t=1}^n \alpha E(m)_t^2 = \alpha \sum_{t=1}^n E(m)_t^2 \quad (7.30)$$

If m_t are not statistically independent (when trend is included in the model), calculate total $Var(m)$ for the observation period:

$$Var(m) = \sum_{t=1}^n Var(m)_t + 2 \sum_{t=1}^n \sum_{j=t+1}^n Cov(m_t, m_j) \quad (7.31)$$

3. Calculate gamma distribution parameter a and b by using Equation (7.3).
4. Estimate GP distribution parameter λ for each entity from historical accident counts of the entity, for estimation by MEB method.
5. Calculate EB and MEB safety estimates $E(m|K)$ for the observation period by using Equations (7.9) and (7.21).
6. Calculate variances of the EB and MEB safety estimates $Var(m|K)$ by using Equations (7.10) and (7.22).

7.5.2 A calculation procedure for prediction of safety

The procedure to predict safety in the future period using both the EB and MEB methods, is as follows:

1. Calculate $E(m)'$ for a given future period, by using the fitted accident prediction model.
2. Calculate safety $E(m|K)'$ for the future period:

$$E(m|K)' = \frac{E(m)'}{n_r E(m)} \times n_r E(m|K) = \frac{E(m)'}{E(m)} \times E(m|K) \quad (7.30)$$

where n_r = ratio of the future period to the observation period

3. Calculate the variance of the prediction:

$$Var(m|K)' = \left(\frac{E(m)'}{n_r E(m)} \right)^2 \times (n_r)^2 Var(m|K) = \left(\frac{E(m)'}{E(m)} \right)^2 \times Var(m|K) \quad (7.31)$$

7.5.3 Calculation example

Estimation of safety for a given observation period

Data:

Observation period : 5 years (1999-2003)

Junction ID : X17

Type of junction : X-signalised junction

Total accident count during observation period $K = 11$

GP distribution parameter $\lambda = -0.131$

a. Estimation using the short-term (5-year) model for X-signalised junctions

Average daily junction flow during the period:

$$Q_p = 33,197 ; Q_s = 17,202$$

CLASS= 1 (major junction)

CITY = 0 (location is not in the city area)

Calculation of, $E(m)$ and $Var(m)$:

$E(m)_t$ values were calculated by putting the above values into the short-term (5-year) accident prediction model for X-signalised junctions given in Table 6.35 for signalised junctions:

$$E(m) = \mu = 1.137 \times 10^{-3} (Q_p Q_s)^{0.434} e^{(0.624CLASS - 0.649CITY)}$$

$$\text{Therefore, } E(m) = 13.299$$

The model had dispersion parameter $\alpha = 0.416$ (Table 6.8), and

$$Var(m) = \alpha E(m)^2 = 73.575$$

Calculation of EB and MEB safety estimates:

The gamma distribution parameters a and b :

$$a = \frac{E(m)}{Var(m)} = 0.181 ; b = \frac{E(m)^2}{Var(m)} = 2.404$$

The EB safety estimate:

$$E(m|K) = \frac{K + b}{1 + a} = 11.352$$

$$Var(m|K) = \frac{K + b}{(1 + a)^2} = 9.614$$

The MEB safety estimate:

$$E(m|K) = \frac{\int_0^{\infty} m^{b+1} e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} = 11.270$$

$$Var(m|K) = \frac{\int_0^{\infty} (m - E(m|K))^2 m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} = 7.769$$

Instead of using the above equations, the EB safety estimate can also be calculated by using the same formula for calculating MEB safety estimate by taking $\lambda = 0$. Both formulas would result in similar number.

b. Estimation using the simple long-term model for X-signalised junctions

The simple long-term model for X-signalised junctions was developed using year 1992 ($t = 1$) as base year. Therefore, $t = 8$ at year 1999 and so on.

Traffic flows at year 1999, 2000, 2001, 2002, 2003 were:

$$Q_{p_8} = 32,592 ; Q_{p_9} = 32,781 ; Q_{p_{10}} = 33,517 ; Q_{p_{11}} = 33,450 ; Q_{p_{12}} = 33,644$$

$$Q_{s_8} = 16,888 ; Q_{s_9} = 16,986 ; Q_{s_{10}} = 17,368 ; Q_{s_{11}} = 17,333 ; Q_{s_{12}} = 17,433$$

CLASS= 1 (major junction)

CITY = 0 (location is not in the city area)

Calculation of $E(m)_i$, $E(m)$, $Var(m)_i$ and $Var(m)$:

$E(m)_i$ values were calculated by putting the above values into the accident prediction model given in Table 6.12 for simple long-term model for X-signalised junctions:

$$E(m)_i = \mu_i = 1.371 \times 10^{-4} (Q_{p_i} Q_{s_i})^{0.479} (t)^{-0.117} e^{(0.324CLASS - 0.555CITY)}$$

Therefore,

$$E(m)_8 = 2.269 ; E(m)_9 = 2.250 ; E(m)_{12} = 2.270 ; E(m)_{11} = 2.241 ; E(m)_{12} = 2.230$$

$$\text{so, } E(m) = \sum_{i=8}^{12} E(m)_i = 11.260$$

$Var(m)_i = \alpha E(m)_i^2$ and $\alpha = 0.328$ (Table 6.3), then

$$Var(m)_8 = 1.688 ; Var(m)_9 = 1.660 ; Var(m)_{10} = 1.690 ; Var(m)_{11} = 1.646 ;$$

$$Var(m)_{12} = 1.631$$

$$\text{so, } \sum_{i=8}^{12} Var(m)_i = 8.315$$

Since a trend is included in the model then m_i are not statistically independent.

Therefore,

$$Var(m) = \sum_{i=8}^{12} Var(m)_i + 2 \sum_{i=8}^{12} \sum_{j=i+1}^{12} Cov(m_i, m_j)$$

with,

$$\begin{aligned} Cov(m_8, m_9) &= 0.758 ; Cov(m_8, m_{10}) = 0.765 ; Cov(m_8, m_{11}) = 0.755 ; \\ Cov(m_8, m_{12}) &= 0.751 ; Cov(m_9, m_{10}) = 0.758 ; Cov(m_9, m_{11}) = 0.749 ; \\ Cov(m_9, m_{12}) &= 0.755 ; Cov(m_{10}, m_{11}) = 0.752 ; Cov(m_{11}, m_{12}) = 0.742 \end{aligned}$$

$$\text{so, } \sum_{i=8}^{12} \sum_{j=i+1}^{12} Cov(m_i, m_j) = 7.529, \text{ and}$$

$$Var(m) = 23.373$$

Calculation of EB and MEB safety estimates:

The gamma distribution parameters a and b were calculated:

$$a = \frac{E(m)}{Var(m)} = 0.482 ; b = \frac{E(m)^2}{Var(m)} = 5.425$$

The EB safety estimate was calculated:

$$E(m|K) = \frac{K + b}{1 + a} = 11.085$$

$$Var(m|K) = \frac{K + b}{(1 + a)^2} = 7.481$$

The MEB safety estimate was calculated:

$$E(m|K) = \frac{\int_0^{\infty} m^{b+1} e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} = 11.050$$

$$Var(m|K) = \frac{\int_0^{\infty} (m - E(m|K))^2 m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} = 6.292$$

Prediction of safety for a given future period

Data:

The same set of data from the preceding example on estimation of safety was used.

Future period : 3 years (2004, 2005 and 2006)

a. Prediction using the short-term (5-year) model for X-signalised junctions

Average daily junction flow during the period:

$$Q_p = 35,987 ; Q_s = 18,648$$

Calculation of $E(m)'$:

Using the fitted accident prediction model:

$$E(m)' = 14.264 \text{ accidents in five years or } E(m)' = 8.558 \text{ accidents in three years}$$

Prediction of safety for year 2004-2006:

Prediction of safety using EB method:

$$E(m|K)' = \frac{E(m)'}{E(m)} \times E(m|K) = 7.305$$

$$Var(m|K)' = \left(\frac{E(m)'}{E(m)} \right)^2 \times Var(m|K) = 3.981$$

Prediction of safety using MEB method:

$$E(m|K)' = \frac{E(m)'}{E(m)} \times E(m|K) = 7.252$$

$$Var(m|K)' = \left(\frac{E(m)'}{E(m)} \right)^2 \times Var(m|K) = 3.217$$

b. Prediction using the simple long-term model for X-signalised junctions

Since 1992 ($t = 1$) is the base year, therefore, $t = 13$ at year 2004 and so on.

Traffic flows at year 2004, 2005 and 2006:

$$Q_{p_{13}} = 34,418 ; Q_{p_{14}} = 35,722 ; Q_{p_{15}} = 34,418$$

$$Q_{s_{13}} = 17,834 ; Q_{s_{14}} = 18,510 ; Q_{s_{15}} = 17,834$$

Calculation of $E(m)'$:

Using the fitted accident prediction model:

$$E(m)_{13} = 2.258 ; E(m)_{14} = 2.320 ; E(m)_{15} = 2.602$$

$$\text{so, } E(m) = \sum_{t=13}^{15} E(m)_t = 7.180$$

Prediction of safety for year 2004-2006:

Prediction of safety using EB method:

$$E(m|K)' = \frac{E(m)'}{E(m)} \times E(m|K) = 7.068$$

$$Var(m|K)' = \left(\frac{E(m)'}{E(m)} \right)^2 \times Var(m|K) = 3.041$$

Prediction of safety using MEB method:

$$E(m|K)' = \frac{E(m)'}{E(m)} \times E(m|K) = 7.046$$

$$Var(m|K)' = \left(\frac{E(m)'}{E(m)} \right)^2 \times Var(m|K) = 2.558$$

7.6 Results

The EB and MEB methods as illustrated above were applied to first estimate the safety over the 1999-2003 period pertaining to the set of junctions used in constructing the accident prediction model, and then to predict the safety of these junctions for the years 2004, 2005 and 2006. The fitted accident prediction models, which consist of short-term (5-year) models for X-signalised, T-signalised, X-unsignalised, and T-unsignalised junctions and also simple long-term models for X-signalised and T-signalised junctions as presented in Table 6.35 were used to calculate the safety of the reference population $E(m)$ and its variance $Var(m)$.

The GPD parameter λ for each site was estimated in two ways: estimated from the longest available period of historical accident record of respective junctions (between 5 to 12 years) and estimated from the 5-year (modelling period) historical accident record of respective junctions. For junctions with no recorded accident over these periods of estimation, a zero value was assumed for λ , which basically means that the underlying accident counts are Poisson-distributed. It is noted that from 1032 study junctions, the majority (93%) had accident records for as long as 12 years while the rest (7%) had accident records between 5 to 11 years. In the further analysis, the MEB estimation using λ that was predicted from the longest available accident record is called MEB1 method and the MEB estimation using λ that was estimated from the 5-year (modelling period) accident record is called MEB2 method. Examples of the result of safety estimation and prediction by EB and MEB methods are presented in Tables 7.1 and 7.2 while complete results are included in Appendices F and G.

