1-1-2013

Integrating Information From Prior Research Into A Before-After Road Safety Evaluation Through Bayesian Approach And Data Sampling

Chen Yongsheng
Ryerson University

INTEGRATING INFORMATION FROM PRIOR RESEARCH INTO A BEFORE-AFTER ROAD SAFETY EVALUATION THROUGH BAYESIAN APPROACH AND DATA SAMPLING

By
Yongsheng Chen
Doctor of Philosophy, Beijing University of Technology, Beijing, China 2001
Master of Science, Beijing University of Technology, Beijing, China 1996
Bachelor of Science, Beijing University of Technology, Beijing, China 1993

A dissertation
presented to Ryerson University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Program of
Civil Engineering

Toronto, Ontario, Canada

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Integrating Information from Prior Research into a Before-after Road Safety Evaluation through Bayesian Approach and Data Sampling

Doctor of Philosophy in Civil Engineering, 2013

By
Yongsheng Chen
Ryerson University

ABSTRACT

Before-after road safety evaluation (B/A) to measure safety treatment effect is a key mission in road safety management, and has fueled considerable research. However, previous research in this area has been overwhelmingly dedicated to safety model estimation with less emphasis on other methodological issues. As a result, there continues to be uncertainty in the validity of treatment effect estimates. This study seeks, with innovative paradigms, a systematic solution by solidifying methodologies for every essential step of a thorough B/A process to secure its ultimate validity.

Methodologies of data sampling and processing, and before and after model development, both vital procedures that have been historically neglected, are investigated. A pre-test data sampling approach to select reference groups is established in the context of B/A application. A post-assignment propensity score matching method is developed in order to further eliminate statistical bias while the treatment effect indicator – collision reduction ratio (CRR) – is being estimated.

Rather than focus on single safety model development as is common in traffic safety research, this study seeks all viable knowledge by employing various safety measures including collision and safety surrogates, by embedding several adaptable random distributions, by fitting models through both "Frequentist" and "Bayesian" approaches, and by exploring a variety of model forms and components. Accordingly, the output of this study is not a "best" single model, but rather an amalgamation of diversified models. The diversity is shown to be attractive in terms of information conveyed, especially for the B/A process.

Finally, this study succeeds in finding a methodology to integrate all of the diverse knowledge sources. The Bayesian Model Averaging (BMA) method is investigated and developed to integrate a variety of statistical significant models without exclusion, in forging a unified model.

All methodologies explored and developed in this study are essential to secure the validity of the B/A process. As important, they are substantially connected to each other. Should one method be deficient, the remaining steps cannot guarantee validity of B/A process. As a whole, these methodologies, if properly developed and applied, constitute a logical chain to estimate treatment effect with minimal errors and high validity.

ACKNOWLEDGMENTS

Working on a Ph.D. program, even my second one, is still overwhelming but thankfully it is absolutely worthy and I feel gratitude for many people who diligently shored me up to accomplish this dissertation research.

First of all, I am heartily grateful to my supervisor, Dr. Bhagwant Persaud. His extraordinary vision, insightful guidance and high scientific standards safeguarded success of this dissertation. He is steady influence behind each step I ever made in this research. Particularly, I appreciate his consistent trust, kindness and patience. I'm looking forward to having his continued support in my future career development and life.

Data are the cornerstone for all studies conducted for this dissertation. It is with this recognition that I express sincere gratitude to the Traffic Management Centre, Transportation Services, City of Toronto, and the Office of Traffic Safety, City of Edmonton for providing key Canadian data sets. My gratitude also extends to the providers of the Italian roundabout data, including Professor Raffaele Mauro of the University of Trento, Local Police of the City of Cuneo, and Local Police of the City of Novara of Italia.

I am also indebted to Mr. Craig Lyon for his guidance on statistical methods, tools and traffic safety applications, and to Dr. Emanuele Sacchi and Dr. Marco Bassani of Politecnico di Torino for facilitating the Italian roundabout data and for their collaboration in the research on safety surrogates that has already yielded a peer-reviewed publication in a journal with a high impact rating.

My gratitude also goes to Dr. Stevanus Tjandra, of the Edmonton Office of Traffic Safety, for his consideration and support at the last but crucial period of my dissertation research, and to Dr. Xiaoduan Sun, Professor of Civil Engineering, University of Louisiana, for her

strong reference that helped me in joining Dr. Persaud's research team, and for her long time trust and support.

Finally, I would thank my family. Their contributions are not visually present on any line of this dissertation but are fundamentally everywhere. Their support was enormous and pivotal for me to get this job done.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AADT	Annual Average Daily Traffic
B/A	Before-after Road Safety Evaluation
BMA	Bayesian Model Averaging
CMF	Collision Modification Factor
CM-Function	Collision Modification Function
CRR	Collision Reduction Rate
CURE	Cumulative Residuals
EB	Empirical Bayesian
FB	Full Bayesian
FI	Fatal and Injury (Collisions)
GEE	Generalized Estimating Equation
GLM	Generalized Linear Modeling
GOF	Goodness-of-fit
HSM	Highway Safety Manual

k	Dispersion Parameter, a parameter describing the relationship between mean and variance
MAD	Mean Absolute Deviation
MCMC	Markov Chain Monte Carlo
MPB	Mean Prediction Bias
MSE	Mean Squared Error
MSPE	Mean squared Prediction error
MTO	Ministry of Transportation, Ontario
n	Number of years of collision used for the study
NB	Negative Binomial
PDO	Property Damage Only (collisions)
PPS	Probability Proportional to Size
RA	Random Assignment
RG	Reference Group
RP	Reference Population
RS	Random Sampling

SAS	Statistical Analysis Software
SPF	Safety Performance Function
SS	Statistical Significant
TG	Treated Group
ZINB	Zero Inflated Negative Binomial
ZIP	Zero Inflated Poisson

CHAPTER 1 INTRODUCTION

1.1 OVERVIEW OF BEFORE-AFTER ROAD SAFETY EVALUATIONS

The implementation of before-after road safety evaluation (before-after evaluation or B/A) to measure the effects of safety remedies is one of the two key missions in road safety analysis. Another is network screening, which identifies sites with potential for safety treatment (Persaud et al., 2010a). As one of two pivots in this domain, B/A has fueled considerable research work, with a large body of published literature. The empirical Bayes (EB) method (Hauer, 1985; Persaud et al., 2010a), for instance, is one of the most well established approaches in before-after evaluations to date. Recently, the full Bayesian (FB) approach has also generated many efforts as a viable option for the conduct of before-after evaluations (Persaud et al., 2010a; Lan et al., 2009; Lan, 2010; Yanmaz-Tuzel and Ozbay, 2010; El-Basyouny and Sayed, 2010).

Regardless of the "maturity" of B/A mechanisms, some essential pieces are still missing. This dissertation accordingly aims to improve the methodology of before-after evaluations by filling such research gaps.

1.1.1 How Before-after Evaluations Work

A B/A process utilizes four datasets: treated group, reference population, reference group and comparison group. The treated group is the group receiving a certain treatment. The reference population, in traffic safety practice, is the total collection of intersections or roadway segments of a jurisdiction with same features as the treated group before treatment. For example, for the treatment of protected left turn provision, the reference population is the total signalized intersections in a city without protected left turn, or, for the treatment of median barrier, the reference population is the entire highway network in a state/province without median barrier. Reference groups are the legitimate samples selected from reference population. The comparison group is a subset of the same type sites as treated groups and is usually used to compare observed

before and after period collisions between the comparison and treated groups to enhance treatment effect estimation. Figure 1-1 illustrates the relationship among these different datasets.

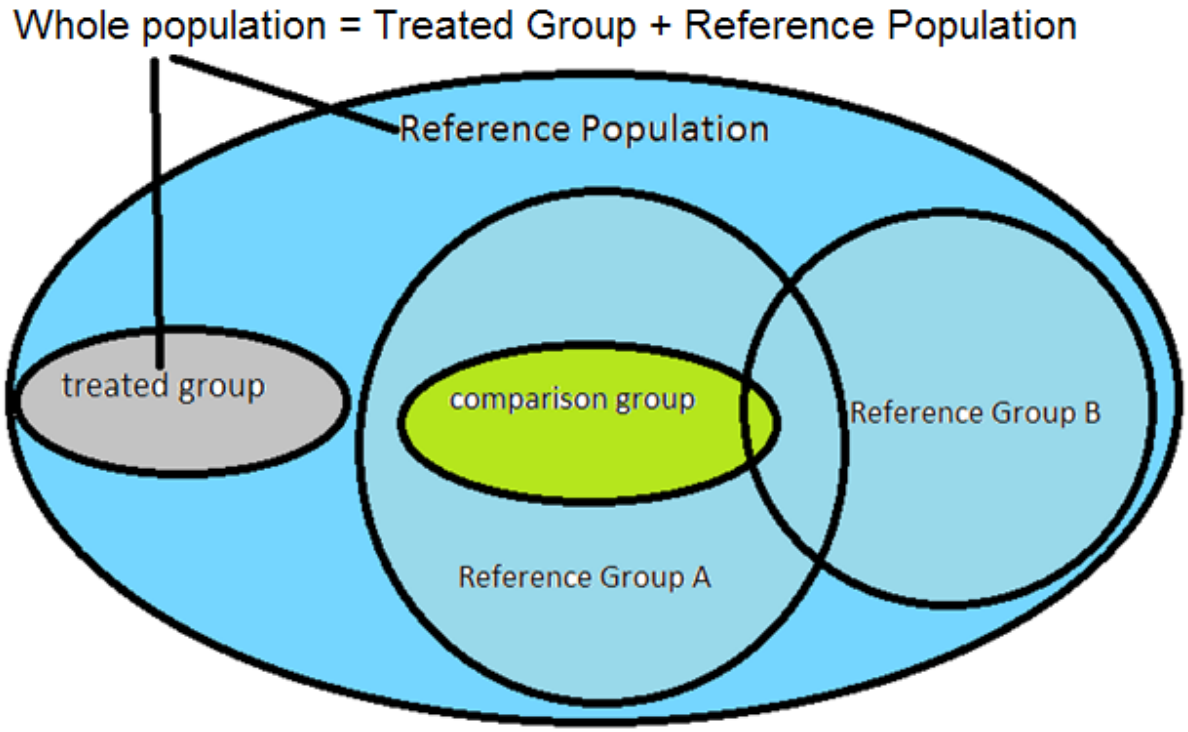


Figure 1-1 Different Data Groups Involved in Before-after Evaluation for Protected Left Turn Control

The working mechanism of B/A utilizes the above four datasets in different ways. Among them, the treated group (TG) is the core of B/A processing, since the final result of B/A, namely the treatment effect, is estimated by comparison for the treated group itself between observed collisions after treatment and the "postulated" collisions without treatment (Hauer, 1985; Persaud et al., 2012a; Persaud et al., 2012b).

Reference population and reference groups are utilized to develop safety models, which are introduced to offer referential information for before-after evaluations. Reference Group A and B at Figure 1-1 represent different samples obtained from one reference population. Both the EB

and FB methods, the best- established approaches, require a reference population (RP) or reference group (RG).

The comparison group is not necessarily present for all B/A approaches. It is only for comparison group (C/G) method (Hauer, 1985) that a comparison group is a necessity. For the currently applied EB and FB methods, the conventional practice does not require the physical presence of a comparison group; instead, postulated collisions from safety models replace the role of observed collisions in the comparison group in the B/A process.

Furthermore, B/A working procedures are different, depending on the different methods. More details will be provided in the following chapters, but in short, the EB method is a 2-step process that combines prior and current information to derive an estimate for the expected safety of a site that is being evaluated. In contrast, a Bayesian approach such as FB has a one-step process to do the same job as the EB method (Lan et al., 2009).

All B/A schemes work in an identical manner to seek the collision reduction rate (CRR) that compares collisions that are observed in the after-treatment period for a treated group to postulated collisions without treatment. The latter is statistically estimated via extrapolation from the before-treatment period (before) collisions of a reference group, as shown in Figure 1-2.

1.1.2 Prior Knowledge for Before-after Evaluations

It can be noted from Figure 1-2 that the B/A methodology is rooted in "prior knowledge". Traditionally, the prior knowledge used to assist B/A processes is developed by a safety predictive model, also called as a safety performance function (SPF). However, this is not the only model used. A recent trend can be observed, in which there is a shift towards indirect safety indicators; that is, safety surrogates, such as conflicts, gap-acceptance distribution, speed, speed differentials, and traffic violations (Gettman and Head, 2001).

Although this dissertation is still primarily focused on safety models, safety surrogates will also be investigated as alternative prior knowledge. Theoretically, any type of the above-mentioned

safety surrogates can be used in B/A processes. However, this dissertation will use only speed as an example.

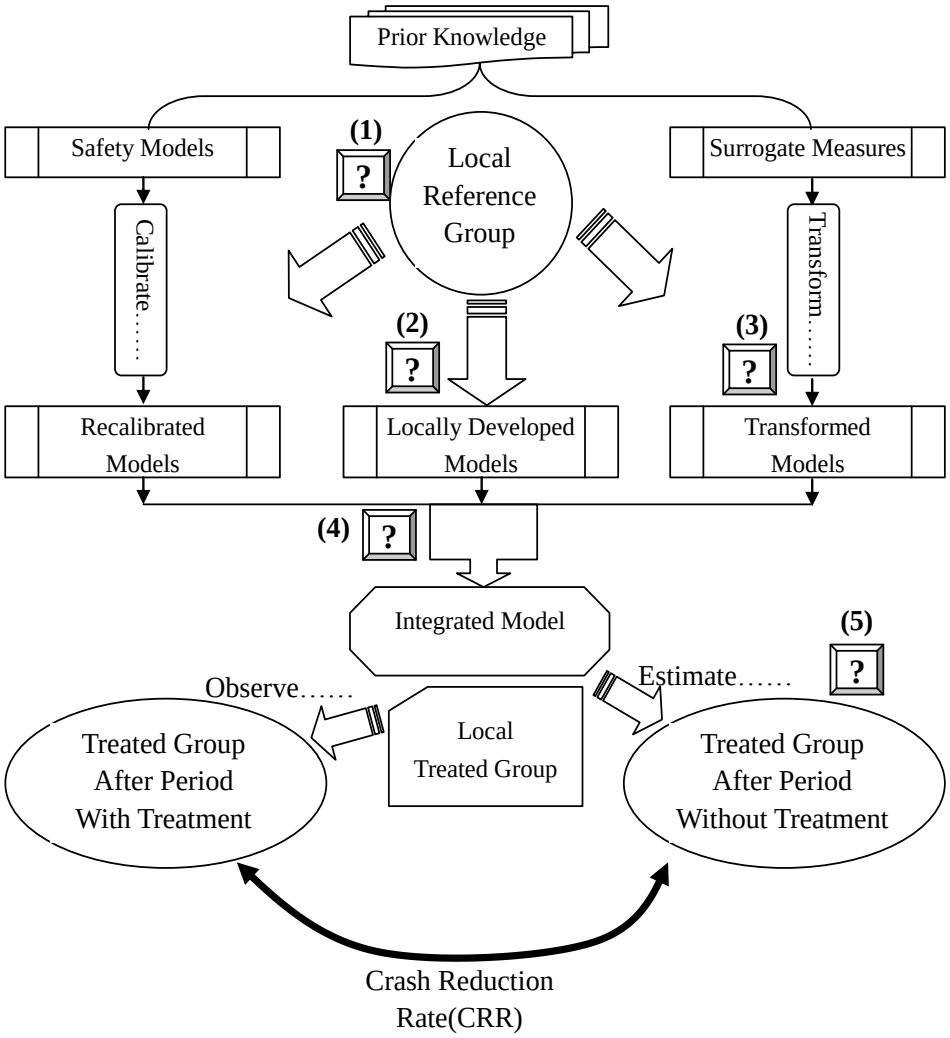


Figure 1-2 Identical Working Flowchart of Before-after Evaluations and Research Gaps

1.2 MISSING PIECES OF CURRENT BEFORE-AFTER EVALUATION SCHEMES

Notwithstanding the lengthy investigations and applications, there are still some gaps in B/A methodology. As highlighted in Figure 1-2, the boxes with "?" marks suggest the five most

pressing issues for researchers and practitioners, which are all bound by one unanswered question: how does one determine the B/A that gives the most robust results? For a safety model, there is "goodness-of-fit" (GOF) to measure the validity of a model (Huaer and Bamfo, 1997; Persaud et al., 2012; Persaud et al., 2011a; Persaud et al., 2011b; Persaud et al., 2010b; Ando and Tsayb, 2010; Ye et al., 2011). For B/A, it is reasonable to expect equivalent "goodness-of-compare" (GOC, vs. GOF) measures or procedures, but these are not systematically established as yet in current B/A schemes.

These five issues are conceptually discussed below, while a deeper investigation from a statistical perspective will be conducted in Section 1.3.

1.2.1 Selection of Local Reference Group

As mentioned, the basic idea of before-after evaluations is to calculate the treatment effect through comparison of observed collisions after treatment with postulated after period collisions without treatment. The latter, postulated collision frequency given that there is no treatment, is estimated from an reference group. One way to use reference groups is actually indirect rather than direct. It employs available SPFs developed from data of other (for some practitioners, might be even unknown) jurisdictions. That is to say, such before-after evaluations rely on a "remote and/or unknown" reference group. This pinpoints the first methodological issue of a B/A mechanism: how are genuine local reference groups selected?

This is the first issue that this dissertation addresses.

1.2.2 Identification of the Structure of Locally Developed Models

The question of interest is: What constitutes predictors and their function forms for a reference group? These are developed from safety models, which define the model structure and reflect the nature of the data. Different local jurisdictions have different data, which provide different model structures. However, given that there is no consistency, a model may have transferability issues.

Consequently, the second issue that this dissertation will address is to determine how to develop a structure for local models in order to resolve this dilemma.

1.2.3 Application of Safety Surrogates into Before-after Evaluation

In the B/A process described in Figure 1-2, the final objective in applying safety surrogates is to merge their information with those from safety models. Unfortunately, there is no current solution that is capable of quantitatively utilizing safety surrogates in collision measures, especially for application to B/A processes.

This void thus becomes another issue for this research to address.

1.2.4 Integration of Different Models into Before-after Evaluation

The fourth methodological issue in Figure 1-2 is based on the principle that all prior knowledge sources are intended to become unified afterwards. The question is: does this unification process work as a filter, or as an integrator? That is, should this step select one model while neglecting others, or should it somehow merge these models together?

This constitutes a major issue for the entire research study.

1.2.5 Determining Whether Reference Group Matches Treated Group

The last step for the B/A is a comparison to obtain the CRR, as shown in Figure 1-2. The comparison takes place, for treated group, between after period collision observations and estimations of what would have happened without the treatment, derived from the reference group. Traditionally, there is no numerical evaluation of whether a reference group "matches" the treated group in terms of this type of comparison, or, if not, what steps should be carried out to do so.

This is the last issue addressed in the dissertation.

1.3 MOTIVATION AND RESEARCH OBJECTIVES

The general goal of this dissertation research is to achieve a higher GOC for B/A processes, or, as expressed in conventional statistical terms, to minimize the bias of before-after evaluations by securing the internal and external validity of before-after evaluations. The external validity stems from the appropriate selection of a reference group and a referential knowledge base while the internal validity is achieved by appropriate assignment of treated groups (Dattalo, 2010). Since the assignment of treated groups, or network screening, is not a topic of this dissertation, the internal validity is instead pursued through a post-assignment matching process on the reference group. In realizing this goal, the above-mentioned five issues as outlined in Figure 1-2 are to be addressed. To supplement the previous conceptual descriptions, this section will accordingly re-investigate these five issues from a statistical perspective, so as to focus on the dissertation research objectives.

1.3.1 Pre-Test Data Sampling on Select Local Reference Group

From a statistical perspective, the B/A process is a "test" (Dattalo, 2010). This means that the before period is the "pre-test" stage, so that selection of an reference group is statistically a "pre-test" data sampling procedure.

The local reference group has multiple roles. First, any model needs to be calibrated before being used in the local context, as recommended by the Highway Safety Manual (HSM) (AASHTO, 2010). Hence, local data collection will at least require a calibration database. However, one should always consider developing a local model whenever possible since this is pertinent to local traffic system characteristics. That is to say, local data can be used as the basis for local model development as well.

1.3.2 Identification of Multi-level Structure for Locally Developed Models

The current HSM framework for collision prediction is actually not directly derived from a fully specified SPF equation. Instead, it is built on a base SPF and several collision modification factors (CMFs) as follows:

$$N_{predicted} = N_{spf\,x} \times (CMF_{1x} \times CMF_{2x} \times ..... \times CMF_{yx}) \times C_x \quad (1 - 1)$$

where

$N_{predicted}$ = predicted average crash frequency for a specific year for site type x,

$N_{spf\,x}$ = predicted average crash frequency determined for the base conditions of the SPF developed for site type x,

CMF_{yx} = crash modification factors specific to an SPF for site type x, and

C_x = calibration factor that adjusts the SPF to local conditions for site type x.

The objective of a locally developed model is thus to seek an appropriate structure transferable to the fundamental components of Equation 1-1 while being applicable to the local context. To compromise on these two aspects, a multi-level model (also referred to as hierarchical model) will be applied in which the first-level structure is consistent while its sub-categorical components address the local context (Goldstein, 1999; Chin and Huang, 2008; Lee et al., 2008).

1.3.3 Utilizing Knowledge from Surrogates towards Collision Measures

The intuitive statistical solution for utilizing knowledge from surrogates towards collision measures is a regression model that associates the former with the latter, which will be explored in this dissertation. Moreover, an alternative solution in the event that no statistical model is available will be investigated as well. Generally, this is a rank-based algorithm that conveys the ranking of the surrogates to the ranking of collision measures and finally estimates the relevant collisions in accordance with their rankings.

1.3.4 Model Averaging to Integrate Different Knowledge Sources

As for the selection of multiple models, conventional practices tend to retain only an optimal model determined by the "GOF" and to abandon all other candidate models. However, this is fundamentally against the principle of a robust B/A process in that excluding other models means neglecting many knowledge sources, with the consequence that the subsequent B/A steps are restricted to a narrow reference base.

Furthermore, when diverse models are considered for selection, including calibrated and locally developed ones, there is no consistent GOF measure efficient enough to identify the best choice. Different measures may sometimes lead to different recommendations, and pros and cons usually overlap. The GOF tests in Section 3.1.3 will provide good examples in addressing this issue.

The conceptually superior solution for this issue is an integrator to merge all models together without mass exclusion. From a statistical perspective, this is the modeling averaging approach (Claeskens and Hjort, 2009). Generally, the approach estimates coefficients of a unified model through weighted average algorithms from equivalent coefficients of candidate models.

1.3.5 Post-assignment Propensity Score Matching to Enhance Validity

The above research components all strive for reference group and referential information, which pertains to external validity (Dattalo, 2010). An ideal B/A requires internal validity as well. The foundation of internal validity is that both treated groups and reference groups should be homogeneous in all aspects, except for the implemented treatment itself. By this means, the comparison result validly reflects only the outcome of the treatment, instead of other effects. This internal validity is originally achieved via random treated group assignment that ensures that the treated group does not have "innate" heterogeneity compared to the reference group.

However, as stated at the beginning, this dissertation focuses on before-after evaluations rather than on assignment of sites to the treated group, which has been done before and is impossible modify. So the internal validity of before-after evaluations can only be pursued by an alternative

"post-assignment" approach. This approach may not be always be necessary but it would be an important alternative solution when heterogeneity still exists after treated group assignment, or when statistically rigorous treated group assignments were not conducted due to practical limitations. For these cases this dissertation investigates a post-assignment statistical process called "propensity score matching" to enhance the validity of the reference group. This is basically a calibration procedure based on the propensity score to adjust for the impact due to the dissimilarity of the reference group and the treated group.

In summary, after further exploration of the research objectives from a statistical perspective, the entire structure of this dissertation can be re-established per Figure 1-3.