Table 7.1 EB and MEB safety estimation over 1999-2003 period

No	ID	K	$E(m)$	$Var(m)$	a	b	EB estimates		MEB1 estimates			MEB2 estimates		
							$E(m K)$	$Var(m K)$	λ	$E(m K)$	$Var(m K)$	λ	$E(m K)$	$Var(m K)$
1	X1	11	10.409	45.068	0.231	2.404	10.889	8.846	0.081	(N=12)	10.890	0.171	10.902	11.836
2	X2	3	9.222	35.376	0.261	2.404	4.287	3.400	0.138	(N=12)	4.612	0.134	4.601	4.240
3	X3	7	7.567	23.818	0.318	2.404	7.137	5.416	-0.002	(N=12)	7.134	0.121	7.201	6.534
4	X4	9	9.133	34.694	0.263	2.404	9.028	7.146	-0.158	(N=12)	9.006	-0.604	1.789	8.568
5	X5	13	11.545	55.448	0.208	2.404	12.749	10.552	0.007	(N=12)	12.750	-0.063	12.769	9.530
6	X6	18	12.810	68.258	0.188	2.404	17.180	14.465	0.215	(N=12)	16.825	0.085	17.053	16.752
7	X7	19	20.204	169.804	0.119	2.404	19.128	17.094	0.121	(N=12)	19.181	-0.186	19.084	12.533
8	X8	18	13.418	74.888	0.179	2.404	17.304	14.675	-0.135	(N=12)	17.436	-0.366	18.580	-8.825
9	X9	25	14.373	85.936	0.167	2.404	23.477	20.113	-0.142	(N=12)	23.797	-0.426	0.804	3.740
10	X10	8	14.435	86.678	0.167	2.404	8.919	7.645	-0.195	(N=12)	8.657	0.072	9.049	8.668
11	X11	26	16.979	119.916	0.142	2.404	24.881	21.795	0.269	(N=12)	24.136	-0.052	24.967	19.919
12	X12	4	7.210	21.624	0.333	2.404	4.803	3.602	0.092	(N=12)	4.947	-0.249	4.545	2.525
13	X13	14	10.103	42.461	0.238	2.404	13.251	10.704	0.064	(N=12)	13.175	0.228	12.950	15.840
14	X14	0	11.482	54.841	0.209	2.404	1.988	1.644	0.185	(N=12)	2.348	na	1.988	1.645
15	X15	7	5.395	12.108	0.446	2.404	6.505	4.500	0.150	(N=12)	6.434	-0.323	6.633	3.064
16	X16	4	9.021	33.853	0.266	2.404	5.057	3.993	0.138	(N=12)	5.348	-0.069	4.945	3.590
17	X17	11	13.299	73.575	0.181	2.404	11.352	9.614	-0.131	(N=12)	11.270	-0.354	11.155	5.948
18	X18	0	4.628	8.910	0.519	2.404	1.582	1.041	0.034	(N=12)	1.619	na	1.583	1.042
19	X19	5	4.979	10.313	0.483	2.404	4.993	3.367	-0.093	(N=12)	4.976	0.000	4.993	3.367
20	X20	4	4.830	9.703	0.498	2.404	4.276	2.855	-0.361	(N=12)	4.157	-0.069	4.236	2.604

Notes:

- Calculations were based on short-term (5-year) model for X-signalised junctions
- MEB1 estimates: safety estimates based on λ that was estimated from accident counts data of the longest available period N (between 5 to 12 years)
- MEB2 estimates: safety estimates based on λ that was estimated from accident counts data of 5-year period (1999-2003)
- na in λ column means no λ can be estimated from the accident counts data due to no accident recorded during the period, therefore safety estimate was based on the assumption that $\lambda = 0$ (Poisson-distributed)

Table 7.2 EB and MEB safety prediction for 2004-2006 period

No	ID	$K_{(1999-2003)}$	K_{04-06}	$E(m)$	$E(m)'$	Predicted EB		Predicted MEB1		Predicted MEB2	
						$E(m K)'$	$Var(m K)'$	$E(m K)'$	$Var(m K)'$	$E(m K)'$	$Var(m K)'$
1	X1	11	2	10.409	6.698	7.007	3.663	7.008	4.191	7.015	4.901
2	X2	3	3	9.222	5.934	2.758	1.408	2.968	1.768	2.961	1.756
3	X3	7	2	7.567	4.869	4.592	2.243	4.591	2.235	4.634	2.706
4	X4	9	5	9.133	5.877	5.809	2.959	5.795	2.329	1.151	3.548
5	X5	13	13	11.545	7.430	8.204	4.370	8.205	4.422	8.217	3.946
6	X6	18	8	12.810	8.243	11.055	5.990	10.827	8.833	10.974	6.937
7	X7	19	22	20.204	13.001	12.309	7.079	12.343	8.880	12.281	5.190
8	X8	18	11	13.418	8.634	11.135	6.077	11.220	4.885	11.956	-3.654
9	X9	25	9	14.373	9.249	15.108	8.329	15.313	6.613	0.517	1.549
10	X10	8	7	14.435	9.289	5.739	3.166	5.571	2.310	5.823	3.589
11	X11	26	14	16.979	10.926	16.011	9.025	15.532	15.224	16.066	8.248
12	X12	4	3	7.210	4.640	3.091	1.491	3.183	1.718	2.925	1.046
13	X13	14	3	10.103	6.501	8.527	4.433	8.478	4.924	8.333	6.559
14	X14	0	6	11.482	7.389	1.279	0.681	1.511	0.950	1.279	0.681
15	X15	7	4	5.395	3.472	4.186	1.863	4.140	2.297	4.268	1.269
16	X16	4	6	9.021	5.805	3.254	1.653	3.441	2.075	3.182	1.487
17	X17	11	7	13.299	8.558	7.305	3.981	7.252	3.217	7.178	2.463
18	X18	0	3	4.628	2.978	1.018	0.431	1.042	0.451	1.018	0.431
19	X19	5	6	4.979	3.204	3.213	1.394	3.202	1.232	3.213	1.394
20	X20	4	5	4.830	3.108	2.751	1.182	2.675	0.721	2.726	1.078

Notes:

- Calculations were based on short-term (5-year) model for X-signalised junctions
- Predicted MEB1: safety prediction based on MEB estimates 1
- Predicted MEB2: safety prediction based on MEB estimates 2

Comparison of the results generated by the EB and MEB methods showed that the magnitude of safety $E(m|K)$ estimated by the MEB method is larger than, equal to or smaller than that estimated by the EB method as depending on the magnitude of λ , K and $E(m)$ as follows:

- λ is positive and $K < E(m)$, then $E(m|K)_{MEB} > E(m|K)_{EB}$
- λ is positive and $K > E(m)$, then $E(m|K)_{MEB} < E(m|K)_{EB}$
- λ is negative and $K < E(m)$, then $E(m|K)_{MEB} < E(m|K)_{EB}$
- λ is negative and $K > E(m)$, then $E(m|K)_{MEB} > E(m|K)_{EB}$
- λ equals zero, then $E(m|K)_{MEB} = E(m|K)_{EB}$

Further examination of the data revealed that, when λ was estimated from the longest available period (MEB1 method), 23 out of more than a thousand data points used to estimate safety in this research (based on short-term models) may experience MEB estimation problems as described previously. These problems include: $E(m|K)$ that does not lie between K and $E(m)$, negative $Var(m|K)$ and probability of $E(m|K) > E(m)$ exceeding one. Details of the problematic data points are presented in Table 7.3.

It can be seen that 8 data points (X168, T220, T272, T347, T414, T501, T577, T742) experienced single problem of probability of $E(m|K) > E(m)$ exceeding one. Further analysis showed that the problem for T272, T501 and T577 were not serious as the estimates of $E(m|K)$ were outside the fluctuating zone of estimation, as shown in Figure 7.7. A similar situation was observed for X168, T220, T347, T414 and T742 which means that the problem for those sites was also not serious. A probability value of one for those data points was then assumed.

Table 7.3 Data points with MEB estimation (based on short-term models) problems

No	ID	K	$E(m)$	$Var(m)$	λ	$E(m K)$	$Var(m K)$	$P(E(m K) > E(m))$
1	X168	4	2.104	1.842	-0.406	3.592	0.140	1.012
2	T220	2	0.356	0.136	-0.173	1.134	0.074	1.171
3	T272	2	0.199	0.042	-0.106	0.737	0.034	1.176
4	T347	2	0.223	0.053	-0.106	0.761	0.062	1.121
5	T414	4	0.292	0.092	-0.069	1.377	0.260	1.007
6	T501	2	0.199	0.042	-0.115	0.787	0.008	1.241
7	T577	2	0.199	0.042	-0.106	0.737	0.034	1.176
8	T742	2	0.397	0.169	-0.173	1.161	0.128	1.113
9	X39	12	5.676	13.401	-0.348	11.270	-0.076	1.043
10	T7	6	1.860	2.130	-0.322	5.496	-2.328	1.186
11	T233	4	0.356	0.136	-0.186	4.835	-13.127	2.718
12	T521	2	0.125	0.017	-0.106	0.995	-0.388	2.065
13	X24	23	13.937	80.802	-0.502	0.694	0.164	0.000
14	X93	7	2.795	3.251	-0.483	2.136	6.480	0.298
15	X109	14	10.011	41.685	-0.682	0.727	0.131	0.000
16	T634	7	1.109	1.319	-0.461	0.402	0.433	0.018
17	X57	10	8.852	32.593	-0.483	23.076	-287.429	1.624
18	X144	6	3.459	4.976	-0.483	8.212	-18.990	1.520
19	X190	14	15.450	99.290	-0.493	-3.125	-66.728	-0.075
20	T12	10	4.186	10.791	-0.454	-1.046	-15.979	-0.176
21	T74	2	0.488	0.147	-0.329	2.332	-2.245	2.421
22	T145	4	1.943	2.324	-0.406	4.157	-1.548	1.199
23	T157	10	2.826	4.917	-0.464	0.279	-1.545	-0.023

Figure 7.7 T272, T501 and T577 MEB estimate analysis,
 $K = 2$, $E(m) = 0.199$ and $Var(m) = 0.042$

Four data points (X39, T7, T233, T521) experienced problem of negative variance besides problem of probability of $E(m|K) > E(m)$ exceeding one. However, the estimates of $E(m|K)$ for those sites were lying between K and $E(m)$. The situations for those sites are illustrated in Figures 7.8 to 7.11.

Figure 7.8 X39 MEB estimate analysis,
 $K = 12$, $E(m) = 5.676$ and $Var(m) = 13.401$

Figure 7.9 T7 MEB estimate analysis,
 $K = 6$, $E(m) = 1.860$ and $Var(m) = 2.130$

Figure 7.10 T233 MEB estimate analysis,
 $K = 4$, $E(m) = 0.356$ and $Var(m) = 0.136$

Figure 7.11 T521 MEB estimate analysis,
 $K = 2$, $E(m) = 0.125$ and $Var(m) = 0.017$

It can be seen in Figure 7.8 that for $K = 12$, $E(m) = 5.676$ and $Var(m) = 13.401$, a serious problem of estimation may appear for $-0.5 < \lambda < -0.35$ where there were sudden changes in the estimates of $E(m|K)$ and $Var(m|K)$. At the λ of X39 ($\lambda = -0.348$), the estimates of $E(m|K)$ was still outside the fluctuation zone, so the problem was not serious. Therefore, $E(m|K)$ for X39 could still be estimated using their actual λ s and setting the negative variance to zero. On the other hand, the

problem for T7, T233 and T521 may be serious as indicated by Figures 7.9, 7.10 and 7.11. It can be seen in the figures that λ of T7 ($\lambda = -0.322$), T233 ($\lambda = -0.186$) and T521 ($\lambda = -0.106$) were lying at the fluctuating zone of estimation. Thus, it was decided to assume $\lambda = 0$ for T7, T233 and T521.

The next 4 data points (X24, X93, X109, T634) experienced the single problem of $E(m|K)$ value not lying between K and $E(m)$ bounds. The problem seems to be serious as the estimates of $E(m|K)$ were very far from K or $E(m)$. Seven more data points (X57, X144, X190, T12, T74, T145, T157) experienced serious error as having problems with the estimates of $E(m|K)$, negative $Var(m|K)$ and probability of $E(m|K) > E(m)$ exceeding one. Thus, the safety estimate for those sites was based on Poisson-distributed assumption.

When λ was estimated from 5-year accident data (1999-2003), there were 37 data points which may experience MEB estimation problems. These problems include: $E(m|K)$ that does not lie between K and $E(m)$, $Var(m|K)$ that is larger than $E(m|K)$, negative $E(m|K)$, negative $Var(m|K)$, and probability of $E(m|K) > E(m)$ exceeding one. As many as 23 data points were experiencing serious estimation problems in that the estimates $E(m|K)$ lie within the fluctuating zone, therefore the λ values for these 23 data points were set to zero. No serious problems were experienced by the remaining 14 data points.

However, when estimation of safety was based on the simple (long-term) accident prediction models (for X-signalised and T-signalised junctions only), the number of problematic data points were fewer than when the estimation was based on the short-term models, which thus indicated that the simple (long-term) models could be better than the short-term models. In this case, when λ was estimated from the longest available period of accident data, only 3 data points experienced serious problems (X144, X190, T12) while 1 data point (T7) experienced no serious problems. On the contrary, the number of problematic data points when estimation

was based on the short-term models for X- and T-signalised junctions were 13. When λ was estimated from 5-year accident data, 6 data points (X9, X144, X158, X194, T12, T183) experienced serious problems while 2 data points (T7 and T91) experienced no serious problems.

Therefore, based on the number of problematic data points, MEB1 method is more preferred over MEB2 method since MEB1 method resulted in fewer number of problematic data points. Also, the simple (long-term) accident prediction models are preferred over the short-term (5-year) accident prediction models as the former resulted in fewer number of problematic data points.

7.7 Validation of the results of prediction

The results of the estimation were used for predicting the expected number of accidents (safety) for year 2004, 2005 and 2006 according to the procedure explained earlier. For comparison purpose, the expected number of accidents was predicted according to 4 methods:

- Accident prediction model (APM) method, where $\hat{K} = E(m)'$
- Empirical Bayes (EB) method, where $\hat{K} = E(m|K)'_{EB}$
- Modified empirical Bayes 1 (MEB1) method, where λ was predicted from the longest available period of accident counts record and $\hat{K} = E(m|K)'_{MEB}$
- Modified empirical Bayes 2 (MEB2) method, where λ was predicted from accident counts record from the 5-year modelling period (1999-2003) and $\hat{K} = E(m|K)'_{MEB}$

The results of prediction by the 4 methods were then compared by using their sum-squares of error (*SSE*) and root-mean square error (*RMSE*) values. It is noted that:

- The better method shall have the smallest *SSE* and *RMSE* values
- The better method shall have the larger number of individual junctions with the smallest individual square error (*SE*) values

Table 7.4 presents the *SSE* and *RMSE* values of the 4 methods. It can be seen in the table that the EB, MEB1 and MEB2 methods performed better than the APM method in that they had smaller *SSE* and *RMSE* values than those for APM method. The MEB1 method performed better than the other methods for all type of junctions with exception for X-signalised junctions accident prediction based on the long-term model where the EB method performed better. On the other hand, the MEB2 method performed better than the EB method only for X-signalised and X-unsignalised junctions accident predictions based on the short-term models.

Table 7.4 Comparison of methods using *SSE* and *RMSE*

Models used	Junction types	Performance indicator	APM	EB	MEB1	MEB2
Long-term models	X-signalised	<i>SSE</i>	2874.389	2178.329	2232.051	2413.569
		<i>RMSE</i>	3.763	3.276	3.316	3.448
	T-signalised	<i>SSE</i>	650.039	536.496	527.314	554.358
		<i>RMSE</i>	1.890	1.717	1.702	1.745
Short-term models	X-signalised	<i>SSE</i>	2992.811	2331.203	2263.002	2301.057
		<i>RMSE</i>	3.840	3.389	3.339	3.367
	T-signalised	<i>SSE</i>	744.339	657.395	633.979	687.518
		<i>RMSE</i>	2.022	1.901	1.866	1.944
	X-unsignalised	<i>SSE</i>	162.882	130.915	112.980	120.914
		<i>RMSE</i>	1.515	1.358	1.261	1.305
	T-unsignalised	<i>SSE</i>	335.906	309.965	307.945	320.822
		<i>RMSE</i>	0.777	0.746	0.744	0.759

The comparison also indicated that, for X-signalised and T-signalised junctions, the long-term models were better than the short-term models as the predictions based on the long-term models gave generally lower *SSE* and *RMSE* values than those based on the short-term models.

Tables 7.5 and 7.6 show the performance of EB, MEB1 and MEB2 methods by comparing the individual *SE* values of EB method with MEB1 method and MEB2 method, respectively.

Table 7.5 Performance comparison of EB and MEB1 methods

Models used	Junction types	Comparison of squared error (SE)			
		$SE_{EB} > SE_{MEB1}$	$SE_{EB} = SE_{MEB1}$	$SE_{EB} < SE_{MEB1}$	Total
Long-term models	X-signalised	92	23	88	203
		45.3%	11.3%	43.3%	100%
	T-signalised	74	47	61	182
		40.7%	25.8%	33.5%	100%
Short-term models	X-signalised	101	19	83	203
		49.8%	9.4%	40.9%	100%
	T-signalised	78	32	72	182
		42.9%	17.6%	39.6%	100%
	X-unsignalised	25	39	7	71
		35.2%	54.9%	9.9%	100%
	T-unsignalised	80	399	78	557
		14.4%	71.6%	14.0%	100%

Table 7.6 Performance comparison of EB and MEB2 methods

Models used	Junction types	Comparison of squared error (SE)			
		$SE_{EB} > SE_{MEB2}$	$SE_{EB} = SE_{MEB2}$	$SE_{EB} < SE_{MEB2}$	Total
Long-term models	X-signalised	77	46	80	203
		37.9%	22.7%	39.4%	100%
	T-signalised	51	87	44	182
		28.0%	47.8%	24.2%	100%
Short-term models	X-signalised	80	46	77	203
		39.4%	22.7%	37.9%	100%
	T-signalised	53	91	38	182
		29.1%	50.0%	20.9%	100%
	X-unsignalised	10	54	7	71
		14.1%	76.1%	9.9%	100%
	T-unsignalised	29	491	37	557
		5.2%	88.2%	6.6%	100%

Table 7.5 shows that the MEB1 method performed better than the EB method as, for all types of junctions, the number of junctions having SE_{EB} greater than SE_{MEB1} were larger than the number of junctions having SE_{EB} smaller than SE_{MEB1} . Comparison of MEB2 method with EB method as shown in Table 7.6 may also indicated that the MEB2 method generally performed better than the EB method as, the number of junctions having SE_{EB} greater than SE_{MEB2} were larger than the number of junctions having SE_{EB} smaller than SE_{MEB2} for more of the junction types (with exception for X-signalised junctions predictions based on long-term model and T-unsignalised junctions).

Table 7.5 shows when the accident predictions were based on the long-term models, as many as 45.3% of the X-signalised junctions had SE_{EB} greater than SE_{MEB1} and as many as 43.3% of the X-signalised junctions had SE_{EB} smaller than SE_{MEB1} . For the case of T-signalised junctions, about 40.7% of the junctions had SE_{EB} greater than SE_{MEB1} and about 33.5% of the junctions had SE_{EB} smaller than SE_{MEB1} . When the accident predictions were based on the short-term models as many as 49.8% of the X-signalised junctions and 42.9% of the T-signalised junctions had SE_{EB} greater than SE_{MEB1} , and as many as 40.9% of the X-signalised junctions and 39.6% of the T-signalised junctions had SE_{EB} smaller than SE_{MEB1} .

The results also show that about 35.2% of the X-unsignalised junctions and 14.4% of the T-unsignalised junctions had SE_{EB} greater than SE_{MEB1} , and about 9.9% of the X-unsignalised junctions and 14% of the T-unsignalised junctions had SE_{EB} smaller than SE_{MEB1} . However, the proportion of the X-unsignalised and T-unsignalised junctions having $SE_{EB} = SE_{MEB1}$ was quite large (54.9% and 71.6%, respectively) as many of the unsignalised junctions had zero accidents during the study period. This had caused the λ value of these junctions to be assumed as equal to zero, which means that the estimation/prediction using MEB method will result in the same value as those using EB.

Therefore, from the above comparisons, the MEB1 method may be regarded as the better method to predict accident counts at the study junctions during year 2004 – 2006. The MEB1 method performed better than the APM, EB, and MEB2 methods such that it had generally the lowest SSE and $RMSE$ values. When it was compared to the EB method alone, the MEB1 method performed better such that it had more number of junctions with individual SE value lower than that of the EB method besides having generally lower SSE and $RMSE$ values. Although the MEB2 method also had generally more number of junctions with individual SE value lower than that of the EB method, it could not be entirely regarded as a better method than the EB method since its SSE and $RMSE$ values were higher than the EB method for more types of junction.

7.8 Summary

The EB and MEB methods can be considered as a refinement of accident prediction tools to yield a site-specific safety estimate. The differences between EB and MEB methods lie in the assumption on the distribution of the accident counts of an entity. In the EB method, the accident counts are assumed to be Poisson-distributed, which means that the λ parameter is restricted to be equal to zero whereas in the MEB method the accident counts are GP distributed, which means that λ can be smaller or greater than zero as depending on the inherent variation. This has become the main advantage of MEB over EB as no assumption regarding the variation of the data needs to be specified.

The magnitude of safety $E(m|K)$ that is estimated by MEB method will be larger than, equal to or smaller than that estimated by EB method depending on the magnitude of λ , K and $E(m)$ in the following manner:

- If λ is positive and $K < E(m)$, then $E(m|K)_{MEB} > E(m|K)_{EB}$
- If λ is positive and $K > E(m)$, then $E(m|K)_{MEB} < E(m|K)_{EB}$
- If λ is negative and $K < E(m)$, then $E(m|K)_{MEB} < E(m|K)_{EB}$
- If λ is negative and $K > E(m)$, then $E(m|K)_{MEB} > E(m|K)_{EB}$
- If λ equals zero then $E(m|K)_{MEB} = E(m|K)_{EB}$

Unlike the EB method, the use of MEB method does not necessarily guarantee that the value of $E(m|K)$ will be positioned between K and $E(m)$. Moreover, the estimate of $Var(m|K)$ may also be greater than the estimate of $E(m|K)$. However, these conditions have been shown to be not serious for positive λ .

When λ takes on a negative value, there is also a problem of overly high or low estimate of $E(m|K)$, negative variances and probability of $E(m|K) > E(m)$ that is more than unity which would be very serious. When such a condition occurs, the

Poisson assumption is suggested to be used instead. Thus, MEB estimates should be used carefully in order to avoid these potential problems.

Comparison of the EB and MEB methods by using *SSE* and *RMSE* values had indicated that the MEB1 method generally performed better than the EB and MEB2 methods. A similar conclusion was obtained when comparing the number of sites with smaller *SE* between EB, MEB1, and MEB2 methods whereby the MEB1 and MEB2 methods gave relatively more sites with smaller *SE* than the EB method. However, although the MEB2 method also had generally more number of junctions with individual *SE* value lower than that of the EB method, it could not be entirely regarded as a better method than the EB method since its *SSE* and *RMSE* values were higher than the EB method for more types of junction. The superiority of the MEB1 method relative to the MEB2 method indicated that it is better to use λ that is estimated from a longer available accident count sequence than λ that is estimated from a shorter period.

The comparisons have also indicated that, for X-signalised and T-signalised junctions, the long-term models were better than the short-term models as the predictions based on the long-term models gave generally lower *SSE* and *RMSE* values than those based on the short-term models.