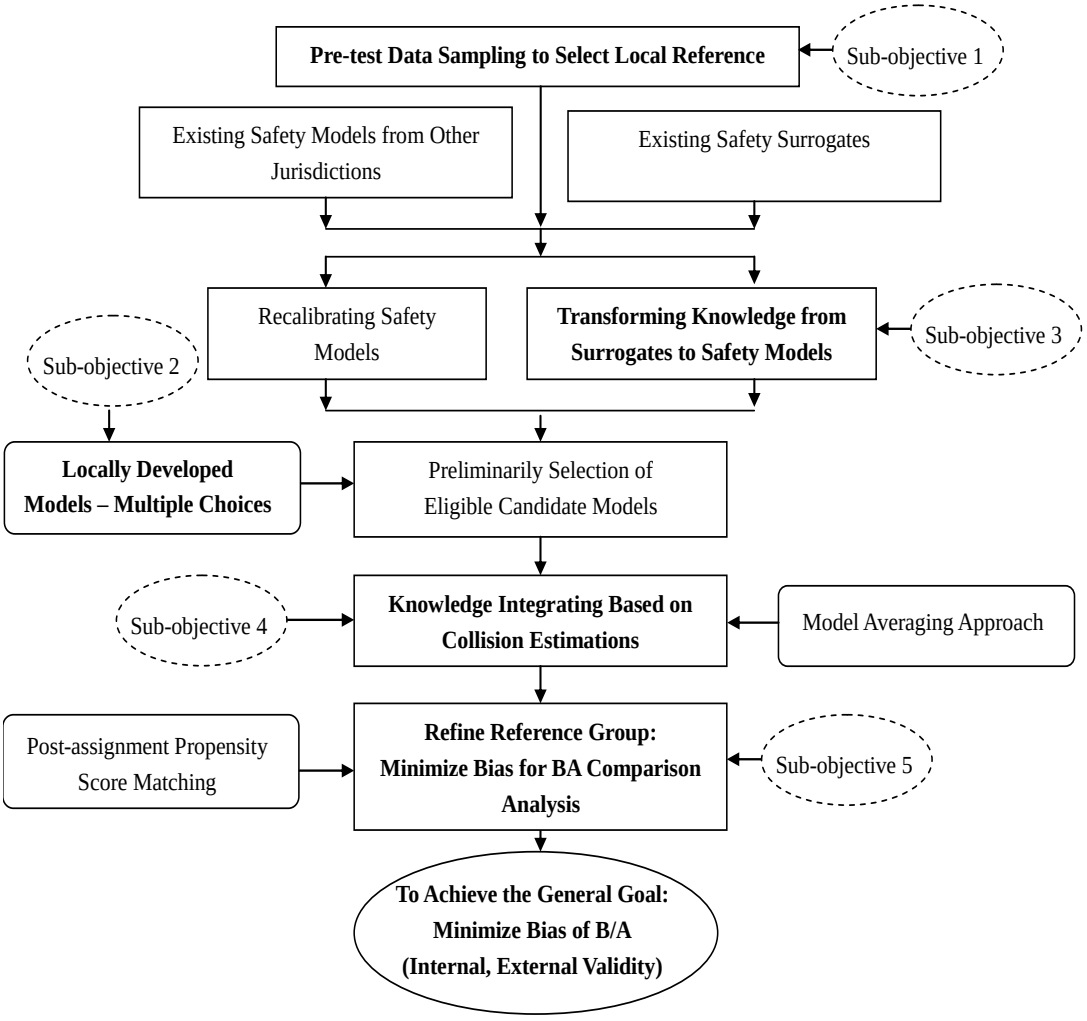


Figure 1-3 Research Flowchart of Dissertation – Objectives and Statistical Approaches

1.4 ORGANIZATION OF DISSERTATION

This dissertation is composed of 9 chapters. A literature review and sample data summaries are provided in the relevant chapters. All processes are developed by the use of the SAS statistical package (SAS Institute Inc., 2012).

Chapter 2 introduces Component 1: pre-test data sampling to select a local reference group. This is a two-stage procedure. The first stage is to determine an approximate sample size controlled by a given Type I error rate (probability of incorrectly identifying a statistically significant effect, denoted as "α"), and a given model power error (probability of not identifying a statistically significant effect when one exists, denoted as "B"; Accordingly, (1- B) is called as "model power level", usually simplified as "power") (Dattalo, 2010). The result of the first stage provides the basis for the second stage, which comprises data sampling to select a local reference group. In this dissertation, sequential stratified sampling is applied to achieve objective such as being representative of the entire population, coverage of all sub-categories of the population, and adequacy for significance of local safety model development.

Chapters 3 and 4 present Component 2: local safety performance function (SPF) development with a variety of different models. Chapter 3 develops SPFs via standard approach while Chapter 4 explores diversified SPF development, which is aimed at including a variety of local SPFs in order to capitalize on the diversity of knowledge sources. This part of research will explore and favor multi-level SPFs but retains other types of SPFs, including calibrated HSM models and single-level full local models, as optional choices. The preferable multilevel local models has first level average daily traffic (ADT)-only model with shape parameters and intercepts which are all functions of sub-hierarchical models with other covariates, including items with a local context.

Chapter 5 demonstrates Component 3: the utilization of knowledge from safety surrogates as a substitute for collision measures. There are two scenarios: with or without statistical models. The ideal scenario is the former when there is adequate data to support surrogate-based safety model estimation. This dissertation investigates the speed of modern roundabouts as the sample

surrogate and speed-based roundabout safety models are developed. Considering data inadequacy, an indirect approach is investigated for using the ranking of safety surrogates as a substitute for ranking of collisions, and for quantitatively estimating the collisions according to these rankings.

Chapter 6 addresses Component 4: a model averaging algorithm to integrate all collision estimations from different sources and approaches. In condideration of the heterogeneous nature of candidate models or estimations, a Bayesian model averaging (BMA) algorithm is applied. This algorithm seeks a unified model in which the coefficients are respectively estimated from equivalent coefficients of all candidate models based on a weighted averaging mechanism for which loglikehoods are employed as weights. The multi-level model structure introduced in Chapter 3, facilitates the averaging of calibrated and locally developed models.

Finally, all knowledge sources are merged together and a unique integrative model is formed.

Chapter 7 presents Component 5: the post-assignment statistics to refine the efficiency of reference groups in B/A processes. It is retrospective to the procedure in Chapter 2 but moves forward, with a post-assignment statistical process on the reference groups to further lower the comparison bias of the B/A. In this dissertation, an algorithm known as propensity score matching is applied. The principle is to measure the heterogeneity of the reference group versus the treated group and then apply calibrations accordingly for referential estimation.

Chapter 8 finalizes the last step of B/A process - treatment effect estimation - by an application example and then provides some brief discussion, comparing the pros and cons of the dissertation methodologies investigated versus more conventional approaches, from both conceptual and statistical perspectives.

Lastly, Chapter 9 concludes this dissertation with a short summary of the accomplishments of the research, conclusions and some suggestions for future studies.

1.5 SAMPLE FACILITY, MEASURE AND DATA

1.5.1 Sample Facility and Treatment Measure

For a B/A process, the central feature is a safety treatment applied to a certain type of facility. Considering data availability and rationality, this research has selected urban 4-legged signalized (4SG) intersection as the sample facility, and introduced left turn protection of signalized intersection (also called exclusive left turn signal) as the sample treatment. Hence the B/A process in this dissertation study has the following key characteristics:

- Phasing before - permitted left turn control,
- Phasing after – protected, protected/permitted left turn control,
- TG – intersections originally with permitted left turn, then converted to protected, protected/permitted left turn control,
- RP - Reference population, all other 4-legged signalized intersections except for treated group in a city or region, and
- RG – a sample extracted from a reference population, applied as representative of the population.

1.5.2 Summary Statistics of Treated Group Data

Treatment group data for the city of Toronto in Ontario, Canada was selected for study in this dissertation. For treatment of "left turn protection", Toronto has a treated group of 61 intersections. Table 1-1 provides the summary statistics for this group of data.

Table 1-1 Summarized Statistics of Treated Group Data

Dataset (# of sites)	Variable	Phase	Minimum Value	Maximum Value	Mean (Standard Deviation)
4-leg, signalized at-grade intersections from Toronto, Ontario, Canada (61 treated sites)	Multi-vehicle total collisions	Before	0	247	79.9(66.6)
		After	0	255	74.4(56.7)
	Multi-vehicle injury collisions	Before	0	105	35.8(27.6)
		After	0	82	26.5(18.2)
	Years	Before	1	7	4.0(1.9)
		After	1	7	4.0(1.9)
	Major AADT	Before	14489	74990	35267(11719)
		After	11504	73697	35069(11941)
	Minor AADT	Before	1466	42723	18096(9729)
		After	1466	37491	18501(9915)
	No. of approaches with left-turn lanes	-	0-8; 1-4; 2-6;3-5;4-38		
	No. of approaches with right-turn lanes	-	0-25; 1-13; 2-13;3-5;4-5		
	Intersection class[a]	-	3-26; 5-19; 6-4; 8-8; 12-4		

Note: a. Toronto intersection classification based on the functional classes of crossed roads: 1-express/express, 2-major arterial/expressway, 3-major arterial/major arterial, 4-expressway/minor arterial, 5-major arterial/minor arterial, 6-minor arterial/minor arterial, 7-unknown, 8-major arterial/collector, 9-minor arterial/collector, 10-collector/collector, 11-express/local, 12-major arterial/local. 13-minor arterial/local, and 14-collector/local.

1.5.3 Summary Statistics of Reference Population Data

The reference population is the entire collection of 4SG intersections except for the 61 treated sites in Toronto. This group comprises 1629 sites. In addition, the entire collection of 4SG intersections in Edmonton, Alberta, Canada was also selected as a supplemental reference population per the requirements of the methodological aspects of the research. Table 1-2 provides the summary statistics of these data.

Chapter 2 will focus on the methodological research to extract the reference groups, and the samples which are supposed to represent the reference population. Summarized statistics for the samples will be accordingly given in Chapter 2.

A detailed introduction and summary statistics for the sample data used for the safety surrogate research will be included in Chapter 4.

Table 1-2 Summarized Statistics of Reference Population

Dataset (# of sites)	Variable	Minimum Value	Maximum Value	Mean (Standard Deviation)
4-leg, signalized at-grade intersections from Toronto, Ontario, Canada (1629)	Multi-vehicle total collisions	0	370	62.0 (57.5)
	Multi-vehicle injury collisions	0	120	16.8 (17.0)
	Years	6	6	6 (0)
	Major AADT	1322	34364	13822 (5657)
	Minor AADT	14	27936	3914 (3930)
	No. of approaches with left-turn lanes	0-396; 1-242; 2-486; 3-177; 4-328		
	No. of approaches with right-turn lanes	0-919; 1-380; 2-225; 3-59; 4-46		
	Intersection class	1-2; 2-25; 3-194; 4-6; 5-254; 6-106; 7-1; 8-438; 9-145; 10-23; 11-0; 12-339; 13-91; 14-5		
4-leg, signalized at-grade intersections from Edmonton, Alberta, Canada (515); (499 with data on turn lanes)	Multi-vehicle total collisions	0	555	75.5 (82.2)
	Multi-vehicle injury collisions	0	195	22.8 (25.5)
	Years	6	6	6 (0)
	Major AADT	4720	70331	24674(10849)
	Minor AADT	102	34926	9634 (7004)
	No. of approaches with left-turn lanes	0-93; 1-74; 2-158; 3-45; 4-129		
	No. of approaches with right-turn lanes	0-167; 1-95; 2-138; 3-20; 4-79		
	Area	urban-251; suburban-264		

CHAPTER 2 PRE-TEST DATA SAMPLING TO SELECT LOCAL REFERENCE GROUPS

An often neglected, but essential step in traffic safety analysis is data sampling. Researchers in the traffic safety field tend to pour their efforts into model development or comparison analysis, with less focus on the procedure for data sampling. In before-after evaluation (B/A) procedures, reference groups are used to develop SPFs. While there is plenty of research on SPF development, very few studies have examined the actual selection of reference groups. Practitioners and researchers tend to include the entire reference population or to arbitrarily select any available data sources without carrying out statistical sampling. The drawbacks of this ad-hoc approach are clear: on the one hand, given that the whole reference population is applied, the data items are not always available or worthwhile to collect, especially when many items need field surveying or manual inputs; on the other hand, if arbitrary selection is applied, the reference group may not be sufficient enough to conduct the next step in modeling, or may not be consistently representative of the reference population.

This chapter aims to address these drawbacks by exploring and establishing a data sampling and a data assignment mechanism that are specifically designed to work for B/A processes. This investigation is comprised of three steps.

The first step, described in Section 2.1, is to review data sampling and data assignment related literature and then to recommend the methods most suitable for the data sampling for this dissertation.

The second step, described in Section 2.2, is to estimate an appropriate sample size for the reference group by controlling the modeling power error level.

The third step, described in Section 2.3, is based on the outcome from the previous two steps and conducts random data sampling procedures to select the reference group to meet two goals: the

reference group will have sufficient samples to develop models with the controlled power level; and the reference group will be a legitimate representative of the reference population so that the models developed from the reference group are identical to the models developed from reference population.

After these three steps are conducted, Section 2.4 will examine the data sampling effects by comparing variables of reference group vs. reference population, in order to prove that the reference groups have consistent statistical features with relevant reference population, and therefore, are legitimate representatives of the reference population.

Finally, Section 2.5 summarizes all outcomes of this chapter.

2.1 CONCEPTS, METHODOLOGY AND NECESSITY

This section constructs the theoretical and methodological foundation for all following analysis applications. Sub-section 2.1.1 is the general introduction for random sampling (RS) and random assignment (RA) strategies. Sub-section 2.1.2 describes the selection of appropriate RS and alternative RA approaches which will be applied for this dissertation studies. Sub-section 2.1.3 emphasizes the rationality of data sampling in the context of B/A process.

2.1.1 Basic Concepts

This sub-section introduces the basic concept of RS, RA and their relation with B/A validity or bias.

Dattalo (2010) systematically described B/A oriented random sampling (RS) and random assignment (RA) strategies, among which there are three key concepts: selection bias, external validity and internal validity.

Selection bias is the introduction of errors due to systematic differences in the characteristics of participants and nonparticipants in a study (reference groups and treated groups in B/A processes). Two types of selection biases can be distinguished: sampling and assignment. In sampling bias, error results from failure to ensure that all members of a reference population have a known chance of being selected for inclusion in a sample. In assignment bias, error results from systematic differences in the characteristics of those allocated to an intervention (treated) group versus a control group in an experimental study (Dattalo, 2010). (Note: a "control" group is used for experimental studies; however, traffic safety B/A processing is an observational study. In such observational cases, a reference group is used instead).

RS is how a sample is drawn from a population, and this affects the external validity of a study's results. RA, at the same time, is how participants are allocated to the treated group and reference group, and is related to the internal validity of a study's results (Dattalo, 2010).

2.1.2 Methodology Selection

Both RS and RA have a variety of approaches and methods, this subsection aims to compare the characters and utilities of all those optional RS and RS approaches and then to recommend the one most appropriate for this dissertation investigation

In practice, there are adjustments and/or substitutions for RS and RA, which are called "alternatives", while others used as compensation on top of RS and RA are called "supplements" (Dattalo, 2010). Meanwhile, strategies used before or during RS and RA are categorized as "methodological" while others used as adjustments after RS and RA are classified as "statistical". Dattalo (2010) listed the available strategies based on different combinations, as shown in Table 2-1.

Table 2-1 implies that sampling strategies do not have to be "random"; as a result, the terminology is generalized thereafter in this dissertation: "data sampling" replaces random sampling (RS) while "data assignment" replaces random assignment (RA).

The dissertation research pertains to specific sampling and assignment approaches by taking into account two considerations. The first consideration is the stage when sampling is conducted. At the pre-testing stage, the dissertation needs a "before and during" sampling approach so it has to be methodological rather than statistical. Another consideration is the sample size. Sampling strategies include fixed-sample design, in which sample size is set in advance, or sequential sampling in which sample size is eventually determined (Stephens, 2001). The dissertation research will not fix the sample size before sampling, so it has to be sequential. As a result, the appropriate sampling strategy for this dissertation is a "sequential sampling methodology" per Table 2-1.

Table 2-1 Methodological/Statistical Alternatives and Supplements to RS and RA

Procedure	Strategy	Supplement versus Alternative	Methodological versus Statistical
Sampling	I. Deliberate Sampling	Alternative	Methodological
	II. Sequential Sampling	Alternative	Methodological
	III. Randomization Tests	Alternative	Statistical
	IV. Multiple Imputation	Supplement	Statistical
	V. Mean-score Logistic Regression	Alternative or Supplement	Statistical
Assignment	I. Sequential Assignment and Treatment-as-Usual Combined	Alternative	Methodological
	II. Partially Randomized Preference Trial	Alternative	Methodological
	III. Constructed Comparison Group	Alternative	Statistical
	IV. Propensity Scores Matching	Alternative or Supplement	Statistical
	V. Instrumental Variables Methods	Alternative or Supplement	Statistical

On the other hand, since assignment is not the topic of this dissertation, only after-assignment adjustment is needed, i.e., a statistical approach. Further investigation necessitates that the mitigating of systematic differences between reference groups and treated groups is vital for this dissertation research. Accordingly, "propensity score matching" per Table 2-1 matches the dissertation requirements and was selected.

This chapter addresses the sequential sampling methodology while Chapter 6 concentrates on propensity score matching.

2.1.3 Necessity of Data Sampling

The fundamental purpose of this dissertation is to determine a better B/A methodology for safety treatment effect evaluation. As for local practices, not all data items are easily accessible. Some documented or conventional inventory data items, such as collisions and traffic volume, might be obtained with ease while other observed data items, such as turn lanes of intersections, usually require field surveys or labor-intensive manual means and can only be processed site by site, which can be costly.

To reduce the difficulties and enhance the feasibility of data acquisition, sample size control is essential. A viable solution is to conduct data sampling based on the population with documented data items, then obtain representative samples, and finally supplement the observed items only with respect to the samples.

The reference population used in this dissertation, as shown in Section 2.2, already contains all of the data items. However, for research purposes, data sampling will still be conducted and the modeling procedures would proceed on both population and sample, followed by similarity comparisons, in order to verify the representativeness of the samples.

After theoretical overviews and selection of the most appropriate approaches described in this section, the subsequent sections will address the application of these methods to conduct real data sampling processing.

2.2 DETERMINATION OF DATA SAMPLE SIZE

Determination of the sample size is the first step of a data sampling process. Before conducting any data sampling, sample size must be set up in advance. Otherwise, data sampling loses target and direction.

Theoretically, research targets regarding sample size are not only the low threshold sufficient for model significance, but also the influence of sample size on model performances. However, the model performance issue will be addressed by strategies such as diversified safety modeling, model integration and data matching procedure, so this chapter will focus only on the lower threshold of sample size with respect to model significance.

Accordingly, Sub-section 2.2 explores an innovative method within the traffic safety domain to estimate appropriate sample size based on target model power (a concept to be introduced in depth later). The outcomes will serve the physical data sampling procedures introduced in Sub-section 2.3.

Sub-section 2.2 contains three parts: review of relevant literatures, introduction of population data used for the subsequent analysis and then the conduction of model power and data sample size estimation.

2.2.1 Relevant Past Researches

Compared to the abundant amount of work on safety modeling in the literature, data sampling themed research, such as determination of sample size in road safety, is scarce. This is unfortunate, given that sampling is supposed to be a prerequisite of any model intended to represent the population.

Signorini (1991) determined the sample size for a Poisson regression model, a popular tool in road safety analyses. The asymptotic variance of the maximum likelihood estimate (MLE) of the parameters was used to calculate the sample size required to test the hypotheses on the parameters controlled by a given Type I error rate (probability of incorrectly identifying a statistically significant effect, denoted as "α"), and a given model power error (probability of not identifying a statistically significant effect when one exists, denoted as "B"; Accordingly, (1- B) is called as "model power level", usually simplified as "power") (Dattalo, 2010). Assume that the count responses is Y_i on N individuals subjected to exposure t_i, so that $\lambda_i = E[Y_i]$ as follows:

$$\lambda_i = t_i \exp(\beta_0 + \beta^T x_i) \quad (2\text{-}1)$$

where

β_0 is the intercept,

x_i is a p–dimensional vector of the covariates,

$\beta^T = (\beta_1,\ldots,\beta_p)^T$ is the corresponding p–dimensional vector of the parameters,

x_i and t_i are regarded as realizations of independent random variables X and T, where $X \sim f_X(x)$ and $T \sim f_T(t)$ with mean exposure time μ_T.

Assume exposure time t_i is independent of x_i for each sample; then the likelihood function from the joint distribution of Y, T and X will be:

$$L(\beta_0, \beta) = \prod_{i=1}^{N} f_X(x_i) f_T(t_i) \lambda_i^{y_i} \exp(-\lambda_i) / y_i! \quad (2\text{-}2)$$

Let b_0 and b denote the respective MLEs of β_0 and β, obtained by maximizing the likelihood function $L(\beta_0, \beta)$. As N increases, the standard asymptotic theory states that these converge in distribution to a multivariate normal distribution, with mean $(\beta_0,\beta^T)^T = \beta^*$ and variance–covariance matrix I^{-1}, where I is the Fisher information matrix with elements given by:

$$I_{jk} = -E\left(\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k}\right) \ (j,k = 0,\cdots,p)$$

Hence in this case,

$$I_{jk} = N\mathrm{E}\{\mathrm{T} X_j X_k \exp(\beta_0 + \beta^T X)\}(j,k = 0,\ldots,p)$$

where $X_o = 1$.

Given the independence of *T* and *X*,

$$I_{jk} = N\mu_T \exp(\beta_0)\, E\{X_j X_k \exp(\beta^T X)\} \quad (2\text{-}3)$$

Where μ_t is the mean exposure time. Define the moment generating function with the covariates X and coefficient vector s as m(s) = E {exp ($s^T X$)}.

For i, j=1,…, p, let $m_i = \frac{\partial m}{\partial s_i}$, $m_{ij} = \frac{\partial^2 m}{\partial s_i \partial s_j}$, $m_0 = m_{00} = m$ and $m_{i0} = m_{oi} = m_i$ and form the (p+1)×(p+1) matrix M = (m_{ij}), then:

$$I(\beta_0, \beta) = N\mu_T \exp(\beta_0) M(\beta) \quad (2\text{-}4)$$

Hence, the MLE $\hat{\beta}$* of β* is asymptotically, as N→∞, multivariate normal with mean β* and covariance matrix$(N\mu_T)^{-1} \exp(-\beta_0)\, M^{-1}(\beta)$.

Suppose β_1 is the parameter of interest, and we wish to test the null hypothesis $H_0: \beta = \beta_N = (0, \beta_2, \cdots, \beta_p)$ against the alternative hypothesis $H_1: \beta_A = (\tilde{\beta}, \beta_2, \cdots, \beta_p)$, at a significance level, α with a power at least 1-B. Assuming N is large enough to apply the asymptotic results derived above, the asymptotic variance of $\hat{\beta}_1$ is given by the second diagonal term of I^{-1}. A routine calculation gives:

$$N\mu_T \exp(\beta_0) \geq \left\{ z_a V^{\frac{1}{2}}(\beta_N) + z_B V^{\frac{1}{2}}(\beta_{NA}) \right\}^2 / \tilde{\beta}_2 \quad (2\text{-}5)$$

where V(β)=$\{M^{-1}(\beta)\}_{22}$, the second diagonal term of M^{-1}, evaluated at β, and z_δ is the 1-δ point of the standard normal distribution.

Another approach is the Bayesian method to conduct sample size determination (Bayesian SSD). In the middle 1990s, Joseph et al. (1995) reviewed several properties of Bayesian SSD approaches, summarized some methods and also recommended a general algorithm for Bayesian SSD based on Monte Carlo simulation. However, Joseph et al. (1995) did suggest there was no general rule on how to select criteria used in Bayesian SSD, and this was very case-specified. As they highlighted, Bayesian sample size calculation is highly reliant on computer-intensive methods and statistics evolution. More recently, Wang and Gelfand (2002) conducted a "simulation-based" Bayesian SSD procedure. They utilized a "loop for Monte Caro Integration"

with iterative steps including prior sampling, data collection, prior fitting, model fitting, posterior samples and model performance criteria. Appropriate sample size is determined until model performance is satisfactory. This method has the advantage of being unrestricted to an arbitrary "model power" while having some disadvantages as acknowledged by the authors themselves. These are: it sacrifices explicit SSD formula; it requires full model specification; and, most importantly, it is computationally intensive.

In their technical report, Ivan et al. (2010) employed the above mentioned methodologies (Signorini, 1991) to estimate the sample size for a safety model that used a Poisson log-linear regression for wet pavement friction. At a selected level of significance, 0.05, given the chosen N, their fitted model was checked with the null hypothesis above, and the power calculated as the number of rejections divided by N. In the end, they found an optimal sample size N that accommodates an appropriate power level.

Ye and Lord (2010) examined the effects of sample size for the three most commonly used crash severity models. The sample sizes were 100, 250, 500, 1000, 1500, 2000, 5000, and 10000. The recommended absolute minimum number of observations for the different models was carried out via Monte Caro simulation. Since the sample sizes were discretely chosen in advance, their research work is more an examination of sample size rather than truly a prospective estimation.

Lord and Miranda-Moreno (2008a) sought to verify whether a small sample size and low sample mean values affect the estimation of the posterior mean of the dispersion parameter when the multi-level negative binomial (HNB) model was used to develop crash prediction models. A series of Poisson-gamma distributions was simulated by using different values that described the mean, dispersion parameters, sample size, and the prior distribution. They concluded that crash data characterized by a low sample mean, combined with a small sample size can seriously affect the estimation of the posterior mean of inverse dispersion parameters in HNB models. After sufficient simulation runs, they recommended a series of minimum sample sizes depending on different population sample means (Lower means require higher sample sizes.) and whether the priors are vague or non-vague (Vague priors require higher sample sizes.). One issue of this research is that the sample sizes were pre-set and discrete (20, 100, 500, etc.). Similar to the

research carried out by Ye and Lord (2010), this is also an examination rather than a prospective estimation of the sample size.

2.2.2 Population Data

The populations on which all subsequent analyses for this dissertation were conducted, consisted of 4SG intersections in Toronto and Edmonton. Relevant descriptions and their summary statistics can be found in Section 1.5.

2.2.3 Power and Data Sample Size Analysis

A. Conceptual methods and formulas

Statistically speaking (Dattalo, 2010), the sample size in a study is determined a priori by establishing null and alternate hypotheses with respect to a primary parameter of interest (θ), and then specifying a Type I error rate (α) and power (1-B) to be controlled for a given treatment effect size (θ=Δ). Type I error is the probability of incorrectly identifying a statistically significant effect. A power error is the probability of not identifying a statistically significant effect when one exists. Usually, traditional values of α and B are used (i.e., α=0.05, B=0.20). Using a sample size that is small relative to the selected effect size can result in an "underpowered" study (i.e., unlikely to detect a smaller, but possibly still important effect).