Overall, the use of MEB seems to be able to improve the prediction of safety. More importantly, the use of MEB may reduce the possibility of making improper decision regarding the hazard level of a site as the actual variation of accident history is used in the MEB method. This issue shall be elaborated further in Chapter 8.

CHAPTER 8

IDENTIFICATION AND RANKING OF HAZARDOUS JUNCTIONS

8.1 Introduction

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

The accident prediction models and EB/MEB safety estimation procedures that have been developed earlier in this research can be applied to identify and rank hazardous sites (junctions) for safety investigation as part of a road safety programme. Moreover, these methods should also be useful in before-after evaluation for estimating the safety in the 'before' period and for predicting what safety would be in the 'after' period if no treatment had been applied to the treated sites.

Whereas the application of the EB/MEB methods to estimate and predict safety has been described previously in Chapter 7, this chapter presents the application of the MEB method to identify hazardous junctions. Methods for ranking the junction according to its hazard level for treatment purposes are also described and applied.

8.2 Review on current methods to identify and rank hazardous sites

In a road safety context, a hazardous site is commonly defined as any site that exhibits an accident potential that is significantly high when compared with some norm or average accident potential which is established from other sites with similar

characteristics. The identification of hazardous sites is important so as to avoid wasting resources due to treating the sites that are wrongly identified as unsafe sites and leaving the truly hazardous sites untreated.

The process of identification is usually followed by ranking the hazardous sites according to pre-specified criteria to measure the hazard level given limited budget. In this way, the authorities can determine which sites need to be prioritized for safety treatment. However, it is also common to make an ordered list of all the study sites, and the determination of the hazardous sites is then done based on the ranking result, i.e. the most hazardous sites are those sites with critical ranking.

The following sections review currently available methods used for identifying and ranking hazardous sites.

8.2.1 Methods to identify hazardous sites

Various methods are available to identify hazardous sites. The conventional ones may involve setting a preset limit for accident count or accident rate. In this context, a site is defined as hazardous if its observed accident count or accident rate or both exceed the preset limit. However, it is undoubtedly that this method is very sensitive to random variation in accident counts and to the regression-to-mean problem (Hauer, 1986; Elvik, 1997). Basically, the method will (incorrectly) identify hazardous sites as those sites which have high accident counts or accident rates.

Another approach is to use a critical threshold that is chosen in such a way that it is exceeded by only a small proportion of the sites. Thus, a site is said to be hazardous if its observed accident count or accident rate exceeds this critical threshold. An example of this kind of method is the rate-quality control method described in Stokes and Mutabazi (1996). The critical accident rate is calculated as:

$$R_c = \bar{R} + \delta \sqrt{\frac{\bar{R}}{Q}} + \frac{1}{2Q} \quad (8.1)$$

where

R_c = critical accident rate for a particular location

$\bar{R}$ = average accident rate for all road locations

Q = number of vehicles traversing a particular road section or entering a particular junction during the analysis period

δ = level of statistical significance

Saccomanno *et al.* (2001) defined hazardous site as a site where the observed number of accidents exceeds either the accident prediction (Poisson) model estimate $E(m)$ or EB estimate $E(m|K)$ by at least one standard deviation. They concluded that the EB estimate has yielded fewer hazardous sites than accident prediction (Poisson) model.

A more advanced method is to use a concept called probability of selecting the worst sites (PSS). A site is identified as hazardous if the probability of selecting the site as being the worst site is above a chosen criterion or threshold value. This kind of technique was done, for example, by Hauer and Persaud (1984) who derived the probability of selecting the worst sites for a Poisson-gamma model. Schluter *et al.* (1997) applied this concept using a hierarchical Bayesian model to select the worst sites. The probability of selecting the worst sites was based on three criteria: the first criterion is the posterior probability that the underlying accident rate for a site is larger than the underlying accident rates of the remaining sites, the second one is the probability that the expected number of accidents at a site in the future period will exceed some threshold. The third criterion is based on the expected number of future accidents being greater than some threshold value.

8.2.2 Methods to rank hazardous sites

As noted earlier, hazardous sites can be selected simply by ranking all the sites according to the hazard level where high-risk sites are assigned smaller ranks and

$$R_c = \bar{R} + \delta \sqrt{\frac{\bar{R}}{Q} + \frac{1}{2Q}} \quad (8.1)$$

where

R_c = critical accident rate for a particular location

$\bar{R}$ = average accident rate for all road locations

Q = number of vehicles traversing a particular road section or entering a particular junction during the analysis period

δ = level of statistical significance

Saccomanno *et al.* (2001) defined hazardous site as a site where the observed number of accidents exceeds either the accident prediction (Poisson) model estimate $E(m)$ or EB estimate $E(m|K)$ by at least one standard deviation. They concluded that the EB estimate has yielded fewer hazardous sites than accident prediction (Poisson) model.

A more advanced method is to use a concept called probability of selecting the worst sites (PSS). A site is identified as hazardous if the probability of selecting the site as being the worst site is above a chosen criterion or threshold value. This kind of technique was done, for example, by Hauer and Persaud (1984) who derived the probability of selecting the worst sites for a Poisson-gamma model. Schluter *et al.* (1997) applied this concept using a hierarchical Bayesian model to select the worst sites. The probability of selecting the worst sites was based on three criteria: the first criterion is the posterior probability that the underlying accident rate for a site is larger than the underlying accident rates of the remaining sites, the second one is the probability that the expected number of accidents at a site in the future period will exceed some threshold. The third criterion is based on the expected number of future accidents being greater than some threshold value.

8.2.2 Methods to rank hazardous sites

As noted earlier, hazardous sites can be selected simply by ranking all the sites according to the hazard level where high-risk sites are assigned smaller ranks and

vice versa. The worst sites are then selected from the outcome of the ranking; the sites with smaller ranking are regarded as the more hazardous sites. The simplest way to rank the sites is according to their accident count or accident rate. The site that has the highest count or rate is considered as the worst site. However, as mentioned before, the method is very sensitive to random variation in the accident counts and regression-to-mean problem. To overcome this problem, the expected number of accidents is suggested to be used as the criterion for ranking.

McGuigan (1981) proposed ranking sites according to their potential for accident reduction (PAR), which is the difference between the reported (actual) number of accidents at a site and the expected number of accidents at sites with similar characteristics. Soon after, McGuigan (1982) postulated that using PAR in ranking hazardous sites is more likely to maximise the cost-effectiveness of a road safety improvement programme than using accident count or accident rate. On the other hand, Maher and Mountain (1988) concluded that using accident count as ranking criterion may perform as well as or better than using PAR due to inaccuracy in the estimation of expected number of accidents at a site which is required in the PAR method.

After safety estimation by EB method has become popular, EB estimate has been used as a criterion for the purpose of ranking sites. Hauer (1996) used the EB safety estimate directly to rank sites. Persaud *et al.* (1999) used it in a method called potential for safety improvement (PSI). This method is quite similar to the PAR method, except that the EB safety estimate is used instead of accident count. Here, PSI is estimated as the difference between the EB safety estimate $E(m|K)$ and what is normal for similar sites $E(m)$.

In the PSI method, Persaud *et al.* (1999) proposed three alternatives in calculating the PSI based on the definition of $E(m)$. In the first alternative, $E(m)$ is calculated from the best possible regression model that is also used in the estimation of m . In the second alternative, $E(m)$ is calculated from the regression model for a base condition, which reflects roads built to a standard level of safety. In the third

alternative, $E(m)$ is calculated from the regression model that contains only traffic flow as the independent variable. They subsequently evaluated the three alternatives of the PSI method with the method of ranking by the accident count, accident rate, and EB method. The evaluation was based on the total number of accidents for the 20 worst sites identified by each respective method. In addition, the evaluation was also carried out based on the number of treatable accidents, which is defined as the positive difference between the actual safety of the sites and what is normal for the sites with the same traffic flow, as the parameter for comparison purpose. The results indicated that the PSI method is as promising as direct ranking by EB estimate.

A different approach was adopted by numerous other researchers in selecting sites for treatment by considering economic criteria, such as Persaud and Kazakov (1994). In their method, the ranking criterion was based on the net economic benefit, which is the difference between the safety benefit and treatment cost. Here, a site may be selected for treatment if the economical benefits of treating the sites exceed a threshold value based on the allocated budget.

8.2.3 Conclusions on review

From the review, it can be concluded that the easiest method to identify a hazardous site is by setting a preset limit for accident count or accident rate. Then, the selection of sites for treatment priority is merely based on the result of ranking according to the specified criterion. However, this kind of technique will not give any certainty that the worst sites have indeed been selected.

There shall be more rigour if the identification of hazardous sites is done based on the concept of probability in selecting the worst site (PSS). Upon identification of hazardous sites, prioritization of sites for treatment may still be based on ranking the sites according to some preset criteria. An approach which is based on MEB estimate is proposed herewith.

8.3 The application of MEB estimate for identifying and ranking of hazardous sites

The MEB approach consists of two steps: step one includes identification of hazardous sites based on the probability of selecting a site with accident potential exceeding what is normal for sites similar to the site being investigated, and step two involves ranking of the hazardous sites identified in step one for priority treatment.

Given a site with accident potential represented by $E(m|K)$ - mean of the posterior distribution, and a threshold value represented by $E(m)$ - mean of the prior distribution. Here, $E(m|K)$ or $\hat{m}$ is the MEB estimate. The site is identified as hazardous if there is a significant probability that its accident potential exceeds the value that is normal for similar sites in the reference population. So, mathematically, a site is identified as hazardous if:

$$\left[P(E(m|K) > E(m)) = 1 - \frac{\int_0^{E(m)} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm}{\int_0^{\infty} m^b e^{-m(1-\lambda+a)} (m(1-\lambda) + \lambda K)^{K-1} dm} \right] \geq \delta \quad (8.2)$$

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

Three criteria that make use of the MEB estimate were then considered for ranking the hazardous sites for treatment priority, which are:

- Criterion A: Rank the sites according to the MEB estimate
- Criterion B: Rank the sites according to the potential for safety improvement (PSI)
- Criterion C: Rank the sites according to the level of hazard (H)

It should be noted that there is a slight difference between the PSI method proposed in this research with the one proposed by Persaud *et al.* (1999) in that the MEB estimate was used in lieu of the EB estimate. Thus, PSI is defined as the positive difference between MEB estimate $E(m|K)$ and the normal accident potential for similar sites in the reference population $E(m)$, which is:

$$PSI = E(m|K) - E(m) \quad (8.3)$$

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

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

$$H = \frac{E(m|K)}{E(m)} \quad (8.4)$$

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

8.4 Analysis and result

The method presented in Section 8.3, which are based on MEB estimation, were applied to identify and rank hazardous junctions in the study area over the 1999-2003 period. The λ value was estimated using the longest available period of accident record. The safety $E(m|K)$ predictions for X-signalised and T-signalised junctions were based on the long-term models since they had been shown to perform better than the short-term models, as discussed previously in Chapter 7. The methods identified a total of 59 junctions as being hazardous based on the criterion that the probability of $E(m|K)$ exceeding $E(m)$ is greater than 0.95, which

consisted of 21 X-signalised junctions, 11 T-signalised junctions, 4 X-unsignalised junctions and 23 T-unsignalised junctions. The hazardous junctions are listed in Tables 8.1 to 8.4 for X-signalised, T-signalised, X-unsignalised and T-unsignalised junctions, respectively. The number of hazardous sites increased to 88 if the accepted confidence level was lowered to 0.90.

For comparison, a similar computation of the probability in selecting the hazardous site by the EB method was also done. The probability for the EB method can be computed from Equation (8.2) by taking λ equal to zero. The probability of selecting the hazardous site by the EB method for the 59 hazardous junctions identified by MEB method is also presented in Tables 8.1 to 8.4.

Table 8.1 List of hazardous X-signalised junctions identified by MEB method

No	ID	λ	$E(m)$	MEB method		EB method	
				$E(m K)$	$P(E(m K) > E(m))$	$E(m K)$	$P(E(m K) > E(m))$
1	X24	-0.502	11.858	19.743	0.992	21.529	0.997
2	X109	-0.447	8.229	11.755	0.991	9.398	0.991
3	X39	-0.348	6.338	9.042	0.986	38.978	1.000
4	X134	-0.289	3.243	6.252	0.958	9.056	1.000
5	X181	-0.219	5.308	8.362	0.985	6.714	0.999
6	X89	-0.217	5.342	10.546	1.000	19.743	0.992
7	X9	-0.142	12.268	22.188	1.000	10.617	0.879
8	X8	-0.135	11.371	16.237	0.962	10.026	0.997
9	X51	-0.131	10.279	14.760	0.964	28.079	1.000
10	X192	-0.093	19.031	27.986	0.986	16.978	0.989
11	X25	-0.049	3.172	6.253	0.955	8.118	0.873
12	X121	0.007	10.430	16.940	0.988	27.815	0.974
13	X153	0.020	16.561	38.835	1.000	7.730	0.923
14	X107	0.066	5.172	9.568	0.990	30.396	1.000
15	X145	0.094	2.383	8.874	0.998	8.130	0.998
16	X27	0.107	1.334	6.646	0.995	9.218	0.993
17	X146	0.133	4.627	8.524	0.965	14.366	0.929
18	X94	0.225	2.488	7.666	0.970	21.618	1.000
19	X158	0.256	13.794	25.800	0.990	15.888	0.928
20	X117	0.293	11.057	18.607	0.964	6.175	0.935
21	X113	0.397	9.649	18.142	0.978	6.195	0.941

Table 8.2 List of hazardous T-signalised junctions identified by MEB method

No	ID	λ	$E(m)$	MEB method		EB method	
				$E(m K)$	$P(E(m K) > E(m))$	$E(m K)$	$P(E(m K) > E(m))$
1	T7	-0.322	2.069	4.556	1.000	3.685	0.928
2	T157	-0.253	2.819	5.320	1.000	4.342	0.960
3	T107	-0.235	9.463	19.388	1.000	18.688	0.998
4	T4	-0.206	5.249	16.190	1.000	14.179	1.000
5	T128	-0.166	10.102	15.167	0.968	14.927	0.931
6	T67	-0.108	0.611	4.640	0.999	4.568	0.999
7	T92	-0.063	1.499	3.833	0.979	3.727	0.961
8	T158	-0.035	8.403	13.501	0.967	13.419	0.959
9	T183	-0.035	10.730	16.854	0.970	16.787	0.964
10	T130	0.033	8.144	14.655	0.986	14.774	0.990
11	T39	0.226	0.402	3.729	0.975	3.791	0.999

Table 8.3 List of hazardous X-unsignalised junctions identified by MEB method

No	ID	λ	$E(m)$	MEB method		EB method	
				$E(m K)$	$P(E(m K) > E(m))$	$E(m K)$	$P(E(m K) > E(m))$
1	X244	-0.115	0.593	1.590	0.971	1.436	0.917
2	X205	-0.098	1.860	4.653	0.990	4.461	0.999
3	X254	-0.063	0.593	2.734	1.000	2.487	0.971
4	X217	0.116	1.050	3.462	0.980	3.956	0.998

Table 8.4 List of hazardous T-unsignalised junctions identified by MEB method

No	ID	λ	$E(m)$	MEB method		EB method	
				$E(m K)$	$P(E(m K) > E(m))$	$E(m K)$	$P(E(m K) > E(m))$
1	T220	-0.173	0.356	1.134	1.000	0.810	0.849
2	T742	-0.173	0.397	1.161	1.000	0.876	0.839
3	T645	-0.170	0.707	1.859	0.956	1.696	0.909
4	T689	-0.158	1.059	4.731	0.997	4.217	0.999
5	T431	-0.156	1.027	4.672	0.997	4.156	0.999
6	T501	-0.115	0.199	0.787	1.000	0.516	0.885
7	T272	-0.106	0.199	0.737	1.000	0.516	0.885
8	T347	-0.106	0.223	0.761	1.000	0.566	0.880
9	T577	-0.106	0.199	0.737	1.000	0.516	0.885
10	T244	-0.106	0.356	0.944	0.982	0.810	0.849
11	T607	-0.106	0.732	1.858	0.957	1.730	0.905
12	T510	-0.093	1.261	4.225	0.999	3.985	0.992
13	T457	-0.076	2.178	4.972	0.977	4.854	0.958
14	T414	-0.069	0.292	1.377	1.000	1.177	0.991
15	T725	-0.049	0.344	1.146	0.984	1.059	0.955
16	T386	-0.049	0.199	0.578	0.966	0.516	0.885
17	T210	-0.045	0.613	3.787	1.000	3.542	1.000
18	T634	0.000	1.109	4.309	0.999	4.309	0.999
19	T233	0.000	0.356	1.363	0.989	1.363	0.989

Table 8.4 List of hazardous T-unsignalised junctions identified by MEB method (continuation)

No	ID	λ	$E(m)$	MEB method		EB method	
				$E(m K)$	$P(E(m K) > E(m))$	$E(m K)$	$P(E(m K) > E(m))$
20	T406	0.034	0.615	2.249	0.987	2.358	0.994
21	T214	0.043	1.097	3.105	0.968	3.206	0.981
22	T435	0.114	0.397	1.940	0.987	2.667	1.000
23	T250	0.114	1.261	3.648	0.963	3.985	0.992

Tables 8.1 to 8.4 show when the λ value is negative (index of dispersion less than one), the probability of $E(m|K)$ exceeding $E(m)$ when assuming Poisson distribution (EB method) will be lower than when the MEB method is used. This implies that the site may, in a probabilistic sense, be detected as being less hazardous than it is. Similarly, when the λ value is positive (index of dispersion greater than one), then the probability that $E(m|K)$ exceeding $E(m)$ when assuming Poisson assumption will be larger than when the MEB method is used, which means that the site may be mistakenly identified as being more hazardous than it is.

The above finding is supported by Nicholson (1985), who stated that the use of Poisson assumption when historical accident counts are fairly regular (index of dispersion less than one) will cause the probability of taking no action when it is warranted to be greater as the change in underlying true rate cannot be detected. On the other hand, the use of Poisson assumption when historical accident counts are fairly irregular (index of dispersion greater than one) will cause the probability of taking some actions more than it is warranted to be greater because of the mistake in assuming an accident count fluctuation for a change in the underlying true rate.

Thus, using a Poisson assumption when the actual variation of the historical accident count data is greater or less than required by the Poisson distribution (index of dispersion is greater or less than unity) may have implication towards making incorrect decisions when ranking hazardous locations. In fact, only 32 out of the 59 hazardous junctions identified by the MEB method will also be identified as being hazardous by the EB method. Additional analysis also revealed that there

were 22 other junctions that were not identified as being hazardous by the MEB method but were identified as being hazardous by the EB method.

The 59 hazardous junctions were then ranked according to the three criteria described above and the result is presented in Table 8.5. The results show that the ranking by criterion A is not much different from that by criterion B. However, criterion C gave a very different ranking compared to criteria A and B. The effectiveness of the three criteria was evaluated based on the number of accidents that occurred at the most hazardous junctions during 2004-2006 period.

Tables 8.6 and 8.7 show the number of accidents that occurred during years 2004-2006 (in total accidents/year) at the 5 most hazardous junctions and the 10 most hazardous junctions, respectively. The results indicate that criterion C is not as good as criteria A and B in detecting the most hazardous junctions. Criteria A and B show comparable performance as they were able to detect the top hazardous junctions with the largest number of accidents in 1999-2003 period and 2004-2006 period. This was due to the fact that the top 5 and 10 hazardous junctions lists by using criteria A and B were dominated by signalised junctions while those using criteria C was dominated by unsignalised junctions which had lesser accidents than signalised junctions.

Tables 8.8 to 8.14 show the number of accidents that occurred during years 2004-2006 (in total accidents/year) at the 5 most hazardous junctions and the 10 most hazardous junctions for X-signalised, T-signalised, X-unsignalised and T-unsignalised junctions. It should be noted that there were only 4 hazardous X-unsignalised junctions, therefore the data were summarised in only one table. It can be seen that there was lesser discrepancy in the results of the number accidents occurring at the top 5 and 10 most hazardous junctions between criteria A, B and criterion C. Criteria A and B were still able to detect the top hazardous junctions with the largest number of accidents in 1999-2003 period and 2004-2006 period. It is therefore recommended to use either criteria A or B when ranking hazardous sites in the Singapore road network.

Table 8.5 Ranking of hazardous junctions

No	ID	K (1999-2003)	$K_{0.05}$	$P(E(m K) > E(m))$	$E(m)$	$E(m K)$	PSI	H	Criterion A Rank by $E(m K)$	Criterion B Rank by PSI	Criterion C Rank by H
1	X153	43	9	1.000	16.561	38.835	22.274	2.345	1	1	34
2	X192	29	14	0.986	19.031	27.986	8.955	1.471	2	6	55
3	X158	32	21	0.990	13.794	25.800	12.006	1.870	3	2	43
4	X9	25	9	1.000	12.268	22.188	9.920	1.809	4	5	46
5	X24	23	5	0.992	11.858	19.743	7.885	1.665	5	8	49
6	T107	21	8	1.000	9.463	19.388	9.925	2.049	6	4	37
7	X117	27	12	0.964	11.057	18.607	7.550	1.683	7	9	48
8	X113	47	19	0.978	9.649	18.142	8.493	1.880	8	7	42
9	X121	21	15	0.988	10.430	16.940	6.510	1.624	9	11	50
10	T183	18	5	0.970	10.730	16.854	6.124	1.571	10	13	53
11	X8	18	11	0.962	11.371	16.237	4.866	1.428	11	19	58
12	T4	23	11	1.000	5.249	16.190	10.941	3.084	12	3	22
13	T128	16	10	0.968	10.102	15.167	5.065	1.501	13	18	54
14	X51	17	13	0.964	10.279	14.760	4.481	1.436	14	20	56
15	T130	17	10	0.986	8.144	14.655	6.511	1.799	15	10	47
16	T158	15	9	0.967	8.403	13.501	5.098	1.607	16	17	51
17	X109	14	7	0.991	8.229	11.755	3.526	1.429	17	26	57
18	X89	15	14	1.000	5.342	10.546	5.204	1.974	18	15	38
19	X107	16	6	0.990	5.172	9.568	4.396	1.850	19	21	44
20	X39	12	5	0.986	6.338	9.042	2.704	1.427	20	36	59

Table 8.5 Ranking of hazardous junctions (continuation)