The conventional level of model power (1-B) in sample size calculations is 80%, which means that the sample size, N, is to be selected such that 80% of the possible 95% (1-α) confidence intervals of the estimate will not exceed a given standard deviation. When N is increased, the estimate becomes closer to the true value (Gelman and Hill, 2007). No matter how prevailing the value of 80% is, it is arbitrary. As some past research suggests, this issue can be addressed through Bayesian approaches. There is a variety of other model performance criteria to replace the arbitrary power (Wang and Gelfand, 2002; Joseph et al., 1995). All these Bayesian criteria have flexible local critical values for model performance criteria from posterior samples. By this means, they seem to have conceptual advantages.

However the Bayesian approach, regardless of the advantage of flexibility, is complicated, computation-intensive and relies on advanced computer methods (Wang and Gelfand, 2002; Joseph et al., 1995). Most significantly, as Dmitrienko et al. (2007) indicate, in SAS software there is no direct plug-in procedure to conduct Bayesian sample size determination. To do this, users must use the programming language – IML in SAS (SAS Institute Inc., 2012; Dmitrienko et al., 2007) to develop initial functions, a technically complicated procedure with lots of difficulties.

On the contrary, in applying classical model power and sample size analysis methods, SAS uses a procedure called "GLMPOWER" (SAS Institute Inc., 2012), which is adaptable for generalized linear models.

Finally, this dissertation selected the classical method introduced by SAS Institute Inc. (2012) with "GLMPOWER" procedure to perform prospective power and sample size analyses for a variety of goals, including determining the sample size required to obtain a significant result with adequate probability (power). Here, prospective indicates that the analysis pertains to planning for a future study, in contrast to the retrospective power analysis for a past study. GLMPOWER is one of several power and sample size analysis tools in SAS and is especially relevant for generalized linear models (GLMs) with a variety of complexities.

The SAS GLMPOWER procedure is developed based on GLMs with the following form (SAS Institute Inc., 2012):

$$Y=X\beta+\varepsilon \qquad (2\text{-}6)$$

where Y is the N×1 vector of responses, X is the N×p design matrix (Note: X is not a random variable, contrary to the X defined in earlier equations of this chapter.), N is the sample size, β is the p×1vector of the model parameters and ε is the N×1 vector of the error terms.

A general linear hypothesis to test the effect of univariate models is:

$$H_0: L\beta = \theta_0$$
$$H_A: L\beta \neq \theta_0 \quad (2\text{-}7)$$

where L is a $r_L \times p$ contrast matrix to represent r_L linear functions need to be estimated for β and θ_0 is the null value (usually just a vector of zeros).

The test statistic is:

$$F = \frac{(\frac{SS_H}{r_L})}{\hat{\sigma}^2} \quad (2\text{-}8)$$

where

$$SS_H = \frac{1}{N}(L\hat{\beta} - \theta_0)'(L(X'X)^{-1}L')^{-1}(L\hat{\beta} - \theta_0),$$

$$\hat{\beta} = (X'X)^{-1}X'y, \text{ and}$$

$$\hat{\sigma}^2 = \frac{1}{DF_E}(y - X\hat{\beta})'(y - X\hat{\beta})$$

where DF_E = N-rank(X). Note that DF_E = N-p if X has full rank.

Under H_0, $F \sim F(r_L, DF_E)$. Under H_A, F is distributed as $F(r_L, DF_E, \lambda)$ with non-centrality $\lambda = N(L\beta - \theta_0)'\left(L\left(\ddot{X}'diag(w)\ddot{X}\right)^{-1}L'\right)^{-1}(L\beta - \theta_0)\sigma^{-2}$. Here $\ddot{X}$is in essence the design matrix - the collection of unique rows of X, and w is a weight vector; see more details at SAS Institute Inc. (2012). To further conduct an adjustment for the covariates, let n_v be the number of covariates; then the power of the testing (Muller and Peterson, 1984) is:

$$\text{Power}=P(F(r_L, DF_{E-}\, n_v, \lambda^*) \geq F_{1-\alpha}(r_L, N\text{-rank}(x)-\, n_v) \quad (2\text{-}9)$$

where λ* is calculated by an adjusted error standard deviation σ* as:

$$\lambda^* = N(L\beta - \theta_0)'(L(\ddot{X}'\text{diag}(w)\ddot{X})^{-1}(L\beta - \theta_0)(\sigma *)^{-2}$$

and where in the power analysis design matrix, X is parameterized in three parts: $\{\ddot{X}, w, N\}$. $\ddot{X}$ is q×p essence design matrix, the collection of unique rows of X is referred to as “design profiles”; the q×1 weight vector w reveals the relative proportions of design profiles; N is the sample size and also the number of rows of X (Muller and Benignus, 1992a; Muller et al., 1992b).

The SAS GLMPOWER procedure is operated based on the power equation of Equation 2-9 which connects the significance level (α), power (1-B), surmised response means for subject profiles (often called "cell means"), surmised variability, etc. (SAS Institute Inc., 2012). When one item is set to “null” while the others are pre-set to the surmised values, then the null item is the output of the GLMPOWER procedure. For example, given a certain sample size, Equation 2-9 can be used to estimate power; on the contrary, the sample size would be computed by inverting the power equation in Equation 2-9 given a selected power level.

B. Computational procedures and outputs

Power and sample size analyses were conducted based on the reference population described in Table 1-2. The statistical procedure in SAS (SAS Institute Inc., 2012) to analyze the power and sample size requires users to presume a model form supposed to be developed based on the estimated sample. Since, in this dissertation, the sample data will be used for future safety model development with generalized linear error structure as the default type, the potential model form is set up as follows:

Response variable: collision

Predictors: logarithm (Major AADT), logarithm (Major AADT), No. of left turn lanes, No. of right turn lanes, and local special factor (2-10)

where local factor is the intersection class for Toronto and area for Edmonton.

The analysis consists of two steps. Step 1 locates the appropriate standard deviation which reflects the traditional power level, for instance, B=0.20. This procedure is conducted to set the size of the whole population as the sample size, then to surmise a reasonably wide incremental range of standard deviations (for instance, from 1 to 5, with each incremental step as 0.01),

introduce an operation based on the power equation in Equation 2-9, and obtain a series of power values as outputs. Finally, an appropriate standard deviation can be estimated by inverting the given power level from the list backwards to a relevant standard deviation value. Equivalent to the traditional B=0.20 power value, the standard deviation was estimated as 2.70 for the Toronto reference population, and 4.31 for Edmonton.

Step 2 estimates the approximate sample size by matching power levels. This is based on the rational assumption that the standard deviation of a population equals that of the samples. Hence, it was possible to conduct this step by keeping the standard deviation of the population as a sample standard deviation, which surmises a reasonably wide range of sample sizes, and then applying the power as the output by plotting the profile changes against the sample size. Figures 2-1 and 2-2 illustrate the SAS outputs for the power curves that pertain to the data samples from Toronto and Edmonton.

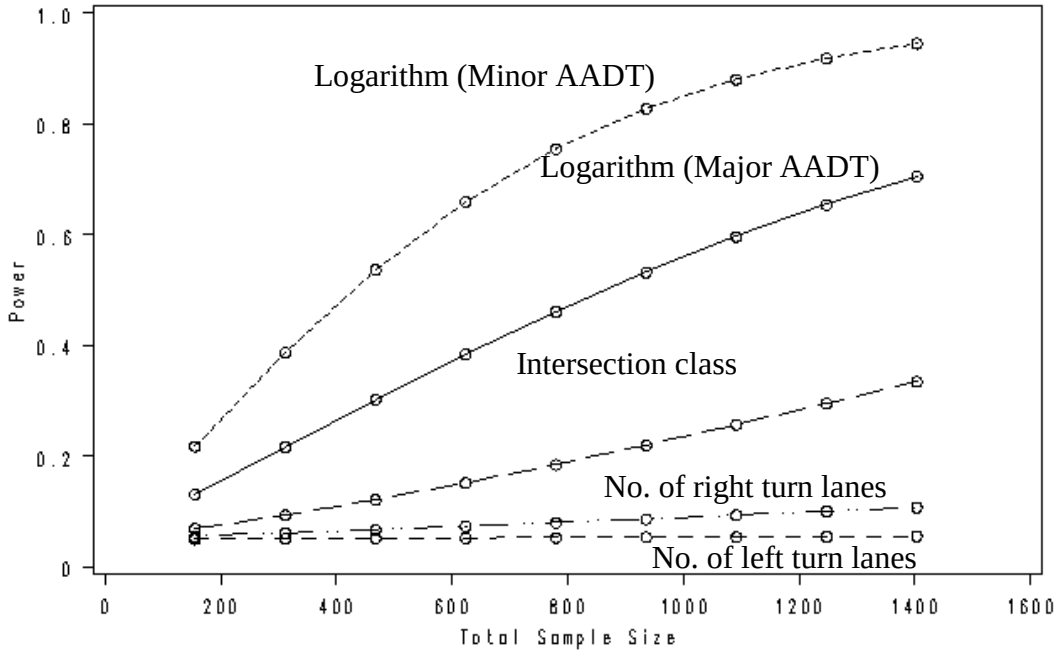


Figure 2-1 SAS Output of Power Curve for Toronto Data

To be specific, the control variable in Figure 2-1 and 2-2 is AADT and not the number of turning lanes. Because of the specific data sampling strategy conducted at Section 2.3, data samples will

definitely cover any number of turning lanes no matter how high or low the sample size is, so the variables for number of turning lanes won't govern the determination of sample size.

Given that the traditional power (1-B) = 0.80, a sample size between 600 and 800 seems likely to satisfy the modeling requirements for Toronto.

Given that the traditional power (1-B) = 0.80, a sample size between 300 and 400 seems likely to satisfy the modeling requirements for Edmonton.

The two AADT variables did yield different power curves. Considering the principle is to select samples as small as possible so as to increase efficiency, the curves leading to smaller sample sizes would govern. Moreover this sample size estimation is preliminary and the final result will be determined by the model fitting in Chapter 3 and Chapter 4. In the case where model fits are insignificant, the sample size will be reset. So the criterion to select smaller sample size won't have irrevocable negative influence.

These estimated sample sizes play just tentative and referential roles for the next step in data sampling. The real sample sizes will be adjusted by a sampling procedure shown in Section 2.4 and will finally be validated by a significance test of the model estimation as shown in Chapter 3.

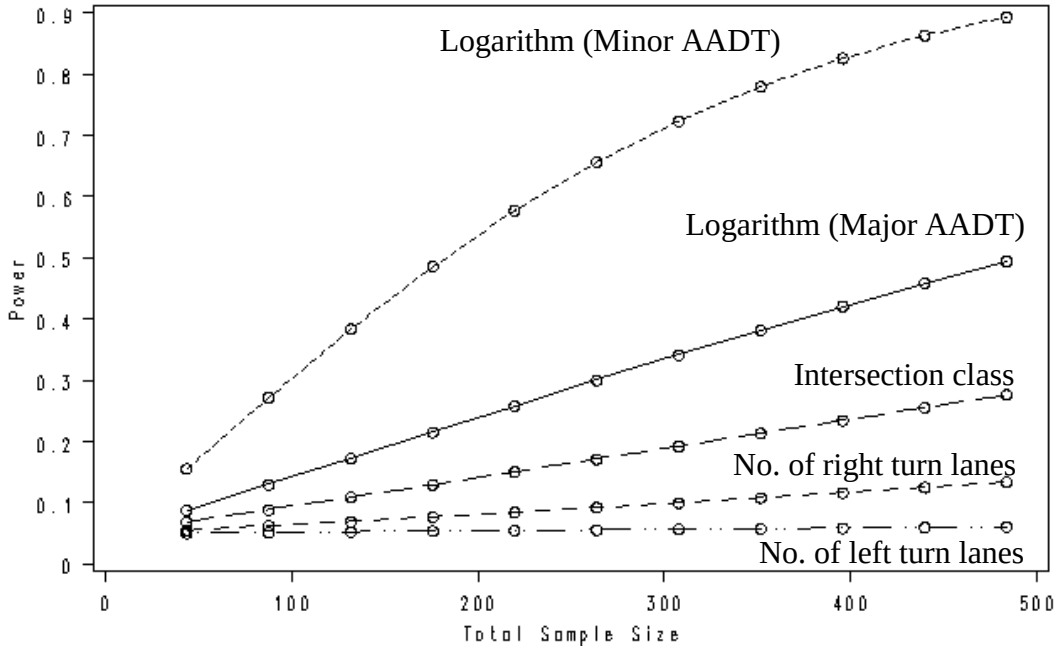


Figure 2-2 SAS Output of Power Curve for Edmonton Data

2.3 DATA SAMPLING

Data sampling procedures were conducted with the above obtained numbers as the initial sampling sizes.

2.3.1 Relevant Past Research in the Road Safety Area

Ivan et al. (2010) selected random locations to incorporate wet pavement friction into safety analyses based on an optimal sample size determined by the method mentioned in Section 2.3.1. They basically applied an RS approach to generate a "found" group, a "random" group with similarities in road characteristics compared with the "found" group, and finally, a merged, "combined" group, based on which different statistical models were developed.

Luo et al. (2008) applied a stratified sampling method to identify samples for black spot selection. Their method assumed that there are a total of N sites in the data collection, which are divided into K layers, with each layer containing N_i sites. This means:

$$\sum_{i=1}^{K} N_i = N \quad (2\text{-}11)$$

Then, n_i sites were randomly selected within each layer; these sites constituted the total sample.

The numbers of n_i were determined by:

$$n_i = n\frac{N_i \cdot \delta_i}{\sum N_i \cdot \delta_i} \quad (2\text{-}12)$$

where n is the estimated total sample size with $\min(n) = \frac{N \cdot [\sum w_i \cdot \delta_i]^2}{N \cdot V_{opt} + \sum w_i \cdot \delta_i^2}$

δ_i is the standard deviation of each layer,

$w_i = N_i/N^2$, and

V_{opt} is the pre-set criterion of variance for crash estimates

2.3.2 Data Sampling Procedure

A. Approach and method for data sampling

There is a paucity of literature on data sampling in the road safety field despite the fact that that this is an essential process for any model that is intended to represent the entire population. Notwithstanding this void, the fundamental methodologies for all types of data sampling are sufficiently mature and have been well established within SAS through a specific procedure called "SURVEYSELECT" (PROC SURVEYSELECT) (SAS Institute Inc., 2012).

As described in Section 2.1, the sequential sampling approach is required for the purpose of this dissertation research. Furthermore, both the Toronto and Edmonton data have categorical items including number of left turn lanes, right turn lanes, and intersection class (or area for Edmonton). That is to say, the reference population data are stratified. Under this circumstance, a stratified sampling approach is desirable; otherwise some categories may be neglected (SAS Institute Inc., 2012). As result, this dissertation has applied a stratified sequential sampling approach.

As for the sampling methods, PROC SURVEYSELECT provides both equal probability sampling and probability proportional to size (PPS) sampling. In the former, each unit in the sampling frame, or in a stratum, has the same probability of being selected for the sample. In the latter, the selection probability of a unit is proportional to its size measure – the unit numbers of the stratum (Lohr, 2009). Equal probability sampling, in the data used for this dissertation, may cause imbalanced distributions throughout all of the categories. Subsequently, the PPS data sampling method (SAS Institute Inc., 2012) is preferable. Thus, stratified sequential PPS data sampling was conducted for this dissertation.

B. Conceptual methods and formulas

With sequential PPS data sampling specified, PROC SURVEYSELECT uses Chromy's method (Chromy, 1979; Williams and Chromy, 1980), which sequentially selects units with probabilities proportional to size and with minimum replacement. Selection with minimum replacement means that the actual number of hits for a unit can equal the integer part of the expected number of hits for that unit, or the next largest integer. Sequential RS controls the distribution of the sample by spreading it throughout the sampling frame or stratum, thus providing implicit stratification according to the order of the units in a frame or stratum. According to Chromy's method of sequential selection, PROC SURVEYSELECT first randomly chooses a starting unit from the entire stratum with PPS. The procedure uses this unit as the first one and treats the stratum observations as a closed loop. The procedure sequentially numbers observations from a random start to the end of the stratum and then continues from the beginning of the stratum until all units are numbered (SAS Institute Inc., 2012).

For a stratified design, let n denote the sample size for the current stratum, N denote the stratum population size, and M_i denote the size measure for unit i in the stratum (while M denotes size measure for the whole stratum). According to the Hanurav-Vijayan algorithm (Hanurav, 1967; Vijayan, 1968), PROC SURVEYSELECT first orders units within the stratum in ascending order by size measure, so that: $M_1 \leq M_2 \leq \ldots \leq M_N$. $Z_j=M_j/M$ is defined for PPS sequential sampling, beginning with the randomly chosen starting unit. Then Chromy's method (Chromy, 1979) partitions the ordered stratum sampling frame into n_h zones of equal size. Beginning with the random start, the procedure accumulates the expected number of hits and computes:

$$E(S_{hi}) = n_h Z_{hi} \tag{2-13}$$

$$I_{hi} = \text{Int}(\sum_{j=1}^{i} E(S_{hi})) \tag{2-14}$$

$$F_{hi} = \text{Frac}(\sum_{j=1}^{i} E(S_{hi})) \tag{2-15}$$

where $E(S_{hi})$ represents the expected number of hits for unit i in stratum h, Int(·) denotes the integer part of the number, and Frac(·) denotes the fractional part. In sequentially sampling considering each unit, Chromy's method determines the actual number of hits for unit i by comparing the total number of hits for the first (i-1) units, $T_{h(i-1)} = \sum_{j-1}^{i-1}(S_{hj})$, with a value of $I_{h(i-1)}$.

If $T_{h(i-1)}=I_{h(i-1)}$, Chromy's method determines the total number of hits for the target unit as follows. If $F_{hi}=0$ or $F_{h(i-1)}>F_{hi}$, then $T_{hi}=I_{hi}$, otherwise $T_{hi}=I_{hi}+1$ with probability $(F_{hi}-F_{h(i-1)})/(1-F_{h(i-1)})$ and the number of hits for the unit, S_{hi}, equals $T_{hi}-T_{h(i-1)}$ (Chromy, 1979).

If $T_{h(i-1)}=(I_{h(i-1)}+1)$, Chromy's method determines the total number of hits for the target unit as follows. If $F_{hi}=0$, then $T_{hi}=I_{hi}$. If $F_{hi}>F_{h(i-1)}$, then $T_{hi}=I_{hi}+1$. Otherwise, $T_{hi}=I_{hi}+1$with probability $F_{hi}/F_{h(i-1)}$. Finally, the number of hits for the unit, S_{hi}, equals $T_{hi}-T_{h(i-1)}$ (Chromy, 1979).

C. Computational procedures and outputs

For data sampling from reference populations, as shown in Table 1-2, SAS PROC SURVEYSELECT with a stratified sequential PPS method is implemented. The variables grouping data into the strata are “intersection class” for Toronto and “area” for Edmonton. Moreover, optimal allocation was applied, in which the total sample size is allocated among the strata in proportion to stratum sizes, variances, and costs. Here, stratum size is the number of samples for each stratum (i.e., the frequency), while the major AADT was selected as stratum variance, and negative minor AADT was selected as the stratum cost. As a result, the generated sample was distributed throughout the strata proportional to each stratum size, while reflecting the scope of the major and minor AADTs of each stratum as well.

The outcomes from the power and sample size analyses were used as the initial sample size. For Toronto, the initial sample sizes were selected as 600 and 700. For Edmonton, the initial sample sizes were 300 and 400. After data sampling, the final sample sizes were slightly adjusted depending on when the sampling process stopped.

Table 2-2 summarizes the sample sizes obtained for the Toronto and Edmonton data.

Table 2-2 Summarized Statistics of Toronto & Edmonton Samples

Dataset (sample size, % of population)	Variable	Minimum Value	Maximum Value	Mean (Standard Deviation)
4-leg, signalized at-grade intersections from Toronto, Ontario, Canada (680, 42%)	Multi-vehicle total collisions	0	370	60.8 (59.3)
	Multi-vehicle injury collisions	0	120	16.5 (17.5)
	Years	6	6	6 (0)
	Major AADT	1322	34364	13213 (6094)
	Minor AADT	24	27936	4117 (4104)
	No. of left-turn lanes	0-170; 1-99; 2-200; 3-73; 4-138		
	No. of right-turn lanes	0-377; 1-157; 2-93; 3-30; 4-23		
	Class	1-2; 2-22; 3-84; 4-6; 5-98; 6-64; 7-1; 8-131; 9-72; 10-23; 11-0; 12-115; 13-57; 14-5		
4-leg, signalized at-grade intersections from Toronto, Ontario, Canada (588, 36%)	Multi-vehicle total collisions	0	370	61.0 (59.0)
	Multi-vehicle injury collisions	0	102	16.7 (17.5)
	Years	6	6	6 (0)
	Major AADT	1322	33504	13005 (5861)
	Minor AADT	14	27936	4040 (3897)
	No. of left-turn lanes	0-147; 1-84; 2-176; 3-62; 4-119		
	No. of right-turn lanes	0-325; 1-138; 2-82; 3-24; 4-19		
	Area	1-2; 2-19; 3-72; 4-6; 5-84; 6-54; 7-1; 8-112; 9-62; 10-23; 11-0; 12-99; 13-49; 14-5		
4-leg, signalized at-grade intersections from Edmonton, Alberta, Canada (400, 78%); (387 with data on turn lanes)	Multi-vehicle total collisions	0	555	73.8 (77.4)
	Multi-vehicle injury collisions	0	195	22.1 (23.7)
	Years	6	6	6 (0)
	Major AADT	4720	70331	24772 (10589)
	Minor AADT	102	34926	9656 (7000)
	No. of left-turn lanes	0-71; 1-58; 2-122; 3-36; 4-100		
	No. of right-turn lanes	0-129; 1-72; 2-107; 3-16; 4-63		
	Class	urban-195; suburban-205		
4-leg, signalized at-grade intersections from Edmonton, Alberta, Canada (300, 58%); (294 with data on turn lanes)	Multi-vehicle total collisions	0	555	76.6 (86.0)
	Multi-vehicle injury collisions	0	195	22.7 (26.07)
	Years	6	6	6 (0)
	Major AADT	5560	70331	24566 (10780)
	Minor AADT	171	30186	9702 (7092)
	No. of left-turn lanes	0-57; 1-44; 2-91; 3-27; 4-75		
	No. of right-turn lanes	0-103; 1-55; 2-80; 3-11; 4-45		
	Area	urban-146; suburban-154		

2.4 SAMPLING RESULT EXAMINATIONS

This section examines the validity of the sampling outcomes for all of the variables in the datasets.

2.4.1 Comparisons of Continuous Variable Values

Table 2-3 describes the comparison of their continuous variables in the sample versus the reference population. If the variable value in the sample equals that of the population, the relevant table cell is labeled as "same"; otherwise the ratio of the value in the sample to that in the population is shown.

Table 2-3 Variable Value Comparisons of Sample versus Population

Dataset (sample size)	Variable	Minimum Value	Maximum Value	Mean
4-leg, signalized at-grade intersections from Toronto, Ontario, Canada (680)	Multi-vehicle total collisions	same	same	98.1%
	Multi-vehicle injury collisions	same	same	98.2%
	Years	same	same	same
	Major AADT	same	same	95.6%
	Minor AADT	171.4%	same	105.2%
4-leg, signalized at-grade intersections from Toronto, Ontario, Canada (588)	Multi-vehicle total collisions	same	same	98.4%
	Multi-vehicle injury collisions	same	85.0%	99.4%
	Years	same	same	same
	Major AADT	same	97.5%	94.1%
	Minor AADT	same	same	103.2%
4-leg, signalized at-grade intersections from Edmonton, Alberta, Canada (400); (387 with data on turn lanes)	Multi-vehicle total collisions	same	same	97.7%
	Multi-vehicle injury collisions	same	same	96.9%
	Years	same	same	same
	Major AADT	same	same	100.4%
	Minor AADT	same	same	100.2%
4-leg, signalized at-grade intersections from Edmonton, Alberta, Canada (300); (294 with data on turn lanes)	Multi-vehicle total collisions	same	same	101.5%
	Multi-vehicle injury collisions	same	same	99.6%
	Years	same	same	same
	Major AADT	117.8%	same	99.6%
	Minor AADT	167.6%	86.4%	100.7%

Table 2-3 reveals that the boundaries of continuous variables between the samples and their original populations are significantly consistent. The mean values of the samples are also approximately equal to their original population means.

2.4.2 T-testing for Collision Variable

Specifically aimed at addressing the most critical variable - total multi-vehicle collisions - t-tests were conducted to statistically measure whether the sample and its population share a consistent mean. The results of the t-tests are shown in Table 2-4.

The results of the t-testing, as evidenced by the P-values in Table 2-4, illustrate that the null hypothesis should be accepted, i.e., each of Toronto and Edmonton samples has the same mean with that of its original population.

Table 2-4 T-testing for Total Multi-vehicle Collisions in Sample vs. Population

Statistical Measures	Toronto			Edmonton		
	Reference Population	Sample (size=680)	Sample (size=588)	Reference Population	Sample (size=400)	Sample (size=300)
Mean	62.02	60.84	61.05	75.55	73.84	76.56
Variance	3310.03	3510.70	3482.92	6754.86	5994.79	7391.44
Observations	1629	680	588	515	400	300
Pooled Variance		3369.09	3355.85		6422.69	6988.98
Hypothesized Mean Difference		0	0		0	0
df		2307	2215		913	813
t Stat		0.45	0.35		0.32	-0.17
P-value (T<=t) two-tail		0.66	0.73		0.75	0.87

2.4.3 Comparisons of Proportional Distributions of Categorical Variables

With regards to categorical variables, since the “intersection class” variable for Toronto and the “area” variable for Edmonton are subgroup variables of the strata, the samples have already been

proportionally distributed through these two variables for PPS sampling. Hence, they do not need to be reviewed any further. Validity reviews are thus carried out on other categorical variables. The proportional distribution of data for number of left turn lanes and right turn lanes for the samples and their original populations are shown in Figures 2-3 and 2-4 (inner and intermediate circles for samples, outer circles for populations).