N ₀	ID	K (1999-2003)	K ₀₄₋₀₅	$P(E(m K) > E(m))$	$E(m)$	$E(m K)$	PSI	H	Criterion A Rank by $E(m K)$	Criterion B Rank by PSI	Criterion C Rank by H
21	X145	10	5	0.998	2.383	8.874	6.491	3.723	21	12	13
22	X146	13	4	0.965	4.627	8.524	3.897	1.842	22	23	45
23	X181	11	7	0.985	5.308	8.362	3.054	1.575	23	31	52
24	X94	9	2	0.970	2.488	7.666	5.178	3.082	24	16	23
25	X27	7	9	0.995	1.334	6.646	5.312	4.984	25	14	4
26	X25	7	1	0.955	3.172	6.253	3.081	1.971	26	30	39
27	X134	7	4	0.958	3.243	6.252	3.009	1.928	27	32	40
28	T157	10	2	1.000	2.819	5.320	2.501	1.887	28	37	41
29	T457	6	2	0.977	2.178	4.972	2.794	2.283	29	34	35
30	T689	7	1	0.997	1.059	4.731	3.672	4.468	30	24	9
31	T431	7	5	0.997	1.027	4.672	3.645	4.551	31	25	8
32	X205	6	3	0.990	1.860	4.653	2.793	2.502	32	35	33
33	T67	5	0	0.999	0.611	4.640	4.029	7.597	33	22	2
34	T7	6	3	1.000	2.069	4.556	2.487	2.203	34	38	36
35	T634	7	2	0.999	1.109	4.309	3.200	3.885	35	28	11
36	T510	6	1	0.999	1.261	4.225	2.964	3.351	36	33	18
37	T92	5	2	0.979	1.499	3.833	2.334	2.556	37	41	31
38	T210	8	1	1.000	0.613	3.787	3.174	6.178	38	29	3
39	T39	4	0	0.975	0.402	3.729	3.327	9.285	39	27	1
40	T250	6	2	0.963	1.261	3.648	2.387	2.894	40	40	26

Table 8.5 Ranking of hazardous junctions (continuation)

No	ID	K (1999-2003)	$K_{0.05}$	$P(E(m K) > E(m))$	$E(m)$	$E(m K)$	PSI	H	Criterion A Rank by $E(m K)$	Criterion B Rank by PSI	Criterion C Rank by H
41	X217	7	0	0.980	1.050	3.462	2.412	3.297	41	39	20
42	T214	5	1	0.968	1.097	3.105	2.008	2.832	42	43	27
43	X254	6	4	1.000	0.593	2.734	2.141	4.611	43	42	7
44	T406	5	1	0.987	0.615	2.249	1.634	3.656	44	44	16
45	T435	8	2	0.987	0.397	1.940	1.543	4.886	45	45	5
46	T645	3	2	0.956	0.707	1.859	1.152	2.629	46	46	30
47	T607	3	1	0.957	0.732	1.858	1.126	2.537	47	47	32
48	X244	3	3	0.971	0.593	1.590	0.997	2.682	48	50	28
49	T414	4	1	1.000	0.292	1.377	1.085	4.711	49	48	6
50	T233	4	3	0.989	0.356	1.363	1.007	3.827	50	49	12
51	T742	2	1	1.000	0.397	1.161	0.764	2.924	51	53	24
52	T725	3	2	0.984	0.344	1.146	0.802	3.332	52	51	19
53	T220	2	0	1.000	0.356	1.134	0.778	3.184	53	52	21
54	T244	2	0	0.982	0.356	0.944	0.588	2.651	54	55	29
55	T501	2	1	1.000	0.199	0.787	0.588	3.953	55	54	10
56	T347	2	0	1.000	0.223	0.761	0.538	3.415	56	56	17
57	T272	2	1	1.000	0.199	0.737	0.538	3.702	57	57	14
58	T577	2	0	1.000	0.199	0.737	0.538	3.702	58	58	15
59	T386	2	0	0.966	0.199	0.578	0.379	2.903	59	59	25

Table 8.6 Evaluation results for the 5 most hazardous junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	30.4	19.3
2	B	PSI	28.8	19.3
3	C	H	6.4	4.0

Table 8.7 Evaluation results for the 10 most hazardous junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	57.2	39.0
2	B	PSI	57.4	39.3
3	C	H	11.6	8.0

Table 8.8 Evaluation results for the 5 most hazardous X-signalised junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	30.4	19.3
2	B	PSI	35.2	24.0
3	C	H	16.8	13.0

Table 8.9 Evaluation results for the 10 most X-signalised hazardous junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	56.4	42.7
2	B	PSI	52.8	39.3
3	C	H	38.6	30.0

Table 8.10 Evaluation results for the 5 most hazardous T-signalised junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	19.0	14.7
2	B	PSI	18.8	14.3
3	C	H	8.6	5.3

Table 8.11 Evaluation results for the 10 most T-signalised hazardous junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	27.2	20.0
2	B	PSI	27.0	19.3
3	C	H	24.8	16.7

Table 8.12 Evaluation results for the hazardous X-unsignalised junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	4.4	3.3
2	B	PSI	4.4	3.3
3	C	H	4.4	3.3

Table 8.13 Evaluation results for the 5 most hazardous T-unsignalised junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	6.6	3.7
2	B	PSI	7.0	3.3
3	C	H	6.8	3.3

Table 8.14 Evaluation results for the 10 most T-unsignalised hazardous junctions

No	Criteria	Ranking criteria	Total K (accidents/year)	Total K_{04-06} (accidents/year)
1	A	$E(m K)$	13.0	6.0
2	B	PSI	13.0	6.3
3	C	H	10.2	5.7

8.5 Summary

The identification and ranking of hazardous junctions is very important towards being able to select the junctions for priority treatment, given limited budget. The identification method was developed by applying the MEB method to estimate safety. Under the MEB method, actual variability in the accident count data is used in safety estimation which thus reduces the potential of making incorrect decision in identifying and selecting hazardous junctions for safety remediation works. The ranking criteria were also set based on the MEB estimates. Three criteria were proposed to rank the hazardous junctions, which were the MEB estimate $E(m|K)$, potential for safety improvement (PSI) and level of hazard (H).

Using the MEB method, 59 hazardous junctions were identified as hazardous, and were ranked according to the $E(m|K)$, PSI and H values. The results showed that there was only a slight difference in the ranking based on $E(m|K)$ and PSI indices.

However, ranking based on H value resulted in clear discrepancy with the results of the other two criteria. Also, the H index seems not to be able to detect the most hazardous junctions as well as the other two criteria. Therefore, the H criterion is not recommended to be used for ranking hazardous sites, and ranking based on $E(m|K)$ and PSI criteria are preferred.

CHAPTER 9

CONCLUSIONS AND RECOMMENDATIONS

9.1 Conclusions

Road safety is commonly associated with accident occurrences from road traffic activities and their harmful effects. A limited few in-depth studies have pointed to the human factors as the principal “causatory” factors in the majority of accidents. However, given the variable nature of human characteristics and behaviours, it is not easy to study the effect of the human factors on accident occurrences nor are there ready remediation actions that can be applied effectively on the road users. A common accident remediation strategy has thus been to apply engineering measures on the road infrastructure in order to reduce the likelihood of road users making errors and to make the road environment to be more error-tolerant. Hence, road safety can be tackled by studying the associative relationship between accident occurrences and the more controllable accident contributing factors such as roadway characteristics, traffic and operational conditions, in the quest to make the road infrastructure as adaptive (in minimising scope for errors) and as forgiving (be error-tolerant) as possible. Of course, such a strategy, which is the approach adopted in this study, necessarily entails some degree of empiricism that also involves a substantial amount of data assembly for mass data analysis.

This research was undertaken with the aim of estimating safety at road junctions. It is appreciated that safety should be assessed not only at the road junctions but also along road links, that is, at road network level. However, the nature of accident occurrences at road links is also very complex which shall broaden the scope of the present research beyond available resources.

Accident prediction models were developed in the form of associative relationships between accident occurrences and roadway characteristics, traffic and operational factors at the junctions. Though the models did not cover the human factors,

considerations were given to the different driving situations encountered by drivers through classifying the study sites into X-signalised, T-signalised, X-unsignalised and T-unsignalised junctions. The accident prediction models give information regarding the safety of the reference population. The safety of the reference population was then combined with information on the historical accident counts of an entity (belonging to the reference population) by means of the original empirical Bayes (EB) method and a modified empirical Bayes (MEB) method (using GP distribution as an alternative to Poisson distribution). Procedures for developing accident prediction models and safety estimation methods were established and applied to Singapore road junctions. The calibrated methods are necessarily limited; an objective is then to test whether the methods as calibrated, though limiting in scope, can deliver a better estimation.

The accident prediction models, for which the outputs were used by the EB/MEB methods, were developed using a negative binomial (NB) regression model. In general, there were two types of accident prediction models that were developed in this research, which were long-term models (models based on long-term 1992-2003 data) and short-term models (models based on short-term 1999-2003 data). The models for X-signalised and T-signalised junctions comprised of long-term models and short-term models while no long-term models were developed for X-unsignalised and T-unsignalised junctions due to unavailability of data.

The long-term models, which consisted of simple long-term model and full long-term model, were formulated by including long-term trend in accident occurrences during 1992-2003 period. The results showed that, generally, accident rates in the study junctions during 1992-2003 period had been decreasing annually by a power trend factor of 11.7% - 12.1% and 15.8% - 19% for X-signalised junctions and T-signalised junctions, respectively.

No statistically significant trends were found for study junctions (except T-unsignalised junctions) during 1999-2003 period, and short-term models were thus formulated by considering constant accident rate during the 1999-2003 period.

Several models were developed by varying the modelled period between 1 year' and 5 years' period such that all the modelled periods started from year 2003 backwards.

The performance of the long-term and short-term models was compared and the better model(s) was/were selected based on the performance of the models when they were used to estimate accident frequencies during their estimation years and to predict accident frequencies at prediction years of 2004, 2005 and 2006.

Analysis revealed that the short-term (5-year) models were found to have generally better performance than the other models, with exception for T-signalised junction models in which the long-term models performed better. Among the short-term models for T-signalised junctions, the short-term (5-year) model was the better model. The short-term (5-year) models and the simple (long-term) models (for X-signalised and T-signalised junctions only) were chosen for estimating and predicting safety at road junctions by empirical Bayes (EB) and modified empirical Bayes (MEB) methods. In the EB and MEB analysis, it was found that, for X-signalised and T-signalised junctions, the long-term models were better than the short-term models as the predictions based on the long-term models gave generally lower sum-squares of error (*SSE*) and root-mean square error (*RMSE*) values than those based on the short-term models.

The research has revealed that, while it is commonly assumed that within a site accident count data are following a (standard) Poisson distribution, the generalised Poisson (GP) distribution is a better fitting model such that, in general, the goodness-of-fit statistics (*SSE* and χ^2) were significantly improved. Also, the GP distribution has the capability to cater to data that are over-dispersed or under-dispersed as is often the case for accident count data.

A major research task has involved modification of the EB method for estimating safety. The modification entailed using the GP distribution to model the accident count data instead of the Poisson distribution used in the original EB method. In the

EB/MEB method, the part of information about an entity's safety as concerning how safety varies in a reference population of similar entities is provided by the accident prediction model. In the absence of any other knowledge about the entity, then the safety of a reference population is also taken as the safety of the entity. However, when historical accident count data of the entity are available, then the EB/MEB method refines the safety estimate. Under the EB/MEB method, safety varies in a subpopulation of entities having not only similar traffic and geometric characteristics but also similar historical accident count data. Thus the EB/MEB safety estimate can be viewed as the result of correcting the output of the accident prediction model to yield a site-specific safety estimate.

The main advantage of MEB over EB method is that MEB method does not require implicit assumption regarding the variation of the historical accident count data as the GP distribution can cater to both under-dispersed ($\lambda < 0$) and over-dispersed ($\lambda > 0$) data as well as equi-dispersed ($\lambda = 0$) data whereas EB method requires the accident count data to be Poisson-distributed (equi-dispersed). More importantly, the use of MEB method can reduce the likelihood of making improper decision regarding the hazard level of a site as the actual variability of accident count history is used.