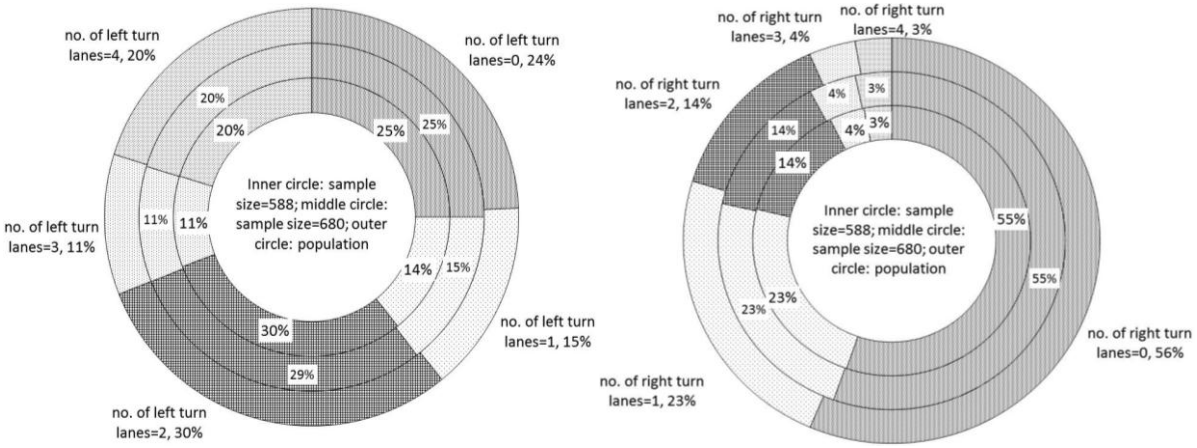


Figure 2-3 Distribution of No. of Turn Lanes for Toronto Samples and Population

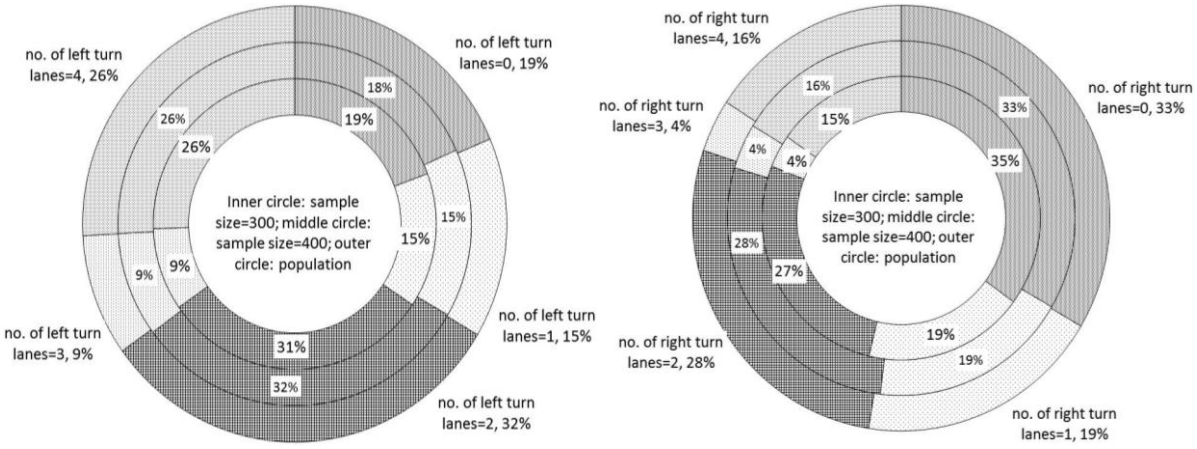


Figure 2-4 Distribution of No. of Turn Lanes for Edmonton Samples and Population

Figures 2-3 and 2-4 demonstrate that the proportional distribution of the samples is mostly consistent with that of their original population. The differences are negligible.

2.5 CHAPTER CONCLUSIONS

An often ignored, but very crucial step, in traffic safety modeling is data sampling. Without data sampling, modelers must either apply all the population data into modeling, which is often infeasible due to data collection difficulty, or apply arbitrarily derived subsets from population, leading to distortion and misrepresentation of the population. This chapter conducted a series of data sampling procedures to address this methodological issue for the pre-modeling stage, and also, to establish databases for subsequent aspects of the dissertation research.

While not prevalent in traffic safety modeling practices, data sampling has actually seen widespread application in many other fields, so the first part of chapter was able to review appropriate theories and strategies and to recommend the most suitable data sampling strategy for B/A process.

After the theoretical review, the second part of this chapter addressed sample size estimation, the prior step for data sampling. The concept of model power error (A power error is the probability of not identifying a statistically significant effect when one exists.) was applied as a control on estimating a sufficient sample size that would lead to a statistically significant model with a controlled power level.

Supported by outcomes of the sample size estimation, the third part of this chapter completed the investigation of the data sampling process. The "stratified sequential PPS (probability proportional to size) method was recommended as the most adaptable to the specific samples applicable to this dissertation study. A relevant SAS procedure was conducted to yield the sampled reference groups.

Following the presentation of the data sampling process, the fourth part of this chapter was aimed at evaluating the validity of sampling outcomes. For this evaluation, different types of variables in the reference group were compared to their counterparts in the reference population through different approaches. Finally, the descriptive statistics of continuous variable values, proportional distribution analyses of categorical variables and t-tests for collision variables lead

to one conclusion: that the samples have consistent boundaries and central tendencies, and similar distributions compared with their original reference population. That is say, they are legitimate representatives of the population and the sampling process is valid.

This chapter set the data foundations for the research outlined in the ones to follow, making modeling, posterior inferences, and final B/A evaluation feasible.

Based on the research outcomes in this chapter, Chapter 3 will be dedicated to local safety model development. Moreover, the statistical consistency of the sampled reference groups compared to the reference population, preliminarily proven in this chapter, will be further validated by model development based on the samples and the reference population, as will be discussed in Chapter 3.

CHAPTER 3 MODEL CALIBRATION AND STANDARD LOCAL SAFETY PERFORMANCE FUNCTION DEVELOPMENT

Treatment effect estimation of a before-after evaluation (B/A) process is a comparison for the treated group of what really happens after treatment versus what would have happened without the treatment (which the reference group is used to estimate). On one side of this comparison, actual collisions after treatment are obtained from straightforward observation and there is no methodological issue at all. However on the opposite side, the postulated collisions given no treatment cannot be obtained directly from treated group itself; instead, it must go through an indirect procedure based on safety models established from reference groups. Since these models are developed prior to comparison, for B/A, the model can be regarded as a source of prior knowledge.

For any local jurisdiction intending to apply a B/A, there are three optional safety models used for treatment effect estimation: exported modeled from outside, locally developed model, or combined/averaged model from both outside and inside. This chapter will be dedicated to outside model calibration to local conditions and the standard development of local models, while Chapter 4 will investigate diversified local safety model development.

The first part of this chapter, described in Section 3.1, evaluates the performance of outside models (in current practice, these are mainly HSM models) calibrated to a local jurisdiction, highlighting the substantial drawbacks of this approach. Then, as described as Section 3.2, theoretical frameworks for local model developments are investigated and optional full model and multi-level paradigms are compared.

Regardless of the theoretical superiority of a multi-level structure, both full models and multi-level models are retained and statistically estimated based on sample data described in Section 3.3. These two styles of model developments are separately described in Section 3.4 and 3.5.

Afterwards, model estimation results are to be compared and discussed. Finally, chapter conclusions and further investigation plans are provided in Section 3.7.

3.1 CALIBRATED HSM MODEL AS A PRIOR KNOWLEDGE SOURCE

As the first part of study, this section does not seek to introduce a new approach of safety model development. Instead, it is intended to evaluate the existing models established by outside sources. For current practice, the default choice of outside models is from the HSM. However, each HSM model must be calibrated before it can be deployed for a local B/A application. This section reviews the procedure of a typical model calibration, makes goodness-of-fit tests for the calibrated models and finally, discusses pros and cons, highlighting limitations of the model calibration approach.

3.1.1 Basic Concepts of Model Calibration

The most common source for determining prior knowledge in before-after evaluations is a collision predictive model, also known as an SPF, for e.g., the HSM SPF as shown in Equation 1-1 (AASHTO, 2010).

The HSM (AASHTO, 2010) stipulates that SPFs must be calibrated before applied to a local region. The calibration approach of HSM SPFs is to introduce a calibration factor, C_x, to accommodate local conditions for site type x. This factor is estimated with a five-step calibration procedure (AASHTO, 2010) as follows:

Step 1 – identify facility types for which the predictive model is to be calibrated,

Step 2 – select sites for calibration of the predictive model for each facility type,

Step 3 – obtain data for each facility type applicable to a specific calibration period,

Step 4 – apply the uncalibrated predictive model (Equation 1-1 without C_x) to predict total crash frequency for each site during the calibration period as a whole, and

Step 5 – compute calibration factors as the ratio of the sum of observations in a local sample to the sum of predictions for that sample from the uncalibrated model obtained as follows:

$$C_x = \frac{\sum_{all\ sites}(observed\ \ crashes)}{\sum_{all\ sites}(predicted\ \ crashes)} \quad (3\text{-}1)$$

3.1.2 Relevant Past Research

Many instances of research work have contributed to, or addressed the HSM calibration procedure.

Tarko (2006) suggested that universal calibration factors be supplemented with factors for subsets defined by users. Persaud et al. (2002) argued that a procedure that pursues a single calibration factor may be inappropriate and disaggregation by traffic volume might be preferable. Sawalha and Sayed (2006) recommended that calibration should be simultaneously undertaken, by using an MLE method for both the base SPF multiplier and the "shape parameter" (the inverse of the dispersion parameter of the negative binomial (NB) distribution assumed in estimating the SPF).

Despite these efforts, the calibration approach recommended by the HSM (AASHTO, 2010) is still the established method in practice so far, thanks to its simplicity and ease of understanding for many practitioners. The calibrated HSM model is still most widely used procedure to develop a prior knowledge source in many local regions.

3.1.3 Goodness-of-fit Tests for Calibrated HSM Model

A. Conceptual methods and formulas

The burning question that arises from Equation 3-1 is: Can a single factor, $C_{x,}$ sufficiently address all local characteristics?

GOF tests are therefore conducted to assess model performance in addressing this question. These consist of two types of measures: calculated overall measures and cumulative residual (CURE) plot that assesses model performance over the entire range of each covariate. The calculated overall measures often used are Pearson product moment correlation coefficient (r), mean prediction bias (MPB), mean absolute deviation (MAD) and mean squared prediction error (MSPE) (Oh et al., 2003; Lord et al., 2010).

The Pearson product moment correlation coefficient (r) is a measure of the linear association between observed and predicted collisions, and its function is:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (\hat{y}_i - \hat{\bar{y}})}{[\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(\hat{y}_i - \hat{\bar{y}})^2]^{1/2}} \quad (3\text{-}2)$$

where

n is the sample size,

y_i and $\hat{y}_i$ are the observed and estimated mean values at site i, respectively, and

$\bar{y}$ and $\hat{\bar{y}}$ are the means of y_i observations and $\hat{y}_i$ prediction, respectively.

The MPB provides a measure for magnitude and direction of the average model bias in comparison with the validation data. Its value is estimated by:

$$MPB = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad (3\text{-}3)$$

The MAD provides a measure of the average "misprediction" of the model. It differs from the MPB in that positive and negative prediction errors do not cancel out each other. Its value is estimated by:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (3\text{-}4)$$

The MSPE is the sum of the squared differences between the observed and predicted crash frequencies divided by the sample size. Its value is estimated by:

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (3\text{-}5)$$

The criteria for assessing these GOF measures are listed in Oh et al. (2003) as follows:

- r – values that are closer to 1.0 suggest a good fit; that is, the plotting of observed versus predicted values is close to a straight-line through the origin with a slope of 1, and
- for the MPB, MAD and MSPE – values closer to 0 suggest a good fit.

CURE plot has now become a standard method of assessing how well models fit the data over the full range of each individual covariate (Hauer and Bamfo, 1997). In this method, the cumulative residuals (the difference between the observed and predicted collisions for each value of a covariate) are plotted in increasing order for each covariate, e.g., AADT. Also plotted are the two standard deviation (2σ) boundaries. If there is no bias in the model, the plotting of cumulative residuals should stay inside these boundaries.

B. GOF tests for HSM models calibrated to local regions

GOF tests were conducted to calibrate HSM SPFs in 4SG intersections for Toronto and Edmonton data. The sample data are summarized in Table 2-2. The calculated overall measures are listed in Table 3-1. The CURE plots of the calibrated models for the major AADT covariates are shown in Figure 3-1.

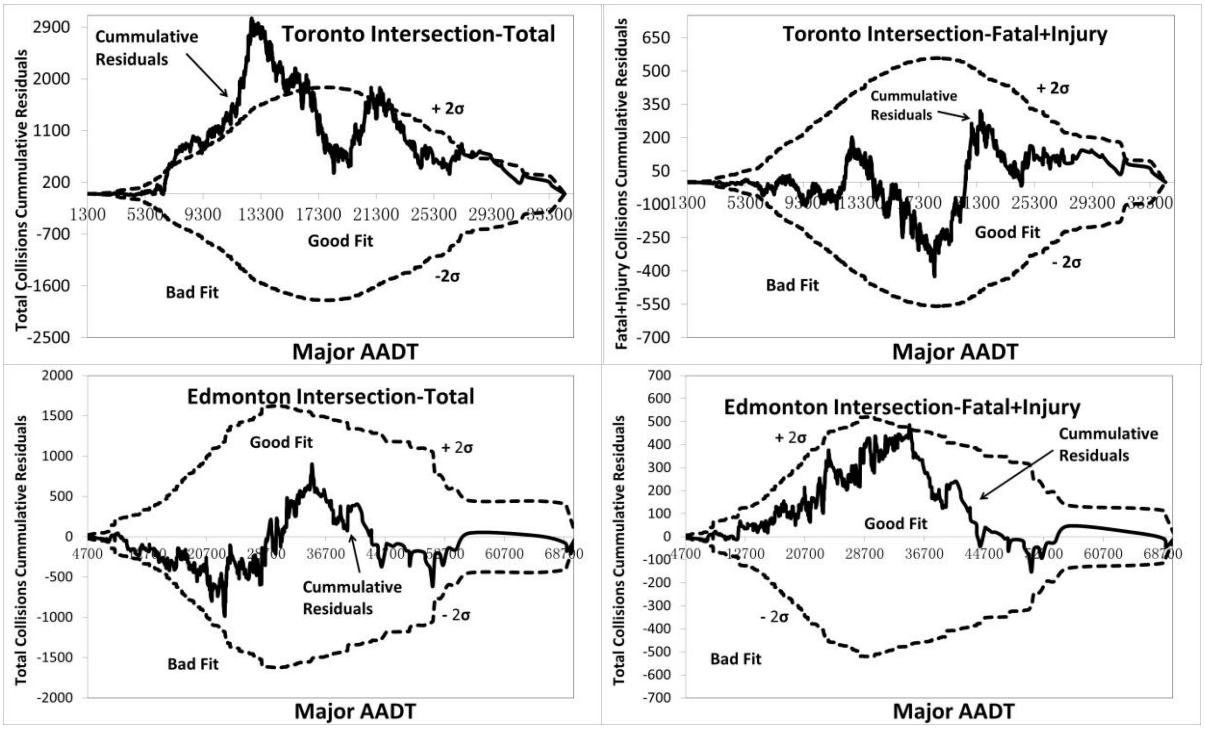
Table 3-1 GOF Measures of Calibrated HSM SPFs for 4SG

Region	Type of Collision[a]	Average Observed Collision	Pearson coefficient (r)	MPB	MAD	MSPE
Toronto	Total	10.34	0.61	-8.12	8.23	146.66
	Fatal/Injury	2.79	0.59	-2.08	2.14	11.29
Edmonton	Total	12.59	-0.01	-7.49	8.75	251.12
	Fatal/Injury	3.80	-0.01	-2.07	2.61	23.34

Note: a. Collision data pertain to multi-vehicle collisions only.

The calculated overall measures in Table 3-1 cannot be interpreted as consistent. While the Pearson coefficient (r) values of the Toronto calibrated models are relatively closer to 1, those of Edmonton are very far from 1, the signal of poor fit. The MAD values, on the other hand, are not sufficiently close to zero, also not a sign for good fit.

The CURE plots reveal extra bias of the calibrated models. For total collisions in Toronto, the cumulative residuals for lower major AADTs stray outside the boundaries (denoted by dashed lines), which indicated substantial bias. Aside from the fact that that the cumulative residuals of fatal plus injury (FI) collisions strayed slightly outside the boundaries, CURE plots for Edmonton calibrated models are generally valid. However, this cannot conceal the poor overall measures for Edmonton calibrated models indicated in Table 3-1.



Note: σ - standard deviation

Figure 3-1 CURE Plots of Calibrated HSM Models

All in all, none of the calibrated models listed above performed adequately with respect to both overall GOF measures and CURE plots. These results suggest that the predictive validity of a calibrated model may be inadequate, either for the overall measures, or over a sizable range of covariates.

3.1.4 Pros and Cons of Calibrated HSM Models as Source of Prior Knowledge

The above GOF tests indicate poor predictive performance of calibrated HSM models. This is likely due to the implicit, but likely incorrect, assumption that calibrated models have the same random structure and shape parameters as the original models developed for another jurisdiction. Another reason is that calibrating the models to data beyond the valid range of the original data could be problematic. For the original HSM (AASHTO, 2010) urban 4SG models, the valid range is 0 to 67,700 for major AADTs and 0 to 33,400 for minor AADTs. The AADTs for Edmonton used in this dissertation partially extend beyond these valid ranges.

Notwithstanding the poor performance of the GOF tests, calibrated HSM models still prevail in current practices, thanks to some of their merits. First, the HSM algorithm of a model for base conditions multiplied by collision modification factors (CMFs) is practically simple and straightforward to general practitioners. Secondly, HSM models were developed based on solid databases and lengthy scrutiny, so it makes sense that they have fundamental adaptability to many regions, with varying levels of success (e.g., GOF tests show calibrated HSM models adapt to Toronto better than Edmonton). Last but not least, many local jurisdictions do not have sufficient data or resources to develop their own models, and so are dependent on the HSM ones.

In summary, calibrated HSM models should still be taken into consideration. Meanwhile, jurisdictions should endeavour, as a top priority, to investigate locally developed SPFs.

3. 2 FRAMEWORK OF LOCAL MODEL DEVELOPMENT

The above section has addressed the necessity of local model development. The first question that then arises is whether local SPFs should follow the same structure of that in HSM models, that is, base conditions multiplied by a series of CMFs. To answer this question, a conceptual investigation of SPF classes, structures and relevant CMF implications is inevitable and should precede statistical efforts.

3.2.1 SPF Classes, Structures and CMFs

The three conventional types of SPF models are baseline models, general AADT models, and models with covariates (full models) (Lord et al., 2008b). Baseline and AADT-only models use AADTs as the only variable. The baseline model is developed based on base-condition datasets while the general AADT-only model is used for general (average) conditions of other potential variables. Both of these cannot be directly applied to specific sites without adjustments through the use of CMFs. In contrast, full models can, in principle, be directly applied to specific sites without the employment of CMFs.

Since the development of full models is still at the exploratory stage, the first version of the HSM (AASHTO, 2010) did not propose their use for the predictive methodology in Part C. Instead, the HSM has recommended crash prediction algorithms with base condition models multiplied by CMFs, as shown in Equation 1-1.

A CMF is a multiplicative factor used to reflect the expected changes in safety performance associated with the corresponding changes in highway design and/or the traffic control features (AASHTO, 2010). Reliable CMFs must be methodologically and statistically valid (Harkey et al., 2008). However many CMFs currently applied have been developed by ‘naïve’ before-after research studies and this has led to questionable results due to the failure to consider “regression to the mean” effects, and/or insufficient data (Sayed and de Leur, 2008). This has actually led to the omitting of many potential CMFs from the final HSM chapters due to overly high statistical

errors, as indicated in the Crash Modification Factors Clearinghouse website http://www.cmfclearinghouse.org (FHWA, 2009a).

Meanwhile, some current CMFs are can be complex can cannot be quickly processed. For example, the HSM CMF for the shoulder width of multi-lane highway segments is calibrated based on 3 sub-categories of shoulder widths, and 3 sub-categories of shoulder types as well as several sub-categories of AADTs. Besides that, adjacent sub-categories of shoulder widths are vaguely differentiated, so that similar shoulder widths may be estimated to have quite different CMFs.

More importantly, one key issue was not addressed by current CMF applications: whether CMFs are multiplicative. That is, is the effect of a CMF when it is applied alone different from its effect when applied with other CMFs? This issue can probably be addressed based on the approach to include CMF variables, along with other variables, all together into safety models and then investigate their correlations by a statistical method. This approach will be explored in subsequent sections.

Another unresolved key issue is whether a CMF is fixed or whether a crash modification *function* is more appropriate. This issue will also be addressed in subsequent next sections.

In short, conventional practice for development and application of CMFs is not ideal and it seems worthwhile to explore an alternative approach.

3.2.2 Full Model Structure with Collision Modification Functions

Regardless how a CMF is developed, it is just one single point estimation. It would be better if it captures, instead, how the effect varies as a function of one or more characteristics that influence the size of the effect. This can only be achieved by using a collision modification *function* (Elvik, 2009).

Studies on collision modification functions (denoted in this dissertation as "CM-Function") have been remarkably sparse so far. Gross et al. (2012a) analyzed safety effectiveness of converting signalized intersections to roundabouts and their analysis indicated that the safety benefit of roundabouts for total crashes decreased as traffic volumes increase, a result that led to the development of a crash modification function. Another paper of Gross et al. (2012b) raised issues associated with estimating the safety effects of multiple treatments, and argued that if multiple treatments are not independent and the CMFs are simply multiplied to estimate the combined effect, the result may be an over- or underestimation of the combined treatment effect. As a solution, they developed a framework for investigating interrelationships between treatments, and a matrix was provided to help identify potential overlapping effects. This series of works led by Gross investigated some key issues for current CMFs and explored preliminarily the CM-Functions with traffic volume as variable; however, the results were limited and didn't lead to well-established CM-functions. Elvik (2009) made some preliminary explorations and fitted points of existing CMFs related to certain factor values, and finally developed a regression curve for a CM-Function. One drawback of this method, however, is that the final regression curve could lose its validity in the light of the fact that some individual CMFs are not statistically significant.

Another plausible method is to develop CM-Functions directly from observed data. Unfortunately, a literature search on this topic yielded no outcomes. This is not surprising, since CM-Functions have been functionally included as part of some SPFs, specifically, full models with covariates. With a full model, the number of collisions, $N_{predicted}$, is predicted by:

$$N_{predicted} = \alpha(AADT)^{\beta_0} \times \exp(\beta_1 \cdot x_1) \times \exp(\beta_2 \cdot x_2) \times ... \times \exp(\beta_m \cdot x_m) \quad (3\text{-}6)$$

where,

$x_1, x_2, \ldots, x_m$ = covariates,

$\alpha, \beta_0, \beta_1, \beta_2, \ldots, \beta_m$ = estimated parameter coefficients, and

AADT = average annual daily traffic.

Then components $\exp(\beta_1 \cdot x_1),..., \exp(\beta_m \cdot x_m)$ are technically CM-Functions respectively for factors from x_1 to x_m.

Lord (2010) compared baseline models multiplied by CMFs versus full models with covariates and finally concluded that full models produced much less variance. This is evidence of the advantages of the use of a full model and supports advocating the full model as the local SPF structure. However, the full model contradicts the requirements of local SPFs in two aspects. First, local SPFs are supposed to parallel equivalent calibrated HSM models with baseline models multiplied by CMFs. The question is: Can a CM-Function component of a full model be easily matched with the relevant CMF of a calibrated HSM model? The answer is no. This is because the majority of CMFs are discrete numbers that have only limited values. For example, the HSM CMF for "Increase lane width from 11 feet to 12 feet" has only single value of 0.95 (AASHTO, 2010; FHWA, 2009a) without values given for lane width changes between 11 and 12 feet or for the dependence of this CMF on shoulder width. On the contrary, a CM-Function is generally one continuous expression. As a result, a mathematical approach to match one to another is actually much more problematic. For example, a CMF with 10 sub-categories might require 10 CM-Functions, which leads to the redundant "inflation" of a full model. More details will be provided in the following sections.

Secondly, the pattern of a full model fails to accommodate the data heterogeneity of local jurisdictions. A full model with all of its covariates reflects the collection of raw variables. Estimated full models from a specific jurisdiction may comprise some unique covariates not found in other models, including calibrated HSM models. The consequence is that such a model fails to safeguard transferability and also dampens merging possibilities with different models.

Furthermore, one major problem with the full model lies in the correlation of variables and the difficulty of modeling all interactions. Thus, full model may give better predictions but may not be so good for estimating the effect of a change in a design feature when designing a road, which is what the HSM predictive algorithm seeks to do. As a result, this dissertation is going to explore other alternative approaches like the multi-level model structure, which deals with interactive variables in more organized way. Also, this is the reason why this dissertation study

will retain the HSM algorithm, full models and multi-level ones all together and won't abandon either of them, since as stated above, each approach has its advantages and disadvantages.

3.2.3 Multi-level Model Structure with Collision Modification Functions

As described in Section 3.2.1, a locally developed SPF should not follow the pattern of a calibrated HSM model with a baseline model multiplied by CMFs. Instead, it should be capable of accommodating CM-Functions. The conventional pattern of the full model is further proven, as shown in Section 3.2.2, to be a less than ideal choice for local safety models.

As a result, a new paradigm in accordance with CM-Functions, other than the full model, should be taken into consideration and have the following three key properties:

1) compliance with CM-Functions,
2) does not "inflate" models, and
3) capability to specifically address locally specific data.

The nature of a single-level model is such that it will not satisfy these three properties. One viable alternative is a multi-level model which includes a first-level AADT-only SPF and some associated sub-models respectively with coefficients of a first-level model. Each sub-level model is associated with one or multiple impact factors, i.e., CM-Functions. With a sub-level hierarchy, the model will not inflate like the full model structure. More interestingly, it can accommodate locally specific data at a sub-level (Chin and Huang, 2008; Goldstein, 1999). Details of this aspect of the research work will be discussed in Section 3.5.

Nevertheless, the use of a full model as a local model option will not be entirely disregarded or eliminated. In Section 3.4, the development of full models will be discussed, while Section 3.5 will focus on multi-level models, followed by Section 3.6 which will provide comparisons and final recommendations.

3.3 DESCRIPTION OF SAMPLE DATA

The information on the data used to develop local full and multi-level models can be found in the datasets in Chapter 2. The Toronto and Edmonton 4SG intersections are described in Table 2-2, and four samples are taken from these two populations, respectively, which are summarized in Table 2-3.

There is evidence of data heterogeneity in these datasets. The Toronto data has a variable called "class" which categorizes intersections into 14 sub-groups, basically dependent on the functional classifications of cross roads (arterial, collector, local, or others). Edmonton does not have an equivalent variable but instead uses "area" to distinguish between urban and suburban environments.

3.4 LOCAL SPFS DEVELOPED WITH FULL MODEL STRUCTURE

3.4.1 HSM SPFs and CMFs for 4-legged Signalized Intersections

Based on Equation 1-1, and also subject to the sample facility and local datasets, a predictive model for multiple-vehicle collisions in urban 4SG intersections for Toronto and Edmonton from the HSM has the following form:

$$\begin{aligned} N_{mv} &= N_{bimv} \times (CMF_{lt} \times CMF_{rt}) \times C_x \\ &= \exp(a + b \times \ln(major\ AADT) + c \times \ln(minor AADT)) \times (CMF_{lt} \times CMF_{rt}) \times C_x \end{aligned} \quad (3\text{-}7)$$

where

N_{mv} = predicted average crash frequency for multiple-vehicle collisions,

N_{bimv} = predicted average crash frequency of base conditions for multiple-vehicle collisions,

Major AADT, minor AADT = AADTs for major and minor roads,

a, b, c = coefficients for SPFs,

CMF_{lt}, CMF_{rt} = CMFs for installation of left- and right-turn lanes, and

C_x = calibration factor for Toronto and Edmonton, respectively.

In Equation 3-7, the available CMFs for left- and right-turn lanes are defined as shown in Table 3-2 (AASHTO, 2010).

3.4.2 Full Models Developed to Match HSM Model

A. Potential components of full model

To match the SPFs in Equation 3-7 and CMFs in Table 3-2, the full model that will predict multi-vehicle collisions must consist of the following variables: major AADT, minor AADT, and number of approaches with turn lanes.

Table 3-2 HSM CMFs for Installation of Turn Lanes at 4SG Intersections

Type	Number of approaches with turn lanes			
	One	Two	Three	Four
Left-turn	0.90	0.81	0.73	0.66
Right-turn	0.96	0.92	0.88	0.85

However, there is the initial problem in that a simple continuous variable of "number of approaches with turning lanes" will not exactly match the relevant CMFs in Table 3-2 which are categorized into four subgroups for each type of turn. Instead, for each type of turn, the four variables have to been manipulated as follows.

For left turn lanes, the base line is no approaches with left turn lanes and the other variables are I_{1lt} (=1 for 1 approach with left turn lanes, 0 for others), I_{2lt} (=1 for 2 approaches with left turn lanes, 0 for others), I_{3lt} (=1 for 3 approaches with left turn lanes, 0 for others), I_{4lt} (=1 for 4 approaches with left turn lanes, 0 for others).

For right turn lanes, the base line is no approaches with right turn lanes, while the other variables are I_{1rlt} (=1 for 1 approach with right turn lanes, 0 for others), I_{2rt} (=1 for 2 approaches with right turn lanes, 0 for others), I_{3rt} (=1 for 3 approaches with right turn lanes, 0 for others), I_{4rt} (= 1 for 4 approaches with right turn lanes, 0 for others).

Except for these common components, there are two more locally specific variables, i.e., class for Toronto and area for Edmonton.

There are three more reasons why this dissertation selected categorical explanatory variables – number of approaches with turning lanes – for full models rather than continuous ones:

First, full models were historically developed but some variables were insignificant and some others had coefficients against well-recognized safety effects (e.g., negative coefficients for access control and lighting), so CMFs were applied instead (Harwood et al., 2000).

Second, the ultimate goal of all locally developed models in this study is to be integrated with the HSM model shown as Equation 3-7. A full model with continuous explanatory variables cannot be integrated with HSM model. More details are covered in Chapter 6.

Third, this dissertation research did attempt models with continuous explanatory variables, but they failed to yield meaningful results, being either statistically insignificant or contrary in indications to well-recognized safety effects. More details can be found at Section 3.6.3.

In conclusion, for the above-mentioned reasons, this dissertation study chose categorical variables for number of turning lanes, and abandoned continuous variables.

B. Full model identification and statistical platform

For the collision data, which are non-negative integers with over-dispersion, the common model form is well described by a mixed Poisson distribution family as follows (Persaud et al., 2010a; Chou and Steenhard, 2009).

The number of collisions Y_{it} for a particular i_{th} site and time period t is Poisson distributed about its mean μ_{it}:

$$Y_{it}|\mu_{it} \sim Po(\mu_{it}) \text{ i = 1, 2, . . . , I and t = 1, 2, . . . ,T} \quad (3\text{-}8)$$

The mean of the Poisson distribution is structured as:

$$\mu_{it} = \exp(x_i\beta + e) = f(X;\beta)\cdot\varepsilon_i \quad (3\text{-}9)$$

where

f(.) = function of the covariates (X or x_i),

β = vector of unknown coefficients, and

e, ε_i = random error terms.

Finally, the precise form of the distribution of the safety model depends on the specific choice of probability distribution of the error term ε_i. Among these distributions, a special case of the Poisson-gamma distribution as Gamma (φ, (1/ φ)) and φ~ Gamma (1,1), also known as the NB distribution, is common in road safety analysis (Lan et al., 2009) and was chosen to develop the full models. Its probability density function (PDF) is given (Lord et al., 2010) as:

$$f(y_{it};\alpha,\mu_{it}) = \frac{\Gamma(y_{it}+a^{-1})}{\Gamma(a^{-1})y_{it}!}\left(\frac{\alpha^{-1}}{\mu_{it}+a^{-1}}\right)^{a^{-1}}\left(\frac{\mu_{it}}{\mu_{it}+\alpha^{-1}}\right)^{y_{it}} \quad (3\text{-}10)$$

where

y_{it} = response variable for observation i and time period t,

μ_{it} = mean response for observation i and time period t, and

α = dispersion parameter of Poisson-Gamma distribution.

SAS statistical software (SAS Institute Inc., 2012) was chosen as the modeling platform, and its "NLMIXED" procedure that applies a non-linear mixed MLE was selected to obtain fitted models for this study.

C. Estimation Results for Full Models

The NLMIXED procedure was run for iterative optimization that would obtain MLE model coefficients for all model parameters, with its significance relying on the final outcomes of the iterations. The rules for significant estimation are stipulated as (SAS Institute Inc., 2012):

- model optimization, as a whole, is finally convergent (output of iteration history finally

displayed as "GCONV convergence criterion satisfied", that is, the relative gradient convergence, GCONV<= 1E-8),

- all parameter estimates are significant at the 10% level judged by the p-values of the t-test, and
- all parameter estimates are finally stable (i.e., each parameter estimate yields a very small gradient which reflects the improvement of optimization, up to the last step of the iteration; stability is achieved when the gradient is less than 0.0001).

The preliminary full models estimated for the Toronto data are summarized in Tables 3-3 to 3-5.

They are not final models yet since the insignificant variables need to be removed and models subsequently re-fitted.

Table 3-3 Preliminary Estimation Results of Full Model for Toronto Reference Population

Parameter[a]	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant (with 90% confidence level[b])?
ln(β_0)	-6.7627	-19.33	<.0001	0.000169	YES
major AADT (β_1)	0.6435	21.19	<.0001	0.001918	YES
minor AADT (β_2)	0.4525	23.14	<.0001	0.002522	YES
I1lt (β_3)	-0.148	-3.67	0.0002	-0.0006	YES
I2lt (β_4)	-0.2121	-6.19	<.0001	-0.00022	YES
I3lt (β_5)	-0.3063	-6.5	<.0001	0.000137	YES
I4lt (β_6)	-0.2196	-5	<.0001	0.000513	YES
I1rt (β_7)	-0.06364	-2.11	0.0347	-0.00059	YES
I2rt (β_8)	-0.09526	-2.5	0.0125	0.00054	YES
I3rt (β_9)	-0.0578	-0.87	0.3828	0.000106	NO
I4rt (β_{10})	0.005848	0.08	0.9374	0.000445	NO
class (β_{11})	-0.07431	-12.23	<.0001	-0.00077	YES
dispersion parameter, α	0.1995				

Note: a. See definitions for all parameters from "A. Potential components of full model" of Section 3.4.2. These definitions apply for all rest tables in this chapter.

Note: b. This confidence level applied for all rest tables in this chapter.

Table 3-4 Preliminary Estimation Results of Full Model for Toronto Sample (Size=680)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln(\beta_0)$	-7.4401	-13.6	<.0001	-0.00087	YES
major AADT (β_1)	0.6734	14.71	<.0001	-0.00823	YES
minor AADT (β_2)	0.4877	15.84	<.0001	-0.0072	YES
I1lt (β_3)	-0.1026	-1.53	0.1258[a]	-0.00003	YES
I2lt (β_4)	-0.2384	-4.25	<.0001	-0.00053	YES
I3lt (β_5)	-0.2521	-3.26	0.0012	0.0001	YES
I4lt (β_6)	-0.1894	-2.63	0.0087	-0.00033	YES
I1rt (β_7)	-0.04141	-0.82	0.4153	-0.00015	NO
I2rt (β_8)	-0.1576	-2.46	0.0143	0.000843	YES
I3rt (β_9)	-0.1404	-1.38	0.1688	0.000353	NO
I4rt (β_{10})	-0.04799	-0.41	0.68	-0.0006	NO
class (β_{11})	-0.06467	-6.73	<.0001	-0.00507	YES
dispersion parameter, α	0.2317				

Note: a. Since p-value of β_3 is very close to 10%, it could be verified as significance.

Although whole convergence and stability of parameter estimations in all of the Toronto modeling were achieved, not all of the parameters satisfied the criteria of statistical significance as determined by the p-values. As a result, only partial covariates could be included in the final models. Further estimations must be carried out by eliminating all of the insignificant variables and re-running the estimation procedure.

The final estimated models are shown in Equations 3-11 to 3-13, with the final estimation results listed in Tables 3-6 to 3-8. The final model for the Toronto reference population has 9 significant variables (except for the intercept) while the model for a sample size of 680 has 8 variables, and model for the sample size of 588 has 9 variables, which are different from those of the reference population.

Due to the different variables removed, the final models won't have consistent components. This disadvantage will be further discussed later in this section.

Table 3-5 Preliminary Estimation Results of Full Model for Toronto Sample (Size=588)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln(\beta_0)$	-7.1059	-11.89	<.0001	0.00024	YES
major AADT (β_1)	0.6509	13.07	<.0001	0.002255	YES
minor AADT (β_2)	0.4756	13.59	<.0001	0.003133	YES
I1lt (β_3)	-0.1199	-1.66	0.0977	0.000043	YES
I2lt (β_4)	-0.156	-2.6	0.0095	-0.00035	YES
I3lt (β_5)	-0.1154	-1.36	0.1741	0.000316	NO
I4lt (β_6)	-0.1371	-1.8	0.0718	0.000125	YES
I1rt (β_7)	-0.139	-2.55	0.011	-0.00006	YES
I2rt (β_8)	-0.1948	-2.83	0.0048	0.000217	YES
I3rt (β_9)	-0.2318	-2.05	0.041	0.000105	YES
I4rt (β_{10})	-0.02768	-0.22	0.8262	0.000018	NO
class (β_{11})	-0.06761	-6.2	<.0001	-0.0028	YES
dispersion parameter, α	0.2308				

Table 3-6 Final Estimation Results of Full Model for Toronto Reference Population

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln(\beta_0)$	-6.7563	-19.34	<.0001	-6.7563	YES
major AADT (β_1)	0.6421	21.19	<.0001	0.6421	YES
minor AADT (β_2)	0.4527	23.18	<.0001	0.4527	YES
I1lt (β_3)	-0.1484	-3.68	0.0002	-0.1484	YES
I2lt (β_4)	-0.214	-6.26	<.0001	-0.214	YES
I3lt (β_5)	-0.3108	-6.66	<.0001	-0.3108	YES
I4lt (β_6)	-0.2252	-5.34	<.0001	-0.2252	YES
I1rt (β_7)	-0.05946	-2.04	0.0417	-0.05946	YES
I2rt (β_8)	-0.08932	-2.45	0.0145	-0.08932	YES
class (β_{11})	-0.07379	-12.21	<.0001	-0.07379	YES
dispersion parameter, α	0.1996				

For the Toronto reference population, the final model is:

$$\mu_{it} = \beta_0 \cdot (majorAADT)^{\beta_1} \cdot (\min orAADT)^{\beta_2} \cdot \exp(\beta_3 \cdot I_{1lt}) \cdot \exp(\beta_4 \cdot I_{2lt}) \cdot \exp(\beta_5 \cdot I_{3lt}) \cdot \exp(\beta_6 \cdot I_{4lt}) \cdot \exp(\beta_7 \cdot I_{1rt}) \cdot \exp(\beta_8 \cdot I_{2rt}) \cdot \exp(\beta_{11} \cdot class) \quad (3\text{-}11)$$

While the Toronto sample size of 680 has a final model as:

$$\mu_{it} = \beta_0 \cdot (majorAADT)^{\beta_1} \cdot (\min orAADT)^{\beta_2} \cdot \exp(\beta_3 \cdot I_{1lt}) \cdot \exp(\beta_4 \cdot I_{2lt}) \cdot \exp(\beta_5 \cdot I_{3lt}) \cdot \exp(\beta_6 \cdot I_{4lt}) \cdot \exp(\beta_8 \cdot I_{2rt}) \cdot \exp(\beta_{11} \cdot class) \quad (3\text{-}12)$$

Table 3-7 Final Estimation Results of Full Model for Toronto Sample (Size=680)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-7.4175	-13.53	<.0001	0.000264	YES
major AADT (β_1)	0.6684	14.61	<.0001	0.00228	YES
minor AADT (β_2)	0.4875	15.81	<.0001	0.001824	YES
I1lt (β_3)	-0.1102	-1.66	0.0966	0.000054	YES
I2lt (β_4)	-0.2499	-4.51	<.0001	0.000144	YES
I3lt (β_5)	-0.2759	-3.68	0.0002	-0.00005	YES
I4lt (β_6)	-0.2139	-3.14	0.0018	0.000055	YES
I2rt (β_8)	-0.126	-2.14	0.0326	3.01E-06	YES
class (β_{11})	-0.06264	-6.57	<.0001	0.002258	YES
dispersion parameter, α	0.2324				

Table 3-8 Final Estimation Results of Full Model for Toronto Sample (Size=588)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-6.8695	-11.9	<.0001	-0.00324	YES
major AADT (β_1)	0.6249	12.96	<.0001	-0.02878	YES
minor AADT (β_2)	0.4669	14.02	<.0001	-0.0239	YES
I2lt (β_4)	-0.07446	-1.61	0.1083	0.000898	YES
I1rt (β_7)	-0.1556	-2.97	0.0031	0.001638	YES
I2rt (β_8)	-0.2181	-3.34	0.0009	-0.00054	YES
I3rt (β_9)	-0.2628	-2.4	0.0167	0.000269	YES
class (β_{11})	-0.06733	-6.17	<.0001	-0.02043	YES
dispersion parameter, α	0.2328				

For the sample size of 588 in the Toronto case, and after multiple rounds of eliminating insignificant variables, the final model is:

$$\mu_{it} = \beta_0 \cdot (majorAADT)^{\beta_1} \cdot (\min orAADT)^{\beta_2} \cdot \exp(\beta_4 \cdot I_{2lt}) \cdot \cdot \exp(\beta_7 \cdot I_{1rt}) \cdot \exp(\beta_8 \cdot I_{2rt}) \cdot \exp(\beta_9 \cdot I_{3rt}) \cdot \exp(\beta_{11} \cdot class) \quad (3\text{-}13)$$

The preliminary results of full model estimations based on the Edmonton population and samples are summarized in Tables 3-9 to 3-11.

Table 3-9 Preliminary Estimation Results of Full Model for Edmonton Population

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-10.9303	-12.99	<.0001	0.00031	YES
major AADT (β_1)	0.7147	9.03	<.0001	0.002934	YES
minor AADT (β_2)	0.611	14.4	<.0001	0.002795	YES
I1lt (β_3)	0.1742	1.38	0.1679	0.000038	NO
I2lt (β_4)	-0.05545	-0.52	0.6049	0.000085	NO
I3lt (β_5)	-0.1091	-0.76	0.4494	0.000139	NO
I4lt (β_6)	-0.119	-0.91	0.3621	0.000139	NO
I1rt (β_7)	-0.03173	-0.31	0.7592	-0.00004	NO
I2rt (β_8)	0.1279	1.35	0.1783	-0.0002	NO
I3rt (β_9)	0.1074	0.56	0.576	0.00031	NO
I4rt (β_{10})	0.4728	3.45	0.0006	0.000066	YES
area (β_{11})	0.3383	4.78	<.0001	0.000533	YES
dispersion parameter, α	0.5221				

Table 3-10 Preliminary Estimation Results of Full Model for Edmonton Sample (Size=400)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-11.1054	-11.5	<.0001	-0.00747	YES
major AADT (β_1)	0.7502	8.11	<.0001	-0.26977	YES
minor AADT (β_2)	0.5788	12.32	<.0001	-0.17501	YES
I1lt (β_3)	0.3123	2.21	0.0278	0.354887	YES
I2lt (β_4)	0.05484	0.45	0.6527	0.158601	NO
I3lt (β_5)	-0.07746	-0.48	0.6319	0.250461	NO
I4lt (β_6)	-0.0105	-0.07	0.9426	-0.02142	NO
I1rt (β_7)	0.1102	0.95	0.3404	-0.36874	NO
I2rt (β_8)	0.1526	1.46	0.1452	-0.23619	NO
I3rt (β_9)	0.108	0.52	0.6041	-0.36343	NO
I4rt (β_{10})	0.5426	3.64	0.0003	-0.07337	YES
area (β_{11})	0.2977	3.78	0.0002	-0.0832	YES
dispersion parameter, α	0.5026				

Although whole convergence and stability of parameter estimations were achieved in all of the Edmonton modeling, not all the parameters satisfied the criteria of statistical significance as determined by the p-values. As result, only partial covariates could be included in the final models. Further estimations must been carried out by eliminating all of the insignificant variables and re-running the estimations.

Table 3-11 Preliminary Estimation Results of Full Model for Edmonton Sample (Size=300)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-11.9122	-10.55	<.0001	0.000036	YES
major AADT (β_1)	0.8314	7.85	<.0001	0.000375	YES
minor AADT (β_2)	0.602	10.71	<.0001	0.000316	YES
I1lt (β_3)	0.04705	0.29	0.7688	-7.14E-06	NO
I2lt (β_4)	-0.1437	-1.07	0.2858	0.000028	NO
I3lt (β_5)	-0.2419	-1.32	0.1863	-2.44E-06	NO
I4lt (β_6)	-0.2025	-1.21	0.2271	6.61E-06	NO
I1rt (β_7)	-0.08107	-0.59	0.5539	-0.00003	NO
I2rt (β_8)	0.07555	0.61	0.541	0.000029	NO
I3rt (β_9)	0.01867	0.07	0.9426	-4.84E-06	NO
I4rt (β_{10})	0.3918	2.16	0.0316	4.16E-06	YES
area (β_{11})	0.3466	3.75	0.0002	0.000081	YES
dispersion parameter, α	0.5197				

The final estimated models are shown in Equations 3-14 to 3-16, with final estimation results listed in Tables 3-12 to 3-14. The final model for the Edmonton population and that with a sample size of 300 have 4 significant variables (except for the intercept) while the model with a sample size of 400 has 5 variables.

Table 3-12 Final Estimation Results of Full Model for Edmonton Population

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-11.1061	-13.97	<.0001	-0.00005	YES
major AADT (β_1)	0.7402	9.8	<.0001	-0.00047	YES
minor AADT (β_2)	0.6008	15.17	<.0001	-0.00056	YES
I4rt (β_{10})	0.3271	3.33	0.0009	-0.00004	YES
area (β_{11})	0.3655	5.29	<.0001	-0.00008	YES
dispersion parameter, α	0.5326				

The final model form for the Edmonton population and that with a sample size of 300 is:

$$\mu_{it} = \beta_0 \cdot (majorAADT)^{\beta_1} \cdot (\min orAADT)^{\beta_2} \cdot \exp(\beta_{10} \cdot I_{4rt}) \cdot \exp(\beta_{11} \cdot area) \quad (3\text{-}14)$$

While the model with a sample size of 400 is:

$$\mu_{it} = \beta_0 \cdot (majorAADT)^{\beta_1} \cdot (\min orAADT)^{\beta_2} \cdot \exp(\beta_3 \cdot I_{1lt}) \cdot \exp(\beta_{10} \cdot I_{4rt}) \cdot \exp(\beta_{11} \cdot area) \quad (3\text{-}15)$$

Table 3-13 Final Estimation Results of Full Model for Edmonton Sample (Size=400)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-11.533	-12.91	<.0001	0.000096	YES
major AADT (β_1)	0.7831	9.02	<.0001	0.001843	YES
minor AADT (β_2)	0.5959	13.48	<.0001	0.000102	YES
I1lt (β_3)	0.3105	2.91	0.0038	0.000548	YES
I4rt (β_{10})	0.414	3.79	0.0002	-0.00155	YES
area (β_{11})	0.327	4.23	<.0001	0.000168	YES
dispersion parameter, α	0.5081				

Table 3-14 Final Estimation Results of Full Model for Edmonton Sample (Size=300)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-11.7452	-11.13	<.0001	-0.00074	YES
major AADT (β_1)	0.8323	8.29	<.0001	-0.0081	YES
minor AADT (β_2)	0.569	11.36	<.0001	-0.00696	YES
I4rt (β_{10})	0.2967	2.26	0.0246	-0.00022	YES
area (β_{11})	0.3665	4.06	<.0001	-0.00068	YES
dispersion parameter, α	0.5291				

D. Assessments of Full Model Estimation Results

The full models for Toronto and Edmonton demonstrate considerable heterogeneity. Besides the original locally specific variables – class for Toronto and area for Edmonton, the models still contain different components. Neither has a configuration that is completely equivalent to Equation 3-8, the HSM baseline SPF multiplied by CMFs.

None of these full models is a genuine "full" model. They are actually "partial" models in various ways. Equations 3-11 to 3-13 for Toronto and Equations 3-14 to 3-16 for Edmonton revealed that all of these models have some original variables removed so that each of them loses some predictors (e.g., the final full model for Toronto reference population does not have the variables "I3rt" and "I4rt", which means, for example, this model will generate same prediction for one site of three approaches with right turn lanes and another site of four approaches with

right turn lanes). Moreover, these models have different components, which mean that they won't convey consistent information.

Therefore, a new paradigm needs to be considered as a solution to these problems. This alternative solution should be structurally homogeneous, or at least, able to restructure heterogeneity in a controlled manner. The above estimated single-level models are apparently unable to do so. Consequently, the study will turn its focus to multi-level models.

3.5 LOCAL SPFS DEVELOPED WITH MULTI-LEVEL MODEL STRUCTURE

3.5.1 Basic Concepts

Increasingly more road safety analysts recognize that road safety data are naturally constructed in a multi-level manner. Correspondingly, more efforts have been put forth on models or methods that address multiple data structures.

Some researchers have employed artificial intelligence models (AI), such as neural networks (NNs) or Bayesian NNs to model multilevel data structure. However, a shortcoming of these approaches is the inability to generate explicit functional relationships and statistically interpretable results (Chin and Huang, 2008).

Another approach is to use generalized estimating equations (GEE) as an extension of the generalized linear model (GLM). When dealing with multilevel data structures, the GEE aims to provide estimates with acceptable properties only for the fixed parameters in the model, and considers the existence of any other random parameters as a necessary "nuisance". Hence, the GEE may merely be superior in cases where the exact form of the multilevel data structure is unknown (Chin and Huang, 2008).