The magnitude of safety $E(m|K)$ that is estimated by MEB method being larger than, equal to or smaller than that estimated by EB method was found to depend on the magnitude of λ , K and $E(m)$. It was shown that the value of $E(m|K)$ that was estimated by MEB method is not always positioned between K and $E(m)$. In some cases, the estimate of $Var(m|K)$ may also be greater than the estimate of $E(m|K)$. There also exists conditions (for negative λ) of overly high or low estimate of $E(m|K)$, negative variances and the probability of $E(m|K) > E(m)$ being more than unity, which can be problematic. When such conditions occur, the Poisson assumption is suggested to be used in conjunction with the EB method. Nevertheless, it was found that only a small proportion (1.4%) of the junction sites (in this study) had this serious problem.

The research has shown that, in general, the use of the MEB with λ estimated from the longest available period (MEB1) method is able to improve the prediction of safety as compared to the EB method. The MEB1 method performed better than the EB method in which the MEB1 method gave relatively more sites (between 0.4% to 25.3% more) with smaller individual square of error (SE) values. Moreover, the MEB1 method gave the smallest sum-squares of error (SSE) and root-mean square error ($RMSE$) between predicted and observed accident frequencies in 2004-2006 (for all types of junctions) than the original EB method and also than the other methods. Although the MEB with λ estimated from the 5-year period (MEB2) method also had generally more number of junctions with individual SE values lower than that of the EB method, it could not be entirely regarded as a better method than the EB method since its SSE and $RMSE$ values were higher than the EB method for more types of junction.

Finally, a procedure for identification and ranking of hazardous sites based on the MEB method was established. When the method was applied to the study junctions, 59 hazardous junctions were identified and then ranked according to three criteria, which were MEB safety estimate ($E(m|K)$), potential for safety improvement (PSI) and level of hazard (H). Evaluation on the three criteria revealed that ranking based on the $E(m|K)$ and PSI criteria were preferred.

9.2 Research contributions

The following are the main contributions of this research:

1. **The development of accident prediction models based on long-term and short-term accident data using junctions in Singapore as a development platform.**

The accident prediction models provide important tools to estimate safety and to evaluate effectiveness of accident countermeasures through before-and-after studies. Importantly, these accident prediction models provide

information about the safety level of specified populations for input to the EB/MEB safety estimation/prediction methods. The long-term accident prediction models were developed by incorporating a power trend factor, which is useful to ensure that the developed model is not rapidly outdated. The power function is selected because of its simplicity (in terms of its parameter being solved easily without requiring any transformation) and its accordance with the national accident trend. However there was an assumption made in determining the form of the long-term models that the effect of road safety programmes is not changing from year to year while in reality the effect of road safety programmes for certain year may be larger than for other years. As an alternative, the short-term models were developed using an assumption of constant accident rate during the study period. In general, the long-term models, when combined with the EB/MEB methods for safety estimation/prediction, outperformed the short-term models when they were used to predict safety/accidents in 2004, 2005 and 2006.

2. The modification of Empirical Bayes (EB) method in estimating/predicting safety in order to improve safety estimation/prediction.

The non-Poisson nature in the accident counts at some sites has led to the development of the modified EB method that makes use of the Generalised Poisson distribution. The Generalised Poisson distribution was able to fit historical accident data of a site better than the Poisson distribution. It was found that the use of the Generalised Poisson distribution in the modified EB method can give better estimation/prediction of safety. This modified EB method should thus provide a more generalized approach in safety estimation/prediction.

3. Formulating an approach to identify hazardous sites using the concept of probability in selecting the worst site (PSS) based on modified EB method.

In this method, a site is identified as hazardous if there is a significant probability that its accident potential - $E(m|K)$ exceeds the value that is

normal - $E(m)$ for similar sites in the reference population. Essentially, the method consists of two steps: step one entails identification of hazardous sites based on the probability of selecting a site with accident potential exceeding what is normal for sites similar to the site being investigated, and step two involves ranking of the hazardous sites identified in step one for priority treatment.

9.3 Recommendations for future research

The following are some recommendations for future research:

1. Traffic data for unsignalised junctions in Singapore

The research was unable to develop an accident-exposure model for unsignalised junctions in Singapore due to unavailability of traffic count data for unsignalised junctions. Therefore, the collection of traffic count data for unsignalised junctions should be prioritised in future research. Given that collecting the traffic data directly in the field may be overly resource-intensive, an alternative approach is to make use of origin-destination (OD) data for Singapore roads to estimate flows at the nodes (junctions).

2. Accident prediction model for one-way T-signalised junctions in Singapore

The developed model cannot cater to T-signalised junctions with prohibition of traffic from minor leg to major legs (one-way minor road, entry only). It is suggested to collect more numbers on this type of junctions so an accident prediction model for this type of junctions can be developed.

3. Non-Poisson junctions

A few junctions were found to have accident counts being non-Poisson in nature. Though the method developed in this study can cater to non-Poissonity behaviour, it would be interesting to do detailed investigation into various features/characteristics of the junction to understand the reasons for the observed counts being non-Poisson in their distribution.

4. Disaggregated accident prediction model

The research involved development of accident prediction models using only total fatal and injury accidents. It is desirable to disaggregate the accident data according to the nature of accidents (single vehicle, between moving vehicles, etc) and also according to the pre-crash movements of vehicles (for crashes between moving vehicles). However, this research re-affirmed the difficulty in trying to develop disaggregate models as the observed accident numbers by collision types were too low to be able to develop reliable models. Also, the recording of pre-crash vehicle movements has only started recently (since 2003, which is the end of the modelled period).

5. Develop accident prediction model at network level

This research only managed to develop accident prediction model for Singapore road junctions. However, it is very important to have models for the whole road network in Singapore, which will consist of models for road junctions, road links and expressways. The availability of such models will be useful for the authority in deciding the best strategy to be taken in road safety improvement programmes.

6. Examine the transferability issue of the accident prediction models

Basically, the research developed accident prediction models based on the western part of Singapore. It is desirable to check whether the model is applicable to the remaining parts of Singapore, and also to other cities in the world that have similar characteristics as Singapore. This would entail establishing procedures to test whether a certain accident prediction model that is developed using one data set at a certain location can produce reliable predictions for other environment.

7. Develop a reliable method that can be used to check the goodness-of-fit to GP distribution for small sample size

It seems that there is no available test that can be used to test the goodness-of-fit of GP distribution for small sample size. When the sample size is

small, the standard χ^2 test and the score statistic test cannot be used. The KS test by Conover (1972, 1999) is also not appropriate since the GP parameters must not be estimated from the samples. A possible test that can be used is the CR test, but the applicability of the test is very limited such that 1001 samples out of 1186 samples in this study could not be tested. It is therefore very crucial to continue research into developing a test method that can be used to check the goodness-of-fit to GP distribution that can be applied for small sample size as is often the case when dealing with accident count data.

References

Abbas, K.A. (2004), "Traffic safety assessment and development of predictive models for accidents on rural roads in Egypt", Accident Analysis & Prevention, Vol.36, No.2, pp. 149-163.

Abdel-Aty, M. and Wang, X. (2006), "Crash estimation at signalized intersections along corridors: Analyzing spatial effect and identifying significant factors", Paper presented at the TRB Annual Meeting 2006, USA.

Abdel-Aty, M., Keller, J., and Brady, P. (2005), "Analysis of the types of crashes at signalized intersections using complete crash data and tree-based regression", Transportation Research Record, Vol.1908, pp. 37-45.

Abdel-Aty, M. and Keller, J. (2005), "Exploring the overall and specific crash severity levels at signalized intersections", Accident Analysis & Prevention, Vol.37, No.3, pp. 417-425.

Abdel-Aty, M.A. and Radwan, A.E. (2000), "Modeling traffic accident occurrence and involvement", Accident Analysis & Prevention, Vol.32, No.5, pp. 633-642.

Belanger, C. (1994), "Estimation of safety of four-legged unsignalised intersections", Transportation Research Record, Vol.1467, pp. 23-29.

Berhanu, G. (2004), "Models relating traffic safety with road environment and traffic flows on arterial roads in Addis Ababa", Accident Analysis & Prevention, Vol.36, No.5, pp. 697-704.

Bonneson, J.A. and McCoy, P.T. (1993), "Estimation of safety at two-way stop-controlled intersections on rural highways", Transportation Research Record, Vol.1401, pp. 23-29.

Brijs, T., Karlis, D., Van den Bossche, F., and Wets, G. (2003), "A Bayesian model for ranking hazardous sites", Flemish Research Center for Traffic Safety, Diepenbeek, Belgium.

Brude, U. and Larsson, J. (1993), "Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit?" Accident Analysis & Prevention, Vol.25, No.5, pp. 499-509.

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

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

Chin, H.C. and Quddus, M.A. (2003), "Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections", Accident Analysis & Prevention, Vol.35, No.2, pp. 253-259.

Cochran (1954), "Some methods for strengthening the common chisquare tests", Biometrics, Vol.10, No.4, pp. 417-451.

Colgate, M.G. and Tanner, J.C. (1967), "Accidents at rural three-way junctions", RRL Report, Road Research Laboratory, Crowthorne.

Conover, W.J. (1972), "A Kolmogorov goodness of fit test for discontinuous distributions", Journal of the American Statistical Association, Vol.67, No.339, pp. 591-596.

Conover, W.J. (1999), "Practical nonparametric statistics", 3rd ed., John Wiley & Sons, Inc., New York.

Consul, P.C. (1989), "Generalized Poisson distributions: properties and applications", Marcel Dekker.

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

Del Mistro, R.F. (1981), "Accidents at urban intersections - A second study", Technical Report, National Institute for Transport and Road Research, CSIR, South Africa.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994), "Analysis of longitudinal data", Oxford University Press, Oxford.

Doherty, S.T., Andrey, J.C., and MacGregor, C. (1998), "The situational risks of young drivers: The influence of passengers, time of day and day of week on accident rates", Accident Analysis & Prevention, Vol.30, No.1, pp. 45-52.

Dutt, A.K., Reinfurt, D.W., and Stutts, J.C. (1977), "Accident involvement and crash injury rates: An investigation by make, model and year of car", Accident Analysis & Prevention, Vol.9, No.4, pp. 275-283.

El-Sadig, M., Nelson Norman, J., Lloyd, O.L., Romilly, P., and Bener, A. (2002), "Road traffic accidents in the United Arab Emirates: trends of morbidity and mortality during 1977-1998", Accident Analysis & Prevention, Vol.34, No.4, pp. 465-476.

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

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

Evans, A.W. (2003), "Estimating transport fatality risk from past accident data", Accident Analysis & Prevention, Vol.35, No.4, pp. 459-472.

Famoye, F., Wulu, J.T.J., and Singh, K.P. (2004), "On the generalized Poisson regression model with an application to accident data", Journal of Data Science, Vol.2, pp. 287-295.

Golias, J.C. (1997), "Effects of signalisation on four-arm urban junction safety", Accident Analysis & Prevention, Vol.29, No.2, pp. 181-190.

Greibe, P. (2003), "Accident prediction models for urban roads", Accident Analysis & Prevention, Vol.35, No.2, pp. 273-285.

Grime, G. (1987), "Handbook of road safety research", Butterworth & Co.

Hammerslag, R., Roos, J.P., and Kwakernaak, M. (1982), "Analysis of accidents in traffic situations by means of Multiproportional Weighted Poisson Model", Transportation Research Record, Vol.847, pp. 29-36.

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

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

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

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

Hauer, E., Ng, J.C.N., and Lovell, J. (1989), "Estimation of safety at signalised intersections", Transportation Research Record, Vol.1185, pp. 48-61.

Hauer, E. and Persaud, B.N. (1984), "Problem of identifying hazardous locations using accident data", Transportation Research Record, Vol.975, pp. 113-140.