Another way to distinctly address a multilevel data structure is to develop multi-level models.

Multi-level modeling, also called hierarchical or random effect modeling, is a statistical technique that allows multilevel data structures to be properly specified and estimated. Multi-level modeling is defined as "a regression (a linear or generalized linear model) in which the parameters, i.e., the regression coefficients, are given in a probability model" (Chin and Huang, 2008; Goldstein, 1999).

3.5.2 Relevant Past Research

Chin and Huang (2008) developed a two-level hierarchical GLM road safety model. A general expression of the statistical modeling in Equations 3-8 and 3-9 is still effective for multi-level modeling. However, within a multi-level model, the covariate vector X is divided into three components, c(1, X^{L1}, X^{L2}) to represent the factors associated with the intercept, Level 1 (individual level) and Level 2 (group level), respectively. Correspondingly, **β** and **ε** are also divided into different components to serve different functions. Hence, the link function becomes a combination of the models in terms of the two levels as shown in Equation 3-9.

$$\begin{aligned} &\text{Level 1 model:} \quad f^{-1}(\theta) = \beta_0^{L1} + X^{L1}\beta^{L1} + \varepsilon^{L1} \\ &\text{Level 2 model:} \quad \beta_0^{L1} = \beta_{00}^{L2} + X^{L2}\beta_0{}^{L2} + \varepsilon_0{}^{L2} \\ &\qquad\qquad \beta^{L1} = \beta_{01}^{L2} + X^{L2}\beta_1{}^{L2} + \varepsilon_1{}^{L2} \end{aligned} \tag{3-16}$$

The combined model is obtained by substituting the Level 2 model into the Level 1 model,

$$f^{-1}(\theta) = (\beta_{00}^{L2} + X^{L1}\beta_{01}{}^{L2} + X^{L2}\beta_0{}^{L2} + X^{L1}X^{L2}\beta_1{}^{L2}) + (\varepsilon^{L1} + \varepsilon_0{}^{L2} + X^{L1}\varepsilon_1{}^{L2}) \tag{3-17}$$

The link function now consists of two parts: a fixed and a random part. The fixed part represents a deterministic relationship which is fully dependent on covariate X, while the random part is stochastically determined by the number of disturbance terms.

Some other road safety researchers have also tried structural modeling or similarly complicated approaches (Lee, Chung and Son, 2008; Davis, 2004) that have mainly addressed the hierarchical nature of their own local databases or dealt with model bias.

There are few satisfactory outcomes from the existing cases in the field of road safety that have used multi-level modeling. Therefore, structural or relevant multi-level methodologies still have much unrealized potential in the area of SPF development.

3.5.3 Multi-level Models Identified with Collision Modification Functions

Full model SPFs developed in Section 3.4 can be functionally restructured as a combination of a base SPF and sub-level CM-Functions where the first level model has an AADT-only SPF, while sub-level models are functions that define various coefficients in the first level model, as shown in Equation 3-18.

First-level SPF: $$N_{mv} = \exp(a + B \times \ln(AADT_{maj}) + C \times \ln(AADT_{\min})) = A(AADT_{maj})^{B}(AADT_{\min})^{C}$$

Sub-level CM-Functions:

$$\begin{aligned} A &= \alpha_0 \cdot \exp(\alpha_1 \cdot LT) \cdot \exp(\alpha_2 \cdot RT) \\ B &= \beta_0 \cdot \exp(\beta_1 \cdot local-factor) \\ C &= \gamma_0 \cdot \exp(\gamma_1 \cdot local-factor) \end{aligned} \quad (3\text{-}18)$$

where

A, B, C are parameter coefficients estimated for first-level AADT-only modeling,

α_0, α_1, α_2, β_0, β_1, γ_0, γ_1 are covariate coefficients estimated in CM-Functions,

LT = number of approaches with left-turn lanes,

RT = number of approaches with right-turn lanes, and

"local-factor" is class for Toronto, area for Edmonton.

Equation 3-18 resembles the locally specific covariates, that is, class for Toronto and area for Edmonton, placed into the associated sub-level modeling to shape the parameters of the AADT variables for first-level SPFs. This maintains homogeneity of the first-level modeling while addressing local specifics at a lower level. By this means, the multi-level nature of safety data is addressed. In addition, it reshuffles the components of other covariates by integrating them into the intercept (constant) of first-level SPFs. Since the hierarchy is utterly different to that of the

baseline model multiplied by a series of CMFs shown in Equation 3-7, there is no need to match individual CMFs with multiple variables for specific values of turning lanes. In contrast, a single variable for number of turning lanes was applied. This further simplified the model structure.

Here, the previously mentioned special case of the Poisson-gamma (or NB) distribution for full models in Section 3.4.2 is replicated for the multi-level models. Also the "NLMIXED" procedure of the SAS was re-introduced as a tool to fit non-linear mixed estimations. Finally, the rules for estimation of significance are the same as those shown in Section 3.4.2.

3.5.4 Estimation Results of Multi-level Models

The estimation results of the multi-level models were obtained with similar statistical estimating procedures as used for the full models. The preliminary multi-level models of the Toronto reference population and samples are listed in Tables 3-15 to 3-17.

With the same standard applied to full model estimations, all of the Toronto multi-level models, except for one parameter in Table 3-16, have thorough statistical significance.

Table 3-15 Estimations of Multi-level Model for Toronto Reference Population

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-7.3137	-22.41	<.0001	-0.00058	YES
LT (α_1)	-0.063	-6.04	<.0001	-0.00116	YES
RT (α_2)	-0.02344	-1.76	0.0781	0.000091	YES
β_0	0.5668	13.41	<.0001	-0.00558	YES
class at B (β_1)	0.005428	1.67	0.0959	-0.04109	YES
γ_0	0.6005	16.33	<.0001	-0.00467	YES
class at C (γ_1)	-0.01568	-4.11	<.0001	-0.03262	YES
dispersion parameter, α	0.2012				

The multi-level models of the Toronto reference population with sample size of 588 have a consistent model form as shown in Equation 3-18 and the above estimation tables are already their final results. However, the multi-level model for a sample size of 680 is not significant due to the coefficient β_1.

Table 3-16 Preliminary Estimations of Multi-level Model for Toronto Sample (Size=680)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-7.976	-15.89	<.0001	-0.00377	YES
LT (α_1)	-0.05848	-3.46	0.0006	0.00074	YES
RT (α_2)	-0.03863	-1.78	0.0753	0.00007	YES
β_0	0.5967	9.56	<.0001	-0.03684	YES
class at B (β_1)	0.006556	1.35	0.1773	-0.21536	NO
γ_0	0.6344	11.55	<.0001	-0.03545	YES
class at C (γ_1)	-0.01583	-2.77	0.0058	-0.21515	YES
dispersion parameter, α	0.2346				

Table 3-17 Estimations of Multi-level Model for Toronto Sample (Size=588)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-7.7264	-14.35	<.0001	-0.00674	YES
LT (α_1)	-0.03671	-2.03	0.0427	-0.01807	YES
RT (α_2)	-0.05399	-2.34	0.0198	-0.00988	YES
β_0	0.4881	6.64	<.0001	-0.05996	YES
class at B (β_1)	0.0153	2.52	0.012	-0.45838	YES
γ_0	0.7229	10.27	<.0001	-0.05742	YES
class at C (γ_1)	-0.02561	-3.69	0.0002	-0.46215	YES
dispersion parameter, α	0.2305				

Since a population-based model is significant, we already have a sound, secured, successful model. As for sample-based model that is insignificant, the solution could be to regenerate another sample through an iterative procedure of re-sampling (see Chapter 2) and re-estimations. After several rounds of iterative attempts, one new sample with a size of 680 resulted in a multi-level model that is significant, replacing the previous sample. The new estimations based on the new sample, i.e., the final results for a sample size of 680, are shown in Table 3-18.

Table 3-18 Final Estimations of Multi-level Model for Toronto Sample (Size=680)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-8.064	-15.67	<.0001	0.008477	YES
LT (α_1)	-0.06924	-4.11	<.0001	0.035436	YES
RT (α_2)	-0.03515	-1.62	0.1064	0.022218	YES
β_0	0.6068	9.31	<.0001	0.08131	YES
class at B (β_1)	0.008632	1.71	0.0872	0.119551	YES
γ_0	0.6388	11.25	<.0001	0.097065	YES
class at C (γ_1)	-0.01845	-3.1	0.002	0.312344	YES
dispersion parameter, α	0.2263				

The summarized statistics of the new sample are listed in Table 3-19.

Table 3-19 Summarized Statistics of New Toronto Sample (size=680)

Dataset (sample size)	Variable	Minimum Value	Maximum Value	Mean (Standard Deviation)
4-leg, signalized at-grade intersections from Toronto, Ontario, Canada (680)	Multi-vehicle total collisions	0	370	62.2 (60.5)
	Multi-vehicle injury collisions	0	102	16.7 (17.6)
	Years	6	6	6 (0)
	Major AADT	1322	33504	13203 (5889)
	Minor AADT	42	27936	4129 (4132)
	No. of left-turn lanes	0-171; 1-97; 2-201; 3-71; 4-140		
	No. of right-turn lanes	0-379; 1-160; 2-92; 3-28; 4-21		
	Class	1-2; 2-22; 3-84; 4-6; 5-98; 6-64; 7-1; 8-131; 9-72; 10-23; 11-0; 12-115; 13-57; 14-5		

Multi-level model estimations for Edmonton are listed in Tables 3-20 to 3-22.

Table 3-20 Estimations of Multi-level Model for Edmonton Population

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-10.6091	-13.03	<.0001	0.014161	YES
LT (α_1)	-0.04553	-1.54	0.1253[a]	0.040521	YES
RT (α_2)	0.104	3.34	0.0009	0.051578	YES
β_0	0.9238	7.32	<.0001	0.142353	YES
area at B (β_1)	-0.1324	-1.97	0.0495	0.270247	YES
γ_0	0.3399	3.01	0.0028	0.129502	YES
area at C (γ_1)	0.1881	2.49	0.0131	0.245478	YES
dispersion parameter, α	0.5236				

Note: a. Since p-value of α_1 is very close to 10%, it could be verified as significance.

Table 3-21 Preliminary Estimations of Multi-level Model for Edmonton Sample(Size=400)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-10.993	-11.77	<.0001	0.001836	YES
LT (α_1)	-0.03766	-1.12	0.2622	0.005135	NO
RT (α_2)	0.1132	3.29	0.0011	0.004139	YES
β_0	0.9774	6.88	<.0001	0.019373	YES
area at B (β_1)	-0.1323	-1.75	0.0801	0.035033	YES
γ_0	0.3196	2.51	0.0123	0.014749	YES
area at C (γ_1)	0.1856	2.19	0.0291	0.026112	YES
dispersion parameter, α	0.5106				

Table 3-22 Preliminary Estimations of Multi-level Model for Edmonton Sample(Size=300)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-11.2384	-10.4	<.0001	-0.00105	YES
LT (α_1)	-0.04613	-1.23	0.219	-0.00177	NO
RT (α_2)	0.08509	2.11	0.036	-0.00219	YES
β_0	1.0934	7.04	<.0001	-0.01097	YES
area at B (β_1)	-0.2021	-2.3	0.0224	-0.02929	YES
γ_0	0.2211	1.53	0.1263[a]	-0.00827	YES
area at C (γ_1)	0.2672	2.71	0.0072	-0.02306	YES
dispersion parameter, α	0.5139				

Note: a. Since p-value of γ_0 is very close to 10%, it could be verified as significance.

With the same standard applied to full model estimations, the Edmonton population-based multi-level model is thoroughly statistically significant. The two sample-based models each have one insignificant variable. The Edmonton population-based multi-level model has a consistent model form as shown in Equation 3-18 and the above estimation table is already its final result.

Since the population-based model is significant, we already have a sound, secured, successful model. As for sample-based models that are insignificant, the solution is still regenerating another sample, as was done for Toronto, through an iterative procedure of re-sampling and re-estimations. After several rounds of iterations, one new sample with a size of 400 resulted in a significant multi-level model. Hence, it replaces the previous sample. Similar attempts resulted in another new sample with a size of 300 which has also led to a significant multi-level model. The model estimations based on the new samples are shown in Tables 3-23 and 3-25. The summarized statistics of the new samples are listed in Tables 3-24 and 3-26.

Table 3-23 Final Estimations of Multi-level Model for Edmonton Sample (Size=400)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-10.9139	-11.24	<.0001	-0.00441	YES
LT (α_1)	-0.05555	-1.62	0.1052[a]	-0.01754	YES
RT (α_2)	0.09313	2.53	0.0117	-0.01099	YES
β_0	1.0701	7.11	<.0001	-0.04553	YES
area at B (β_1)	-0.239	-2.83	0.005	-0.07292	YES
γ_0	0.2111	1.58	0.1155	-0.03851	YES
area at C (γ_1)	0.3099	3.26	0.0012	-0.06112	YES
dispersion parameter, α	0.5529				

Note: a. Since p-value of α_1 is very close to 10%, it could be verified as significance.

Table 3-24 Summarized Statistics of New Edmonton Sample (Size=400)

Dataset (sample size)	Variable	Minimum Value	Maximum Value	Mean (Standard Deviation)
4-leg, signalized at-grade intersections from Edmonton, Alberta, Canada (400); (386 with data on turn lanes)	Multi-vehicle total collisions	0	555	75.3 (84.6)
	Multi-vehicle injury collisions	0	195	23.0 (26.5)
	Years	6	6	6 (0)
	Major AADT	4720	70331	24615 (10769)
	Minor AADT	171	30506	9434 (6774)
	No. of left-turn lanes	0-71; 1-57; 2-123; 3-35; 4-100		
	No. of right-turn lanes	0-130; 1-72; 2-110; 3-15; 4-59		
	Class	urban-195; suburban-205		

Table 3-25 Final Estimations of Multi-level Model for Edmonton Sample (Size=300)

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
$\ln\alpha_0$	-11.0388	-9.8	<.0001	-0.00846	YES
LT (α_1)	-0.05947	-1.49	0.1361[a]	-0.01442	YES
RT (α_2)	0.1163	2.78	0.0057	-0.01614	YES
β_0	1.0287	6.23	<.0001	-0.08702	YES
area at B (β_1)	-0.1643	-1.86	0.0639	-0.12634	YES
γ_0	0.2727	1.94	0.0536	-0.07205	YES
area at C (γ_1)	0.2212	2.23	0.0267	-0.10352	YES
dispersion parameter, α	0.5455				

Note: a. Since p-value of α_1 is very close to 10%, it could be verified as significance.

Table 3-26 Summarized Statistics of New Edmonton Sample (Size=300)

Dataset (sample size)	Variable	Minimum Value	Maximum Value	Mean (Standard Deviation)
4-leg, signalized at-grade intersections from Edmonton, Alberta, Canada (300); (292 with data on turn lanes)	Multi-vehicle total collisions	0	555	73.6 (78.8)
	Multi-vehicle injury collisions	0	195	22.1 (24.7)
	Years	6	6	6 (0)
	Major AADT	4720	70331	24881 (10990)
	Minor AADT	102	34926	9532 (6973)
	No. of left-turn lanes	0-56; 1-43; 2-91; 3-27; 4-75		
	No. of right-turn lanes	0-99; 1-52; 2-84; 3-11; 4-46		
	Class	urban-146; suburban-154		

In summary, after several rounds of iterative re-sampling and re-estimating, all of the Toronto and Edmonton multi-level models are statistically valid while maintaining consistency per Equation 3-18.

3.5.5 Assessment of Multi-level Model Estimation Results

First, all multi-level models gained final validity with full statistical significance and maintained a consistent model form.

Moreover, the CM-Functions for all impact factors were placed into sub-models; that is, the CM-Functions of left- and right-turn lanes were associated with constant components of first-level SPFs, while locally specific variables showed their impacts in the shape parameters. The structural distinction of the covariates gave rise to deeper insights for safety impact mechanisms, and also emphasized the nature of hierarchy for safety datasets.

3.6 ESTIMATION RESULT COMPARISONS AND DISCUSSIONS

3.6.1 Restructuring of HSM Models

This chapter discussed three concepts for safety models: baseline SPFs multiplied by CMFs, and full and multi-level models. Given the conceptual and practical drawbacks of the first concept, which is currently recommended HSM algorithm, the other two methodologies have been investigated in this chapter.

At first glimpse, multi-level models with the structure shown in Equation 3-18 seemingly do not address equivalent components with configurations of baseline SPFs multiplied by CMFs as shown in Equation 3-7. However, there would be similarities if Equation 3-7 were to be restructured as:

$$N_{mv} = A \times (AADT_{maj})^{b} \times (AADT_{\min})^{c}$$

where $A = \exp(a) \times (CMF_{lt} \times CMF_{rt}) \times C_x$ (3-19)

Now Equation 3-19 looks more like the algorithm of the multi-level model shown in Equation 3-18. Their difference lies in how local information is addressed: for the HSM algorithm of baseline SPFs multiplied by CMFs, as shown in Equation 3-19, local information is addressed by the calibration factor, C_x; in contrast, multi-level models, as shown as Equation 3-18, convey local information by sub-hierarchical CM-Functions and refining shape parameters as functions of local factors. The CM-Functions capture continuous correlation whereas the C_x is just a fixed ratio. By this means, multi-level modeling is conceptually superior to the HSM methodology of baseline SPFs multiplied by CMFs.

3.6.2 Structural Comparisons of Full Model versus Multi-level Model

From a statistical view, the fundamental distinction between full and multi-level models lies in a result obtained from the statistical efforts that the population-based multi-level models presented

in this chapter all gained full statistical significance, while the population-based full models did not. As a result, there were opportunities to re-sample and re-estimate in multi-level models while this process does not have any effect in the full models. The full models failed to wholly attain statistical significance and finally lost some of the variables.

As for the full models themselves, first of all, the paralleling of all the variables that are either common or locally specific resulted in model heterogeneity. Secondly, the full models failed to guarantee statistical significance for all the potential variables. This caused loss of some of the CM-Functions and further aggravated the inconformity of the full models in comparison to HSM model configurations. For many jurisdictions without data and technical adequacy to fully establish their own models, a model imported from another jurisdiction with certain means of conformity is the only solution. Apparently, full models are not the optimal choice to achieve this sharing of models across different jurisdictions. Subsequently, the validity of full models is compromised.

On the contrary, multi-level models restructure the variables to safeguard the homogeneity of first-level SPFs. They address CM-Functions for implication factors through sub-level modeling in a non-parallel pattern, not only continuously and numerically describing the safety impacts, but also constructing more flexible interactions between different covariates. By considering the fact that data heterogeneity prevails for road safety domains, multi-level modeling is an appropriate choice for acquiring CM-Functions without sacrificing model conformity.

In conclusion, the multi-level models developed in this chapter have achieved conceptual superiority over both the HSM algorithm of baseline SPFs multiplied by CMFs and full models. Multi-level modeling is, therefore, the recommended solution.

After appropriate restructuring, the HSM SPFs can be rearranged into Equation 3-19, which is consistent with the multi-level model form described in Equation 3-18. This paves the way for potentially merging calibrated HSM models and locally developed multi-level models. This issue will be further investigated in Chapter 5.

3.6.3 Comparisons of CMFs Yielded from Three Different Models

Except for structural heterogeneity of the three different models – HSM model, full model and multi-level model, they also yielded different results if applied to calculate CMFs for practical application. Some examples were listed in Table 3-27.

Table 3-27 Samples of Collision Modification Factors Generated by Different Models

Collision Modification Factor		Number of Turning Lanes							
		Left Turn				Right Turn			
		1	2	3	4	1	2	3	4
HSM		0.90	0.81	0.73	0.66	0.96	0.92	0.88	0.85
Multi-level Model	Toronto RP	0.94	0.88	0.83	0.78	0.98	0.95	0.93	0.91
	Toronto RG (Sample Size=680)	0.93	0.87	0.81	0.76	0.97	0.93	0.90	0.87
Full Model	Toronto RP	0.86	0.81	0.73	0.80	0.94	0.91	1.00	1.00
	Toronto RG (Sample Size=680)	1.00	0.93	1.00	1.00	0.86	0.80	0.77	1.00

One observation from Table 3-27 is that CMFs for the same factor generated by multi-level models are relatively closer across different datasets, while those by full models are further apart. Moreover, there are several "1.00" CMFs from full models due to absence of relevant variables.

3.6.4 Attempts at Estimating Full Models with Continuous Explanatory Variables

Based on a series of analyses in Section 3.4.2, full models applied categorical variables for number of turning lanes rather than continuous ones, to be consistent with the HSM CMFs.

To reinforce the rationality of this choice, two full models with continuous explanatory variables - number of approaches with left turn lane (lt), number of approaches with right turn lane (rt) - were also attempted. Poisson-gamma models using the "NLMIXED" procedure were estimated based on Toronto and Edmonton reference population data, as shown in Table 3-28 and Table 3-29. The models have the structure:

$$\mu_{it} = \beta_0 \cdot (majorAADT)^{\beta_1} \cdot (\min orAADT)^{\beta_2} \cdot \exp(\beta_3 \cdot lt) \cdot \exp(\beta_4 \cdot rt) \cdot \exp(\beta_{11} \cdot class) \qquad (3\text{-}20)$$

Table 3-28 Estimation Results of Full Model with Continuous Variables for Toronto Reference Population

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-6.7972	-19.31	<.0001	-0.00822	YES
major AADT (β_1)	0.6259	20.54	<.0001	-0.07343	YES
minor AADT (β_2)	0.4701	24.39	<.0001	-0.06692	YES
lt (β_3)	-0.0635	-6.05	<.0001	-0.03548	YES
rt (β_4)	-0.01863	-1.4	0.162	-0.01184	NO
class (β_{11})	-0.07364	-12.06	<.0001	-0.11912	YES
dispersion parameter, α	0.2035				

Table 3-29 Estimation Results of Full Model with Continuous Variables for Edmonton Reference Population

Parameter	Estimate	t Value	Pr > |t|	Gradient	Statistical Significant?
ln(β_0)	-10.9979	-13.17	<.0001	-0.00119	YES
major AADT (β_1)	0.7318	9.48	<.0001	-0.01188	YES
minor AADT (β_2)	0.5984	14.24	<.0001	-0.00971	YES
lt (β_3)	-0.04729	-1.59	0.1135[a]	-0.00158	YES
rt (β_4)	0.1146	3.68	0.0003	0.000052	YES
class (β_{11})	0.3483	4.92	<.0001	-0.00212	YES
dispersion parameter, α	0.5303				

Note: a. Since p-value of β_3 is close to 10%, it could be verified as significance.

The fitting results reinforced the decision not to select continuous variables: for the model based on Toronto reference population, the variable "number of approaches with right turn lane" was not statistically significant; for model based on Edmonton reference population, the p-value for variable "number of approaches with left turn lane" was 0.1135, only "significant" with practical standard of this dissertation (but still shown as weak significance). The more importantly, the coefficient of "rt", β_4, is positive, which is against the well-recognized safety effect.

In conclusion, based on analyses at Section 3.4.2 and above trials, this dissertation will apply categorical variables to represent turn lane information for full models.

3.7 CHAPTER CONCLUSIONS AND FURTHER INVESTIGATIONS

3.7.1 Chapter Conclusions

This chapter has contributed to efforts in seeking viable methodologies for local SPF development. Full and multi-level models have been developed by means of the "NLMIXED" procedure in the SAS software. Based on the performance of these two types of models, the study has found that the multi-level model is more suitable. Full models are found to be unable to use CM-Functions and they fail to maintain homogeneous model components. Such problems would be aggravated with contingent variables eliminated due to statistical insignificance. On the contrary, multi-level models retained homogeneity of first-level SPFs while addressing CM-Functions in sub-level modeling. In addition, multi-level models further demonstrate profound correlations between different impact factors.

In practice, an approach that uses CM-Functions based on multi-level models would overcome the drawbacks of conventional CMFs. Such an approach captures the safety impact of certain factors through a continuous and quantitative approach without problematic and sophisticated categorizations. Multi-level models are more efficient and provide practitioners with first-level AADT-based SPFs and CM-Functions all at once. By this means, practitioners can gauge the crash modification effects of factors through sub-modeling, and then use the outcomes of the sub-modeling to calibrate both the constant multiplier and shape parameters in first-level SPFs to predict safety performance. It should be specifically mentioned that the developed CM-Functions are not any more difficult to apply than conventional HSM CMFs processed in complicated ways.

3.7.2 Further Investigations

For the potential merging of various models, full and multi-level models will be further investigated in Chapter 4 by using different statistical approaches other than the "NLMIXED" procedure, and with different distributions other than the NB distribution that has been applied in

this chapter. Then, there will be various full and multi-level models that can be merged by using a model averaging approach that will be developed in Chapter 6.

CHAPTER 4 DIVERSIFIED LOCAL SPF DEVELOPMENTS

Methodologies for data sampling have been established in Chapter 2. Chapter 3 addressed the development of local SPFs through the traditional analytical model fitting approach, using a Poisson-gamma structure. This chapter will expand local SPF development to other optional approaches with different random distributions, different statistical procedures and through different modeling methods.

The research for this chapter is based on one hypothesis -- that a variety of models can be developed through different ways from a certain dataset, and they can be either selected or merged together, which will be theme of Chapter 6.

In advancing any diversified local modeling effort, it is reaffirmed, as is indirectly implied in Chapter 3, that the fundamental diversity comes from the channels to develop the model, which is either calibrated from existing models or locally developed. The former generally means, in practice, HSM models that have been calibrated through the HSM recommended methodology (FHWA, 2012) which has already been addressed in Chapter 3. The following sections will contribute to the diversification of locally developed models. For the sake of potential model merging (Chapter 6), diversified local model developments will accommodate similar components to relevant calibrated HSM models.

This chapter is an extension of Chapter 3. Hence, some conceptual explanations already presented in Chapter 3 will not be repeated here. Section 4.1 addresses the selection of diversified local model structures. Section 4.2 discusses alternative model development approaches, including traditional "Frequentist" and innovative "Bayesian" approaches. Section 4.3 is focused on the identification of alternative models, while Section 4.4 describes alternative platforms, tools and standards applied for model estimation. Section 4.5 presents outputs of alternative local models. Finally, Section 4.6 summarizes the whole chapter.

4.1 SELECTION OF ALTERNATIVE LOCAL MODEL STRUCTURES

4.1.1 Selection of Model Type

As described in Chapter 3, there are four conventional types of models for SPFs: baseline, general AADT, those with covariates (full models) (Lord et al., 2008b), and multi-level models that are developed in this dissertation study.

Two types of models, full and multi-level, were presented in Chapter 3. In this chapter, alternative SPFs will be investigated for these two types.

4.1.2 Selection of Local Model Components

For the full model, components should consist of both traffic exposure and design features that match as best as possible the elements of the pertinent HSM prediction model. Accordingly, for the sample data from Toronto and Edmonton intersections, the selected local model variables were "major AADT", "minor AADT", and variables that pertain to the number of approaches with turn lanes.

For left turn lanes, the base line is no approach with left turn lanes; the other variables are shown in Table 4-1. Besides that, there is one local special categorical variable for each of the cities; this is "intersection class" for the Toronto sample, and "area" for the Edmonton samples. Table 4-1 lists all variables applied for local full and multi-level model development.

For the multi-level model, components should also consist of both traffic exposure and design features that match as best as possible the elements of the pertinent HSM prediction model. However, this matching only needs to be functional, rather than formal, because the distinctive components of local versus HSM models can be expressed with sub-level models while first-level models maintain homogeneity, as demonstrated in depth in Chapter 3. As a result, multi-level model components can be direct expressions of traffic exposure ("major AADT", "minor AADT"), and design features ("number of approaches with left turn lanes (lt)", "number of

approaches with right turn lanes (rt)"), and the local categorical variable (intersection class for Toronto and area for Edmonton). (See more details in Table 4-1.)

From this aspect, the multi-level model is superior to the full model, with simpler model components.

Table 4-1 Local Model Components

Category	Variable	Value	Model	City
Traffic Exposure	Major AADT	Continuous	Full, Multi-level	Toronto, Edmonton
	Minor AADT	Continuous	Full, Multi-level	
Left Turn Lane	lt	number of approaches with left turn lanes	Multi-level	
	I_{1lt}	1 - 1 approach ;0 - others	Full	
	I_{2lt}	1 - 2 approach ;0 - others	Full	
	I_{3lt}	1 - 3 approach ;0 - others	Full	
	I_{4lt}	1 - 4 approach ;0 - others	Full	
Right Turn Lane	rt	number of approaches with right turn lanes	Multi-level	
	I_{1rlt}	1 - 1 approach ;0 - others	Full	
	I_{2rt}	1 - 2 approach ;0 - others	Full	
	I_{3rt}	1 - 3 approach ;0 - others	Full	
	I_{4rt}	1 - 4 approach ;0 - others	Full	
Local Special Variables	Intersection Class	1-express/express 2-major arterial/expressway 3-major arterial/major arterial 4-expressway/minor arterial 5-major arterial/minor arterial 6-minor arterial/minor arterial 7-unknown 8-major arterial/collector 9-minor arterial/collector 10-collector/collector 11-express/local 12-major arterial/local 13-minor arterial/local 14-collector/local	Full, Multi-level	Toronto
	Area	1-urban 2-suburban	Full, Multi-level	Edmonton

4.1.3 Selection of Local Model Function Form

Miaou et al. (2003) synthesized alternative function forms for AADTs in safety models as shown in Table 4-2. In Table 4-2, $F_{1,it}$ is the AADT of site i, for time t on major approaches, $F_{2,it}$ is the AADT of site i, for time t on minor approaches, $\beta_{0,t}$ is the constant (intercept) of time t, and β_1 , β_2 are coefficients for AADTs.

Miaou et al. (2003) mentioned that functional forms 1 to 4 have been commonly used in previous studies, and that among them, functional form 2 is the most popular. The HSM predictive models apply form 2.

To accommodate potential model merging, the diversified model development will keep functional form 2 as the prime selection. Other alternative forms would impose obstacles for model mixture with calibrated HSM models and consequently are inappropriate for application in the first place. Instead, alternative forms will be attempted only in the event that the first choice does not lead to a satisfactory result.

Table 4-2 Commonly Used and Alternative Function Forms for AADTs in Safety Model

Form Number	Function Form f(.)
1	$\beta_{0,t}(F_{1,it} + F_{2,it})^{\beta_1}$
2	$\beta_{0,t}(F_{1,it})^{\beta_1}(F_{2,it})^{\beta_2}$
3	$\beta_{0,t}(F_{1,it} \times F_{2,it})^{\beta_1}$
4	$\beta_{0,t}\left(F_{1,it} + F_{2,it}\right)^{\beta_1}(\frac{F_{2,it}}{F_{1,it}})^{\beta_2}$
5	$\beta_{0,t}(F_{1,it})^{\beta_1}(F_{2,it})^{\beta_2}\exp(\beta_3 F_{2,it})$
6	$F_{1,it}\lambda_{1,it} + F_{2,it}\lambda_{2,it}$ where, $\lambda_{1,it} = \exp(\beta_{0,t} + \beta_1 F_{2,it})$ $\lambda_{2,it} = \exp(\beta'_{0,t} + \beta_2 F_{1,it})$

4.2 ALTERNATIVE APPROACHES APPLIED FOR MODEL ESTIMATION

The original safety models presented in Chapter 3, as well as those in the HSM, were developed with a traditional **"**frequentist**"** approach. The alternative is a "Bayesian" approach.

The most frequently used statistical model estimation methods are known as frequentist (or classical) methods in which statistical fitting is conducted and often, but not always, fixed parameters are obtained. Hence, there is no way that probabilities can be associated with these parameters (Everitt, 2002; Neyman, 1937). Frequentist methods assume that unknown parameters are fixed constants. They define probability by using limiting relative frequencies. It follows from these assumptions that probabilities are objective and that probabilistic statements cannot be made about parameters because they are fixed (SAS Institute Inc., 2012).

The alternative to the frequentist approach is the Bayesian approach. The Bayesian approach combines prior knowledge from a reference population and information from site-specific observations. Bayesian methods treat parameters as random variables and define probability as "degrees of belief". That is, the probability of an event is the degree to which one believes the event is true. It follows from these postulates that probabilities are subjective and one can make probability statements about parameters (Bernardo, 1994; Congdon, 2001; Lee, 1997; Persaud et al., 2010a).

Suppose one is interested in estimating θ from data $y=\{y_1, y_2, \ldots, y_n\}$ by using a statistical model described by a density $p(y|\theta)$. The Bayesian philosophy states that θ cannot be exactly determined, and uncertainty about the parameter is expressed through probability statements and distributions. The following steps describe the essential elements of Bayesian inference (SAS Institute Inc., 2012; Gelman et al., 2004).

1) A probability distribution for θ is formulated as $\pi(\theta)$, which is known as the "prior distribution", or just simply the "prior". The prior distribution expresses one's beliefs (for example, on the mean, the spread, the skewness, and so forth) about the parameters before the data are examined.

2) Given the observed data y, a statistical model $p(y|\theta)$ is chosen to describe the distribution of y, given θ.
3) Beliefs about θ can be updated by combining information from the prior distribution and the data through the calculation of the "posterior distribution", $p(\theta|y)$.

The third step is carried out by using the "Bayes' theorem", which enables the combining of the prior distribution and the model in the following way:

$$p(\theta|y) = \frac{p(\theta,y)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{\int p(y|\theta)\pi(\theta)d\theta} \quad (4\text{-}1)$$

The quantity $p(y) = \int p(y|\theta)\,\pi(\theta)d\theta$ is the normalizing constant of the posterior distribution. This quantity $p(y)$is also the marginal distribution of y, and is sometimes called the "marginal distribution of the data". The likelihood function of θ is any function proportional to $p(y|\theta)$; that is, $L(\theta) \propto p(y|\theta)$.

Both frequentist (classical) and Bayesian methods have their advantages and disadvantages, and there are some similarities.

For the Bayesian approach, firstly, the model parameters are treated as unknown random variables with inferences based on the posterior distributions of the parameters. This facilitates more flexibility than the estimation of fixed parameters with the "frequentist" approach (Carriquiry et al., 2005; Miaou et al., 2003; Congdon, 2001; Lan et al., 2009; Persaud et al., 2010a). Some other advantages to using Bayesian analysis include the following: (SAS Institute Inc., 2012; Berger, 1985; Berger and Wolpert, 1988; Bernardo and Smith, 1994; Carlin and Louis, 2000; Robert, 2001; Wasserman, 2004).

- It provides a natural and principled way of combining prior information with data, within a solid decision theoretical framework. Users can incorporate past information about a parameter and form a prior distribution for future analysis. When new observations become available, the previous posterior distribution can be used as a prior.
- It provides inferences that are conditional on the data and are exact, without reliance on asymptotic approximation. Small sample inference proceeds in the same manner as if one

had a large sample. Bayesian analysis can also directly estimate any functions of parameters, without using the "plug-in" method (a way to estimate functions by plugging the estimated parameters into the functions).

- It obeys the likelihood principle. If two distinct sampling designs yield proportional likelihood functions for θ, then all inferences on θ should be identical from these two designs. Classical inference does not, in general, obey the likelihood principle.
- It provides interpretable answers, such as "the true parameter has a probability of 0.95 of falling into a 95% credible interval."
- It provides a convenient setting for a wide range of models, such as multi-level models and missing data problems.

However, at the same time, there are also disadvantages for using a Bayesian approach as follows (SAS Institute Inc., 2012; Berger, 1985; Berger and Wolpert, 1988; Bernardo and Smith, 1994; Carlin and Louis, 2000; Robert, 2001; Wasserman, 2004).

- It does not specify how to correctly select a prior. Thus, prior beliefs could be misleading.
- It can produce posterior distributions that are heavily influenced by the priors. From a practical point of view, it might be sometimes difficult to convince subject matter experts who do not agree with the validity of the chosen prior.
- It often requires large computational times, especially in models with a large number of parameters. In addition, simulations provide slightly different answers unless the same random seed is used. Note that slight variations in simulation results do not contradict the early claim that Bayesian inferences are exact. The posterior distribution of a parameter is exact, given the likelihood function and the priors, while simulation-based estimates of posterior quantities can vary due to the random number generator used in the procedures.

Frequentist and Bayesian approaches are different in terms of inference method. The frequentist approach derives model inferences from analytical or numerical optimization algorithms based on likelihood, e.g., maximum likelihood estimation (MLE), while the Bayesian approach derives inferences from simulation, which means slightly different answers from each run are expected for the same problem (SAS Institute Inc., 2012; Berger, 1985).

Frequentist and Bayesian approaches also have similarities. In particular, when the sample size is large, Bayesian inference often provides results for parametric models that are very similar to the results produced by the classical frequentist methods.

In this chapter, both Bayesian and frequentist approaches are applied for model development.

4.3 IDENTIFICATION OF ALTERNATIVE LOCAL MODELS

Equations 3-8 and 3-9 demonstrated the common form of an SPF, which is a mixed Poisson distribution family, and that the precise form of the distribution of the safety model ultimately depends on the specific choice of the probability distribution of ε_i, the error term introduced in Equation 3-9. The preliminary local model development in Chapter 3 solely applied an NB distribution that is actually a special case of the Poisson-gamma distribution with certain characteristics (Lan et al., 2009).

For the alternative model developments with the use of the frequentist approach in this chapter, besides the NB, the zero-inflated Poisson (ZIP) distribution will also be applied. Zero-inflated models, which assume a dual-state process for crash data - the "zero" state and the normal count state - have a long history, being adopted for safety modeling from early efforts such as Shankar et al. (1997), and culminating in the recent modification using a full Bayes hierarchical approach (Aguero-Valverde, 2013). Other areas also applied ZIP models, e.g., injury model due to falls (Ullah et al., 2010). Lord et al. (2005) argued that although zero-inflated models provided good statistical fits, they did not characterize the underlying crash process, so that they should be avoided for safety models. Regardless of these arguments, zero-inflated models will still be retained to provide a wider variety of alternate models used to demonstrate the diversity of model development for a before-after evaluation process.

As discussed in Section 4.2, the Bayesian approach relies on prior belief. For the Bayesian approach, the mixed Poisson distribution family is common in road safety practice (Persaud et al., 2010a; Chou and Steenhard, 2009). Lan et al. (2009) developed and compared the Poisson-gamma and Poisson-lognormal models. Chou and Steenhard (2009) discussed more optional mixed Poisson distribution models suitable for "count data", which would characterize observed

collision frequencies. Based on these early practices, and considering proximity of optional error terms for the development of alternative models in this chapter, Poisson-gamma, Poisson-lognormal, and Poisson-Weibull distributions, with ε_i, the error term introduced in Equation 3-9, assumed as Gamma, lognormal and Weibull distributed, respectively, will be utilized.

The modeling choices for this chapter are summarized in Table 4-3. All of the choices in Table 4-3 will be implemented for two model hierarchies: full and multi-level models.

Table 4-3 Characteristics of Distributions Applied

Approaches	Distribution Structure	Distribution of ε_i	Character of Parameter
Bayesian	Poisson-gamma (NB)	Gamma (φ, (1/ φ))	φ~ Gamma(1,1)
	Poisson-lognormal	Lognormal (- (σ^2/2), σ^2)	σ^2 ~ inverse gamma (0.001, 0.001)
	Poisson-Weibull	Weibull (0, c, (1/Γ(1+1/c)))	c ~ normal (0,sd=1000)
Frequentist	NB (Poisson-gamma)	Gamma (φ, (1/ φ))	φ~ Gamma(1,1)
	Zero-inflated Poisson: collisions≡0 with probability φ_i; otherwise, collisions ~ Poisson distribution with probability (1-φ_i)		

4.4 ALTERNATIVE PLATFORMS, TOOLS AND STANDARDS APPLIED FOR MODEL ESTIMATION

4.4.1 Estimations with the Frequentist Approach

SAS statistical software (SAS Institute Inc., 2012) was chosen as the modeling platform for this study.

For models developed through the frequentist approach, three procedures were applied: generalized linear estimation with NB (the GENMOD procedure), generalized linear estimation

with ZIP (COUNTREG procedure), and non-linear mixed estimation (NLMIXED procedure) (SAS Institute Inc., 2012).

For the GENMOD and COUNTREG procedures, the distribution types applied were NB (as the default) and ZIP. For the NLMIXED procedure, the distribution applied was Poisson-gamma, which is analogous to the NB in this case.

For estimations made by the "GENMOD procedure" and "COUNTREG procedure", a rule that requires all parameter estimates to be significant at the 10% level was stipulated for deeming a model significant.

The NLMIXED procedure runs through iterative optimizations, so its rules for assessing significance rely on the final outcomes of iterations. Thus, the rules for a significant estimation were stipulated as:

- model optimization, as a whole, is finally convergent (output of iteration history finally displayed as "GCONV convergence criterion satisfied", that is, the relative gradient convergence, GCONV<= 1E-8),
- all parameter estimates are significant at the 10% level as judged by the p-values of their t-test, and
- all parameter estimates are finally stable (i.e., each parameter estimate yields a very small gradient that reflects the improvement of optimization, up to the last step of the iteration; stability is achieved when the gradient is smaller than 0.0001).

4.4.2 Estimations with the Bayesian Approach

The MCMC procedure in SAS software was applied for the Bayesian model estimations. This procedure complies with the Markov chain Monte Carlo process, but is applied by SAS as a general purpose simulation tool to conduct Bayesian modeling (SAS Institute Inc., 2012). First, prior distributions of all parameters were assumed as non-informative and normal (0, 1000) per Lan et al. (2009).

"Quasi-Newton" was chosen as the optimization method of the MCMC procedure. The development of Bayesian models by MCMC is actually a simulation procedure that comprises iterations that cease when certain criteria are met, specifically when the next iteration no longer yields significant improvements. In other words, the iteration is converged and at the same time, all parameter estimates are statistically significant. The assessment criteria were accordingly defined as:

- model optimization, as a whole, is finally convergent (output of iterative optimization history finally displayed as "Convergence criterion satisfied", that is, the relative function convergence criterion, FCONV<=1E-6), and
- all parameter estimates are significant as determined by the Heidelberger-Welch Diagnostics (HWD) of the Markov chain convergence (Heidelberger and Welch，1983; Cowles and Carlin, 1996; SAS Institute Inc., 2012). This requires that both the "stationarity test" (verifying that the iteration is finally stationary) and the "half-width test" (verifying that the sample size is adequate) are passed for all parameter estimates.

4. 5 OUTPUTS OF ALTERNATIVE LOCAL MODELS

Various classes, forms, approaches and SAS procedures, and diverse local model development were investigated. For each model among the selections, consecutive estimation attempts were made and insignificant variables were removed until the relevant significance criteria, as listed in Section 4.2.4, were fully satisfied.

The results of all final model estimations based on the reference population and different samples as described in Chapter 2, are summarized in Table 4-4 to Table 4-9. These tables amalgamate the diverse developments in this chapter and original developments from Chapter 3. There were generally insignificant outcomes for the FI collision models, so the decision was taken to focus the rest of the investigation on models for total collisions.

Table 4-4 Final Results of Model Estimations (Toronto, population)

Model Hierarchy	Approach	Procedure	Measure	Output
Baseline model multiplied by CMFs	Calibrated	HSM type calibration	log-likelihood	-7874.5
Full Model	Frequentist	GENMOD NB	Overall significance	✓[a]
			Parameter estimates	2 removed; other p-values < 0.08
			log-likelihood	-7241.7
		COUNTREG ZIP	Overall significance	✓
			Parameter estimates	0 removed; all p-values < 0.0001
			log-likelihood	-1440722
		NLMIXED	Overall significance	✓
			Parameter estimates	2 removed; other p-values < 0.02
			log-likelihood	-7277.5
	Bayesian	Poisson - gamma	Overall significance	✓
			Parameter estimates	2 removed; others passed HWD
			log-likelihood	-7694.19
		Poisson - lognormal	Overall significance	✓
			Parameter estimates	9 removed; others passed HWD
			log-likelihood	-13741.7
		Poisson - Weibull	Overall significance	✓
			Parameter estimates	9 removed; others passed HWD
			log-likelihood	-12971
Multi-level Model	Frequentist	GENMOD NB	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		COUNTREG ZIP	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		NLMIXED	Overall significance	✓
			Parameter estimates	0 removed; all p-values < 0.09
			log-likelihood	-7283.5
	Bayesian	Poisson - gamma	Overall significance	✓(marginal[b])
			Parameter estimates	0 removed; all but two passed HWD
			log-likelihood	-7705.22
		Poisson - lognormal	Overall significance	✕
			Parameter estimates	0 removed; three failed HWD
			log-likelihood	-13089.7
		Poisson - Weibull	Overall significance	✕
			Parameter estimates	No significant output
			log-likelihood	Null

Note: a. ✓-significant, ✕-insignificant;

Note: b. "marginal" means estimations generally significant for "HWD" diagnosis with all passing "Stationarity Test" except for one or two parameter estimates narrowly missing the "Half-width Test" (absolute values of Half-width closer but slightly larger than the threshold - 0.10).

Multi-level models do not conceptually adapt for the GENMOD and COUNTREG procedures since these are only designed for single-level modeling.

Table 4-5 Final Results of Model Estimations (Toronto, sample size=680)

Model Hierarchy	Approach	Procedure	Measure	Output
Baseline model multiplied by CMFs	Calibrated	HSM type calibration	log-likelihood	-3261.47
Full Model	Frequentist	GENMOD NB	Overall significance	✓
			Parameter estimates	3 removed; other p-values < 0.09
			log-likelihood	-3027.7
		COUNTREG ZIP	Overall significance	✓
			Parameter estimates	0 removed; all p-values < 0.0001
			log-likelihood	-599533
		NLMIXED	Overall significance	✓
			Parameter estimates	3 removed; other p-values < 0.002
			log-likelihood	-3035.85
	Bayesian	Poisson -gamma	Overall significance	✓
			Parameter estimates	3 removed; other passed HWD
			log-likelihood	-3226.6515
		Poisson - lognormal	Overall significance	✓
			Parameter estimates	9 removed; other passed HWD
			log-likelihood	-6102.63
		Poisson -Weibull	Overall significance	✕
			Parameter estimates	9 removed; one more failed HWD
			log-likelihood	-6102.59
Multi-level Model	Frequentist	GENMOD NB	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		COUNTREG ZIP	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		NLMIXED	Overall significance	✓
			Parameter estimates	0 removed; all p-values < 0.10
			log-likelihood	-3053.45
	Bayesian	Poisson -gamma	Overall significance	✓(marginal)
			Parameter estimates	0 removed; all but two passed HWD
			log-likelihood	-3229.14
		Poisson - lognormal	Overall significance	✓(marginal)
			Parameter estimates	1 removed; other passed HWD
			log-likelihood	-5915.09
		Poisson -Weibull	Overall significance	✕
			Parameter estimates	No significant output
			log-likelihood	Null

Model comparison criteria are mainly rooted in model's log-likelihood and the number of parameters. Thus, the log-likelihood is sufficient for model comparability. The listed in Table 4-4 to Table 4-9 are full log-likelihoods which include terms that involve binomial coefficients or factorials of the observed counts for discrete distributions. Furthermore, those likelihood values

for Bayesian models are actually the likelihoods estimated at the posterior mean since there is no fixed likelihood for a Bayesian model (SAS Institute Inc., 2012; Claeskens and Hjort, 2009).

Table 4-6 Final Results of Model Estimations (Toronto, sample size=588)

Model Hierarchy	Approach	Procedure	Measure	Output
Baseline model multiplied by CMFs	Calibrated	HSM type calibration	log-likelihood	-2827.884
Full Model	Frequentist	GENMOD NB	Overall significance	✓
			Parameter estimates	5 removed; other p-values < 0.10
			log-likelihood	-2602.2
		COUNTREG ZIP	Overall significance	✓
			Parameter estimates	1 removed; all p-values < 0.0001
			log-likelihood	-516522
		NLMIXED	Overall significance	✓
			Parameter estimates	4 removed; other p-values < 0.10
			log-likelihood	-2633.45
	Bayesian	Poisson -gamma	Overall significance	✓
			Parameter estimates	3 removed; all others but one passed HWD
			log-likelihood	-2782.8785
		Poisson - lognormal	Overall significance	✕
			Parameter estimates	9 removed; 2 others failed HWD
			log-likelihood	-5213.89
		Poisson -Weibull	Overall significance	✓
			Parameter estimates	9 removed; other passed HWD
			log-likelihood	-5213.26
Multi-level Model	Frequentist	GENMOD NB	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		COUNTREG ZIP	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		NLMIXED	Overall significance	✓
			Parameter estimates	0 removed; all p-values < 0.05
			log-likelihood	-2631.1
	Bayesian	Poisson -gamma	Overall significance	✓(marginal)
			Parameter estimates	0 removed; all but two passed HWD
			log-likelihood	-3229.04
		Poisson - lognormal	Overall significance	✕
			Parameter estimates	0 removed; 4 failed HWD
			log-likelihood	-12972.9
		Poisson -Weibull	Overall significance	✕
			Parameter estimates	No significant output
			log-likelihood	Null

Table 4-7 Final Results of Model Estimations (Edmonton, population)

Model Hierarchy	Approach	Procedure	Measure	Output
Baseline model multiplied by CMFs	Calibrated	HSM type calibration	log-likelihood	-2670.63
Full Model	Frequentist	GENMOD NB	Overall significance	✓
			Parameter estimates	6 removed; other p-values < 0.02
			log-likelihood	-2481.4
		COUNTREG ZIP	Overall significance	✕
			Parameter estimates	0 removed; all but one p-values < 0.0001
			log-likelihood	Null
		NLMIXED	Overall significance	✓
			Parameter estimates	7 removed; other p-values < 0.0009
			log-likelihood	-2484
	Bayesian	Poisson -gamma	Overall significance	✓
			Parameter estimates	7 removed; others passed HWD
			log-likelihood	-2556.54
		Poisson - lognormal	Overall significance	✕
			Parameter estimates	7 removed; others failed HWD
			log-likelihood	-7717.55
		Poisson -Weibull	Overall significance	✓
			Parameter estimates	7 removed; others but one passed HWD
			log-likelihood	-7718.12
Multi-level Model	Frequentist	GENMOD NB	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		COUNTREG ZIP	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		NLMIXED	Overall significance	✓
			Parameter estimates	0 removed; all p-values < 0.09
			log-likelihood	-2480.4
	Bayesian	Poisson -gamma	Overall significance	✓(marginal)
			Parameter estimates	0 removed; all but two passed HWD
			log-likelihood	-2554.78
		Poisson - lognormal	Overall significance	✕
			Parameter estimates	0 removed; all but one failed HWD
			log-likelihood	-7531.52
		Poisson -Weibull	Overall significance	✕
			Parameter estimates	No significant output
			log-likelihood	Null

Table 4-8 Final Results of Model Estimations (Edmonton, sample size=400)

Model Hierarchy	Approach	Procedure	Measure	Output
Baseline model multiplied by CMFs	Calibrated	HSM type calibration	log-likelihood	-2063.518
Full Model	Frequentist	GENMOD NB	Overall significance	✓
			Parameter estimates	6 removed; other p-values < 0.02
			log-likelihood	-1925.6
		COUNTREG ZIP	Overall significance	✕
			Parameter estimates	0 removed; all but one p-values < 0.0001
			log-likelihood	Null
		NLMIXED	Overall significance	✓
			Parameter estimates	6 removed; other p-values < 0.004
			log-likelihood	-1911.4
	Bayesian	Poisson -gamma	Overall significance	✓
			Parameter estimates	7 removed; others passed HWD
			log-likelihood	-1981.68
		Poisson - lognormal	Overall significance	✓
			Parameter estimates	9 removed; others passed HWD
			log-likelihood	-7792.27
		Poisson -Weibull	Overall significance	✕
			Parameter estimates	7 removed; others failed HWD
			log-likelihood	-6462.66
Multi-level Model	Frequentist	GENMOD NB	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		COUNTREG ZIP	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		NLMIXED	Overall significance	✓
			Parameter estimates	0 removed; all p-values < 0.11
			log-likelihood	-1922.15
	Bayesian	Poisson -gamma	Overall significance	✓(marginal)
			Parameter estimates	0 removed; all but two passed HWD
			log-likelihood	-1977.54
		Poisson - lognormal	Overall significance	✕
			Parameter estimates	0 removed; three failed HWD
			log-likelihood	-6215.65
		Poisson -Weibull	Overall significance	✕
			Parameter estimates	No significant output
			log-likelihood	Null

Table 4-9 Final Results of Model Estimations (Edmonton, sample size=300)

Model Hierarchy	Approach	Procedure	Measure	Output
Baseline model multiplied by CMFs	Calibrated	HSM type calibration	log-likelihood	-1549.476
Full Model	Frequentist	GENMOD NB	Overall significance	✓
			Parameter estimates	6 removed; other p-values < 0.01
			log-likelihood	-1447.4
		COUNTREG ZIP	Overall significance	✕
			Parameter estimates	0 removed; all p-values < 0.0001
			log-likelihood	Null
		NLMIXED	Overall significance	✓
			Parameter estimates	7 removed; other p-values < 0.02
			log-likelihood	-1459.25
	Bayesian	Poisson -gamma	Overall significance	✓
			Parameter estimates	7 removed; others passed HWD
			log-likelihood	-1494.03
		Poisson - lognormal	Overall significance	✓
			Parameter estimates	9 removed; others passed HWD
			log-likelihood	-5354.47
		Poisson -Weibull	Overall significance	✕
			Parameter estimates	7 removed; others failed HWD
			log-likelihood	-4491.9
Multi-level Model	Frequentist	GENMOD NB	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		COUNTREG ZIP	Overall significance	Conceptual foundation not matched
			Parameter estimates	
			log-likelihood	
		NLMIXED	Overall significance	✓
			Parameter estimates	0 removed; all p-values < 0.13
			log-likelihood	-1448.4
	Bayesian	Poisson -gamma	Overall significance	✓(marginal)
			Parameter estimates	1 removed; all others but one passed HWD
			log-likelihood	-1493.92
		Poisson - lognormal	Overall significance	✕
			Parameter estimates	1 removed; all others but one failed HWD
			log-likelihood	-4360.08
		Poisson -Weibull	Overall significance	✕
			Parameter estimates	No significant output
			log-likelihood	Null

While likelihood values were obtained through SAS procedures for local models, a special maximum likelihood procedure was applied for calibrated HSM models by adapting the one in Persaud et al. (2011b) to estimate the "full log-likelihoods".