Hauer, E., Terry, D., and Griffith, M.S. (1994), "Effect of resurfacing on safety of two-lane rural roads in New York state", Transportation Research Record, Vol.1467, pp. 30-37.

Hiselius, L.W. (2004), "Estimating the relationship between accident frequency and homogeneous and inhomogeneous traffic flows", Accident Analysis & Prevention, Vol.36, No.6, pp. 985-992.

Hong Kong Transport Department (2003), "Road traffic accident statistics 2003", Retrieved 12 March-2006 from http://www.td.gov.hk/FileManager/EN/Content_184/02fig1.32e.pdf

Hong Kong Transport Department (2004), "Road traffic accident statistics 2004", Retrieved 12 March-2006 from http://www.td.gov.hk/FileManager/EN/Content_184/04fig1.32e.pdf

Horn, S.D. (1977), "Goodness-of-fit tests for discrete data: A review and application to a health impairment scale", Biometrics, Vol.33, No.1, pp. 237-247.

Jadaan, K.S. and Nicholson, A.J. (1992), "Relationships between road accidents and traffic flows in an urban network", Traffic Engineering and Control, Vol.33, No.9, pp. 507-511.

Joksch, H.C. (1971), "The relation between injury frequency and severity and vehicle age in pre-1967 automobiles", Accident Analysis & Prevention, Vol.3, No.4, pp. 281-289.

Joksch, H.C. and Wuerdemann, H. (1973), "Estimating the effects of crash phase injury countermeasures: The fatality trend and its modification by countermeasures", Accident Analysis & Prevention, Vol.5, No.1, pp. 1-26.

Jovanis, P.P. and Chang, H. (1986), "Modelling the relationship of accidents to miles traveled", Transportation Research Record, Vol.1068, pp. 42-51.

Karlaftis, M.G. and Golias, I. (2002), "Effects of road geometry and traffic volumes on rural roadway accident rates", Accident Analysis & Prevention, Vol.34, No.3, pp. 357-365.

Kumara, S.S.P. and Chin, H.C. (2003), "Modelling accident occurrence at signalized tee intersections with special emphasis on excess zeros", Traffic Injury Prevention, Vol.4, pp. 53-57.

Kutner, M.H., Nachtsheim C.J., and Neter, J. (2004), "Applied linear regression models", 4th ed., McGraw-Hill.

Lassarre, S. (1986), "The introduction of the variables "traffic volume," "speed" and "belt-wearing" into a predictive model of the severity of accidents", Accident Analysis & Prevention, Vol.18, No.2, pp. 129-134.

Lau, M.Y.K. and May, A.D. (1988), "Injury accident prediction models for signalized intersections", Transportation Research Record, Vol.1172, pp. 58-67.

Leong, H.J.W. (1973), "Relationship between accidents and traffic volumes at urban intersections", Australian Road Research, Vol.5, No.3, pp. 72-90.

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

Lord, D. and Persaud, B.N. (2000), "Accident prediction models with and without trend: Application of the generalized estimating equations procedure", Transportation Research Record, Vol.1717, pp. 102-108.

LTA (2006), "Annual vehicle statistics 2006", Retrieved 5 November 2007 from [http://www.lta.gov.sg/corp_info/doc/MVP01-1%20\(MVP%20by%20type\)%20.pdf?](http://www.lta.gov.sg/corp_info/doc/MVP01-1%20(MVP%20by%20type)%20.pdf?)

MacNab, Y.C. (2003), "A Bayesian hierarchical model for accident and injury surveillance", Accident Analysis & Prevention, Vol.35, No.1, pp. 91-102.

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

Maher, M.J. and Summersgill, I. (1996), "A comprehensive methodology for the fitting of predictive accident models", Accident Analysis & Prevention, Vol.28, No.3, pp. 281-296.

Massie, D.L., Campbell, K.L., and Williams, A.F. (1995), "Traffic accident involvement rates by driver age and gender", Accident Analysis & Prevention, Vol.27, No.1, pp. 73-87.

Maycock, G. and Hall, R.D. (1984), "Accidents at 4-Arm roundabout", TRRL Laboratory Report, Transport and Road Research Laboratory, UK.

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

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

Miaou, S.P. (1994), "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions", Accident Analysis & Prevention, Vol.26, No.4, pp. 471-482.

Miaou, S.P. and Lum, H. (1993), "Modeling vehicle accidents and highway geometric design relationships", Accident Analysis & Prevention, Vol.25, No.6, pp. 689-709.

Mighty Minds (2004), "Singapore Street Directory", Mighty Minds, Singapore.

Mighty Minds (2006), "Singapore Map", Retrieved 1 May-2006 from <http://www.mightyminds.com.sg/mightymindsmap.html>

Mountain, L., Fawaz, B., and Jarrett, D. (1996), "Accident prediction models for roads with minor junctions", Accident Analysis & Prevention, Vol.28, No.6, pp. 695-707.

Mountain, L., Maher, M., and Fawaz, B. (1998), "The influence of trend on estimates of accidents at junctions", Accident Analysis & Prevention, Vol.30, No.5, pp. 641-649.

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

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

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

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

Nicholson, A.J. (1986), "The randomness of accident counts", Accident Analysis & Prevention, Vol.18, No.3, pp. 193-198.

Nicholson, A.J. (1987), "A generalised Poisson distribution and its applicability to accident count analysis", Unpublished article.

Nicholson, A.J. (1990), "Measures of accident clustering", Proceedings of 11th International Symposium on Transportation and Traffic Theory, Yokohama, Japan, pp. 133-151.

Numerical-Algorithms-Group (1994), "The GLIM system: release 4 manual", The Royal Statistical Society, Oxford, UK.

Ogden, K.W., Newstead, S.V., Ryan, P.K., and Gantzer, S. (1994), "Factors affecting crashes at signalised intersection", Report, Accident Research Centre, Monash University, Melbourne, Australia.

Okamoto, H. and Koshi, M. (1989), "A method to cope with the random errors of observed accident rates in regression analysis", Accident Analysis & Prevention, Vol.21, No.4, pp. 317-332.

Persaud, B. and Kazakov, A. (1994), "A procedure for allocating a safety improvement budget among treatment types", Accident Analysis & Prevention, Vol.26, No.1, pp. 121-126.

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

Persaud, B.N. (1991), "Estimating accident potential of Ontario road sections", Transportation Research Record, Vol.1327, pp. 47-53.

Pickering, D., Hall, R.D., and Grimmer, R. (1986), "Accidents at rural T-junctions", TRRL Research Report, Transport and Road Research Laboratory, U.K.

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

Polus, A., Hocherman, I., and Efrat, E. (1988), "Evaluation of the crash rates of male and female drivers", Transportation Research Record, Vol.1172, pp. 42-46.

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

Rayner, J.C.W. and McIntyre, R.I. (1985), "Use of the score statistic for testing goodness-of-fit of some generalised distributions", Biometrical Journal, Vol.1985, No.2, pp. 159-166.

Reinfurt, D.W., Stewart, J.R., and Weaver, N.L. (1991), "The economy as a factor in motor vehicle fatalities, suicides, and homicides", Accident Analysis & Prevention, Vol.23, No.5, pp. 453-462.

Sabey, B.E. and Staughton, G.C. (1975), "Interacting roles of road environment, vehicle and road user in accidents", Paper presented at the 5th International Conference on the International Association for Accident and Traffic Medicine, London.

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

Schluter, P.J. (1992), "A hierarchical Bayesian interpretation of motor vehicle accidents", MSc Thesis, Department of Statistics, University of Canterbury, New Zealand.

Schluter, P.J., Deely, J.J., and Nicholson, A.J. (1997), "Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model", The Statistician, Vol.46, No.3, pp. 293-316.

Scott, P.P. (1986), "Modelling time-series of British road accident data", Accident Analysis & Prevention, Vol.18, No.2, pp. 109-117.

Scuffham, P.A. and Langley, J.D. (2002), "A model of traffic crashes in New Zealand", Accident Analysis & Prevention, Vol.34, No.5, pp. 673-687.

Shankar, V., Mannering, F., and Barfield, W. (1995), "Effect of roadway geometrics and environmental factors on rural freeway accident frequencies", Accident Analysis & Prevention, Vol.27, No.3, pp. 371-389.

Shankar, V., Milton, J., and Mannering, F. (1997), "Modeling accident frequencies as zero-altered probability processes: An empirical inquiry", Accident Analysis & Prevention, Vol.29, No.6, pp. 829-837.

Shankar, V.N., Albin, R.B., and Milton, J.C. (1998), "Evaluating median crossover likelihoods with clustered accident counts: An empirical inquiry using the random effects Negative Binomial model", Transportation Research Record, Vol.1635, pp. 44-48.

Stamatiadis, N. and Deacon, J.A. (1995), "Trends in highway safety: Effects of an aging population on accident propensity", Accident Analysis & Prevention, Vol.27, No.4, pp. 443-459.

Stewart, J.R. (1996), "Applications of classification and regression tree methods in roadway safety studies", Transportation Research Record, Vol.1542, pp. 1-5.

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

Storie, V.J. (1977), "Male and female car drivers: Differences observed in car crashes", TRRL Laboratory Report, Transport and Road Research Laboratory, U.K.

Tanner, J.C. (1953), "Accidents at rural three way junctions", Journal of Institute of Highway Engineers, Vol.2, No.11, pp. 56-67.

Taylor, M.C. (1995), "Factors affecting accidents at urban signalised junctions", Proceedings of The 23rd European Transport Forum, University of Warwick, England, pp. 271-284.

Thomas, I. (1996), "Spatial data aggregation: Exploratory analysis of road accidents", Accident Analysis & Prevention, Vol.28, No.2, pp. 251-264.

TPD (2003), "Annual statistical report on road traffic accidents in Singapore 2003", Traffic Police Department, Singapore.

TPD (2004), "Annual statistical report on road traffic accidents in Singapore 2004", Traffic Police Department, Singapore.

Treat, J.R., Tumbas, N.S., McDonald, S.T., Shinar, D., Hume, R.D., Mayer, R.E., Stanisfer, R.L., and Castillan, N.J. (1977), "Tri-level study of the causes of traffic accidents", Report No. DOT-HS-034-3-535-77, Indiana University, USA.

Tunaru, R. (2002), "Hierarchical Bayesian models for multiple count data", Austrian Journal of Statistics, Vol.31, pp. 221-229.

Turner, S. and Nicholson, A. (1998), "Intersection accident estimation: the role of intersection location and non-collision flows", Accident Analysis & Prevention, Vol.30, No.4, pp. 505-517.

Wagenaar, A.C. (1984), "Effects of macroeconomic conditions on the incidence of motor vehicle accidents", Accident Analysis & Prevention, Vol.16, No.3, pp. 191-205.

Wang, X., Abdel-Aty, M., and Brady, P. (2006), "Crash estimation at signalized intersections: Significant factors and temporal effects", Paper presented at the TRB Annual Meeting 2006, USA.

Wang, W. and Famoye, F. (1997), "Modeling household fertility decisions with generalized Poisson regression", Journal of Population Economics, Vol.10, pp. 273-283.

WHO (2004), "World report on road traffic injury prevention", World Health Organization.

Williams, A.F. (1985), "Nighttime driving and fatal crash involvement of teenagers", Accident Analysis & Prevention, Vol.17, No.1, pp. 1-5.

Wong, Y.D., Hau, L.P., and Fan, H.S.L. (2000), "A study of Singapore road traffic accident characteristics", Unpublished Report, Centre for Transportation Studies, Nanyang Technological University, Singapore.

Wong, Y.D. and Nicholson, A. (1992), "Driver behaviour at horizontal curves: risk compensation and the margin of safety", Accident Analysis & Prevention, Vol.24, No.4, pp. 425-436.