In this procedure, a range for the over-dispersion parameter (k) values is first specified. Then, at every incremental value of k, and for each j = 1 to N sites, the following equations are applied:

$$a = (\frac{1}{k}) \cdot LOG(\frac{1}{k} \cdot \frac{1}{predicted})$$

$$b = (\frac{1}{k} + observed) \cdot LOG(\frac{1}{k} \cdot \frac{1}{predicted} + 1)$$

$$c = \sum_{i=1}^{observed} [LOG(\frac{1}{k} + i - 1) - LOG(i)]$$

where

k = the incremental over-dispersion parameter to which the calculation applies,

predicted = the collision prediction from the model for site j, and

observed = the observed number of collisions at site j.

The log-likelihood is then calculated as:

$$Log likelihood = \sum_{j=1}^{N} a - \sum_{j=1}^{N} b + \sum_{j=1}^{N} c \qquad (4\text{-}2)$$

where N = Total Number of sites.

Within the range of k values, if there is a peak value of the log-likelihood, then it is selected. If there is no peak, then a broader range of potential values of the over-dispersion parameter is used. The estimated log-likelihood values of calibrated HSM models have been listed in Table 4-4 to Table 4-9.

Table 4-10 assembles likelihood values from selected statistically significant models to further compare their fit. Here, higher negative likelihood values, i.e., closer to zero, indicate a better fit. The results were from SAS "NLMIXED" procedure for Frequentist models and "Poisson – gamma" procedure for Bayesian models.

Table 4-10 Comparisons of Log-likelihood Values of Selected Models

Approach	Model	Log-likelihood Values					
		Toronto			Edmonton		
		Reference Population	Sampe (Size=680)	Sampe (Size=588)	Reference Population	Sampe (Size=400)	Sampe (Size=300)
Calibrated	HSM	-7875	-3261	-2828	-2671	-2064	-1549
Frequentist	Full	-7278	-3036	-2633	-2484	-1911	-1459
	Multi-level	-7284	-3053	-2631	-2480	-1922	-1448
Bayesian	Full	-7694	-3227	-2783	-2557	-1982	-1494
	Multi-level	-7705	-3229	-3229	-2555	-1978	-1494

There are two key indications from Table 4-10:

- Locally developed models have clearly better fit than calibrated HSM models
- Model fit for full and multi-level models are not significantly different. For some datasets of Table 4-10 multi-level models have better fit than full models while others show the opposite. All the differences between full and multi-level models are too small to be significant

4.6 CHAPTER CONCLUSIONS

This chapter conducted alternative local model developments, extending the standard local model developments introduced in Chapter 3. The outcomes of this chapter are a series of local SPFs developed with different hierarchies, estimated by either "Frequentist" or "Bayesian" approaches and with different SAS procedures.

Three measurements were used to test statistical significance of models: overall significance, parameter estimates, and log-likelihood values. After insignificant variables removed, the remaining variables passed significance tests, and the models as a whole achieved overall significance, optional local SPFs were developed. All statistically significant models developed in this chapter would be kept as viable knowledge sources that will be utilized for the next steps of the research.

In the next steps, given a variety of potential models for local application, including the calibrated HSM model and some locally calibrated ones, the assessment process will require a final decision to determine which model should be applied. Rather than applying the conventional process to compare, filter, and finally recommend the best single one while abandoning all other models, an alternative approach to merge all considered models will be investigated, as described in Chapter 6.

CHAPTER 5 TRANSFORMING PRIOR KNOWLEDGE WITH SAFETY SURROGATES

While SPFs are the most common source for determining prior knowledge, they are not the only means. When SPFs are not available, or fail to be properly calibrated into a local model, the local jurisdiction has to seek another means. In this case, an indirect measure of safety - safety surrogates, is a viable alternative.

This chapter is focused on safety surrogates. After review and discussion of basic concepts and methodologies in Section 5.1, Section 5.2 establishes the framework to develop the safety surrogates based on different scenarios. After sample facility and data introduction in Section 5.3, two approaches to develop safety surrogates, one via a surrogate-based predictive model, another via non-modeled indirect conversions, are utilized separately in Section 5.4 and 55. Finally, Section 5.6 summarizes the research presented in this chapter.

A paper based on this aspect of the research, which contains additional contributions from others beyond this dissertation research has already been published (Chen at al. 2012).

5.1 BASIC CONCEPTS AND RELEVANT PAST RESEARCH ON SAFETY SURROGATES

Surrogate measures of safety have become a popular research topic in traffic safety domain and there are a variety of different concepts, methods and practices. In order to locate the most appropriate approaches for B/A process, a thorough literature review and discussion is necessary. This is introduced in this section.

5.1.1 Basic Concepts and Classification of Safety Surrogates

In their research report to the USA Federal Highway Administration (FHWA), Gettman and Head (2001) enumerated possible safety surrogates as:

- conflict, defined as an observable situation in which two or more road users approach each other at a certain time and space to such an extent that there is the risk of collision if their movement remain unchanged (Amundsen and Hyden, 1977), and involves:
    - gap time
    - encroachment time
    - deceleration rate (DR)
    - proportion of stopping distance
    - post-encroachment time
    - initial attempted post-encroachment time
    - time to collision (TTC);
- standard measures of effectiveness, including:
    - delay
    - travel time
    - speed related measures, such as speed, speed distribution and speed variance
    - percentage of stops
    - percentage of left turns
    - queue length; and
- other operational measures, including (Perkins and Bowman, 1986; Thompson and Perkins, 1983; Fitzpatrick et al., 1999):
    - DR distribution
    - stop-bar encroachments
    - required braking power distributions
    - distribution of merge points (freeway)
    - merge area encroachments (freeway on-ramp merging in weaving sections)
    - gap-acceptance distributions
    - number of vehicles caught in dilemma zones

    - red- and yellow-light violations by phase
- non-operational measures, mostly design features used for road segments such as super-elevation, curvature, distance since last curve, etc. (Harwood et al., 2000; Perkins and Bowman, 1986; Thompson and Perkins, 1983; Fitzpatrick et al., 1999)

5.1.2 Conflicts Applied as a Safety Surrogate

Much research work has contributed to the most prevalent surrogate - conflict. Specific software, for example, the Surrogate Safety Assessment Model (SSAM), have been developed, frequently alongside simulation platforms, to investigate conflict measures and characteristics, especially for intersections (Gettman and Head, 2001; Gettman et al., 2008; FHWA, 2009b; Tarko et al., 2009; Stevanovic, 2011).

Compared to the abundance of conflict dominated research and practices, there are noticeably few studies that contribute to examining the correlation between surrogates and collisions despite the fact that this is the basis for determining the eligibility of a surrogate. In their technical report, Gettman et al. (2008) employed a regression technique to relate actual crash frequency to conflict frequency predicted by the SSAM which resulted in the following model:

$$Crashes = 0.119 \times Conflicts^{1.419} \quad (5\text{-}1)$$

The R-square of Equation 5-1 is 0.41, which indicates a moderately significant regression.

A white paper titled “Surrogate Measures of Safety” (Tarko et al., 2009) introduced systematically past researches about safety surrogates. Among them, one study by Sayed and Zein (1999) employed regression analyses to develop predictive models that relate the number of traffic conflicts to traffic volume and accidents from 92 intersections. Both conflicts and accidents were assumed to follow a Poisson distribution. A statistically significant relationship was found between accidents and conflicts with an R^2 in the range of 0.70 - 0.77 at signalized junctions, but not at unsignalized intersections.

5.1.3 Alternative Safety Surrogates

A search conducted for literature on surrogates other than conflicts was not very productive and led to few concrete findings, regardless of whether conflict was deemed an insignificant issue or a priority safety concern.

Boonsiripant (2009) wrote his PhD dissertation on this topic by employing speed variation derived from GPS-instrumented vehicles as a safety surrogate for network screening, i.e., identification of black spots. The quantified relationships between surrogate measures and crash frequency were developed by using binary recursive partitioning methods and a GLM approach. After separating high and low stop frequencies (number of stops/trip/mile), he finally developed a series of GLM models that connect the surrogate measure and collisions. One of his models, for example, has the following function:

$$ACC_{HC} = L^{0.706} \times e^{[2.2923+2.0295(AN\_A_i)]} \quad (5\text{-}2)$$

where

ACC_{HC} = estimated crashes for high crash frequency corridors,

L = segment length (mile), and

AN_Ai = acceleration noise, which is defined as the root-mean-square of the acceleration of vehicles.

5.1.4 Assessment of Past Research

In summary, research in the past has displayed methodological gaps in three aspects:

1) It has not accommodated other safety factors very well in relationships between surrogates and crashes,
2) It has not sufficiently connected surrogates to physical features that directly reflect safety treatments, such as geometric features and facility characters, and
3) It has not covered all practical methods for the purpose of converting surrogates into crash measures.

Accordingly, this dissertation will conduct research work to address these three gaps.

5.2 FRAMEWORK FOR SAFETY SURROGATE DEVELOPMENT

5.2.1 Basic Characteristics of Safety Surrogates

The white paper titled "Surrogate Measures of Safety" (Tarko et al., 2009) depicted the relationship between surrogates and crash measures per Figure 5-1.

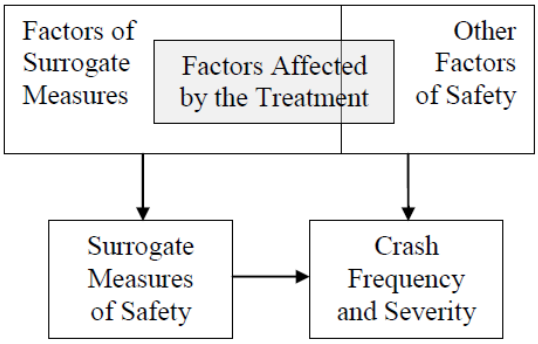


Figure 5-1 Relationship between Surrogate Measures and Crash Features

Figure 5-1 indicates that a genuine relationship between surrogates and crashes needs to take into account other safety factors.

Synthesis of other investigations in this white paper (Tarko et al., 2009) led to the conclusion that a surrogate measure should satisfy two conditions:

1) It must be physically present for crash predictive modeling alongside other factors. Or, there exists a practical method to convert non-crash features into crash frequency or severity, and
2) It can fully capture the effect of a safety treatment. This requires that the surrogate measure be well expressed from physical features that directly reflect the treatment, such as geometric factors or specific facility features.

Regardless of the way that a surrogate is developed, it must have these two characteristics.

5.2.2 Two Scenarios for Safety Surrogate Development

The first condition required of safety surrogates means that they need to be developed under two different scenarios.

The first scenario is to statistically develop a surrogate-based collision predictive model.

The second scenario is to develop an indirect algorithm for the case where the surrogate-based collision predictive model is not available.

Section 5.3 will summarize the sample facility, surrogate measure and datasets that are selected for study in this dissertation. Sections 5.4 and 5.5 will contribute to surrogate measure development based on the two above mentioned scenarios.

5.3 SAMPLE FACILITY, SURROGATE MEASURE AND DATA

As the first step of statistical analysis based on sample data, this section introduces the sample facility applied in this dissertation, the measurement recommended as a safety surrogate, and a description of sample data.

Considering data accessibility, modern roundabouts was selected as the sample facility type for the subsequent analyses and discussions. The raw data of modern roundabouts were obtained from United States and Italy with the assembly of design features, operational measures (e.g., a series of operational speed measures) and collision records.

The first part of this section will address the selection of a safety surrogate appropriate for modern roundabouts. Then the following sub-section will describe in detail the raw data collected from the roundabout sample. The last part of this section will introduce some data items derived from raw data which were necessary for subsequent analyses.

5.3.1 Operational Speed of Roundabouts as Surrogate Measure

The nature of roundabout traffic operations dictates that conflict is not the principal concern, but speed is. There is now substantial evidence that indicates that modern roundabouts can significantly reduce traffic crashes (FHWA, 2008) and that the safety benefits largely result from the fact that they are designed to control traffic speed. It stands to reason, therefore, that the safety performance of roundabouts is related to a measure of their operating speed. That is to say, operating speed is a rational safety surrogate for roundabouts.

5.3.2 Summary Statistics of Raw Data

This study uses data obtained from different approaches for 139 roundabouts from 8 states in the U.S., and 34 roundabouts from 3 cities in Italy. Cross-country datasets solidify the research foundation and widen the scope of the studied methodologies.

U.S. roundabouts can be found in various environments, including urban, suburban and rural contexts, while Italian roundabouts are only present in urban and suburban contexts. Four types of observed speeds - approach, entry, upstream (left side of approach) circulating and upstream exiting speeds - and three types of speed differentials between each pair of adjacent speeds are available for 34 of the U.S. roundabouts and 6 of the Italian roundabouts. These are all median speeds. The U.S. sample database was also used in an earlier NCHRP study (Rodegerdts et al., 2007). Table 5-1 shows the summary statistics of the data.

In Table 5-1, speed differential is the arithmetic difference of two adjacent speed measures (e.g., Speed Differential of Approach vs. Entry = Approach Speed minus Entry Speed). The speed differential data were only taken when both types of speed observations were available. As a result, the frequency of the speed differentials was sometimes less than the minimum frequency of the two relevant speed measures.

Table 5-1 Summary of Raw Data Statistics

Country	Variable	Unit	Min.	Max.	Mean	SD[a]	Freq.[b]
U.S.A.	Total collisions		0	29	4.4	6.0	139
	Collisions with severe injury		0	7	0.5	1.1	139
	Number of years of collision data		0.33	8	3.8	2.3	139
	Entering AADT	veh./day	220	19593	4637	3706	139
	Inscribed circle diameter, D_{INS}	feet	85.3	300	144.1	49.1	139
	Central island diameter, D_{INT}	feet	19.7	214	77.7	44.8	134
	Entry width, W_{ENL}	feet	12	49	22.2	8.3	138
	Circulating width, W_{CR}	feet	11.5	45	26.1	8.3	138
	Exit width, W_{EXL}	feet	12	51	23.0	8.5	128
	Approach speed	mph	18.2	52.2	34.5	8.8	36
	Speed Differential of approach vs. entry	mph	5.8	34.5	17.8	7.4	36
	Entry speed	mph	11.7	26.2	17.0	3.3	49
	Speed Differential of entry vs. circulating	mph	-1.5	6.6	2.2	2.2	39
	Upstream circulating speed	mph	10.1	23.7	15.1	3.6	40
	Speed Differential of circulating vs. exiting	mph	-9.8	0.3	-4.8	2.1	34
	Upstream exiting speed	mph	14.5	30.3	19.3	3.7	40
Italy	Total collisions[a]		0	252	24.7	49.3	34
	Collisions with severe injury[c]		0	65	6.8	12.9	34
	Number of years of collision data		0.8	8.5	6.2	2.1	34
	Entering AADT[c]	veh./day	5000	39000	20506	9455	34
	Inscribed circle diameter, D_{INS}	feet	54.1	529.9	151.7	92.1	34
	Central island diameter, D_{INT}	feet	13.1	492.0	98.1	93.0	34
	Entry width, W_{ENL}	feet	6.3	45.7	19.0	7.5	34
	Circulating width, W_{CR}	feet	17.2	36.7	26.0	5.4	34
	Exit width, W_{EXL}	feet	6.5	38.3	18.8	7.5	34
	Approach speed	mph	20.9	30.8	24.7	3.9	6
	Speed Differential of approach vs. entry	mph	0.4	10.0	4.6	4.0	6
	Entry speed	mph	14.3	23.8	20.0	3.2	6
	Speed Differential of entry vs. circulating	mph	1.5	5.5	2.9	2.0	6
	Upstream circulating speed	mph	12.8	18.7	17.2	2.3	6
	Speed Differential of circulating vs. exiting	mph	-10.5	0.4	-5.3	4.5	6
	Upstream exiting speed	mph	12.4	27.3	22.5	5.9	6

Note: a. standard deviation;

Note: b. frequency, number of sites with data;

Note: c. For Italian data, collisions and entering AADT are for the whole roundabout.

Figure 5-2 depicts the geometric characteristics and the approach locations where the speeds were obtained.

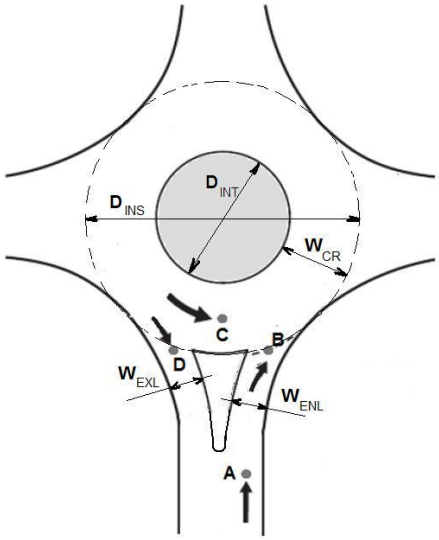


Figure 5-2 Geometric Characteristics of the Approach-level Area and Locations for Speeds

In Figure 5.2, location A depicts the approach speed (measured at least 200 feet upstream of the yield line) (Rodegerdts et al., 2007), B for entry speed, C for upstream circulating speed and D for upstream exiting speed. The U.S. speed data were collected by radar guns (Rodegerdts et al., 2007). For the Italian data, the collection of speed and positional data of the vehicles in traffic were acquired with a speed gun and a video camera. The acquisition system was placed in positions not visible to drivers and only isolated vehicles were considered, thereby excluding information that could be affected by factors such as those linked to the dynamics of traffic flow. In the case of video measurements, a high-speed digital video camera was employed. From the subsequent analysis of the captured frames, and with knowledge of the distance between the selected sections, the average speeds of isolated vehicles were calculated.

5.3.3 Derived Data

The average value of the measured entry, and upstream circulating and upstream exiting speeds were considered as a new speed variable (referred to as "approach average speed", AAS). Logically, the AAS represents the average operating speed inside or at the periphery of an approach. The sum of the absolute values of the three differentials in Table 5-1 (referred to as SDSum) could be used as another measure, which represents the overall level of the speed gaps.

The third derived variable is the speed differential of the approach vs. the AAS (referred to as SDApproachAAS). Table 5-2 illustrates the summary statistics of the derived speed measures.

In Table 5-2, SDSum is the addition of the three speed differentials from Table 5-1. The sum can only be obtained when all three speed differentials are available. Therefore, the frequency of SDSum is less than that of the individual speed differentials.

Table 5-2 Summary statistics of derived speed measurements

Country	Variable	unit	Minimum	Maximum	Mean	Standard Deviation	Frequency
U.S.A.	AAS	mph	12.4	25.7	17.0	3.1	33
	SDSum	mph	12.7	45.3	25.5	7.7	32
	SDApproachAAS	mph	5.8	35.4	18.1	7.8	32
Italy	AAS	mph	13.2	23.2	19.9	3.6	6
	SDSum	mph	5.9	25.0	12.9	7.5	6
	SDApproachAAS	mph	-1.2	11.2	4.8	4.6	6

5.4 SAFETY SURROGATE DEVELOPMENT VIA COLLISION PREDICTIVE MODEL

With respect to the synthesized analysis in Section 5.2, there are two premises for operational speed to act as a roundabout safety surrogate. First, operational speed must be either established as a key predictor in crash predictive modeling or practically converted into collision frequency or a severity measure. Second, operational speed itself must be well predicted from the roundabout design features.

In this section, the crash predictive model, also known as an SPF, will be investigated, which reflects Scenario 1 described in Section 5.2. Section 5.5 will focus on the indirect algorithm that reflects Scenario 2 described in Section 5.2.

5.4.1 Literature Reviews

Recent research that was presented in the NCHRP Report 572 (Rodegerdts et al., 2007) attempted to establish a speed-based approach-level SPF for U.S. roundabouts, with the following structure:

$$Crashes/\,year = \exp(\text{intercept}) \cdot AADT^{b} \cdot \exp(cX) \quad (5\text{-}3)$$

where

AADT=average annual daily traffic,

X=independent speed-related variable, and

b, c=calibration parameters.

However, the estimated model was deemed inadequate on the basis of the weak effects of the speed variables (Rodegerdts et al., 2007), so speed-based SPFs could not be recommended. In contrast, there were a number of successful non speed-based SPFs estimated for U.S. roundabouts in that research. These models were estimated at both roundabout and approach levels. Some of the approach level models contained geometric variables, but for the roundabout level models, the sum of entering AADTs from all approaches was typically the only variable.

Researchers have also developed SPFs for roundabouts in Great Britain, Australia, New Zealand and Sweden (FHWA, 2000; Turner et al., 2006; Brude and Larsson, 2000; Maycock and Hall, 1984). Some of these efforts include, in addition to traffic exposure, variables that reflect geometric features, configuration of vehicles, and speed features (85^{th} percentile speed, speed limits or relative speed difference). Research from New Zealand (FHWA, 2000) also introduced a model that relates speed features and factors such as diameter and visibility.

NCHRP Report 572 (Rodegerdts et al., 2007) also documented and tested the following speed prediction models documented in the FHWA Roundabout Guide (FHWA, 2000):

$V=8.7602 \cdot R^{0.3861}$, for e=+0.02

$$V=8.6164 \cdot R^{0.3673}, \text{ for } e=-0.02 \quad (5\text{-}4)$$

where

V=predicted speed for left-turn circulating, through circulating, exit or entry movements (km/h),

R=radius of vehicle path (m), and

e=super-elevation (m/m) (inner edge of curve is lower than the outer when e is positive).

Recent research by Bassani and Sacchi (2011) developed a multiple linear regression model for Italian roundabouts as shown in Equation 5-5.

$$V_{85} = 0.4433 \cdot D_{INT} + 0.8367 \cdot W_{CR} + 3.2272 \cdot W_{ENL} \quad (5\text{-}5)$$

where

V_{85} = 85-percentile operating speed at circulating roadway (km/h),

D_{INT} = diameter of the central island (m),

W_{CR} = width of the circulatory roadway (m), and

W_{ENL} = width of the entry lane (m).

The key aspects of these two speed prediction models are different. Equation 5-4 is fitted based on the fundamental functions of vehicle dynamics, while Equation 5-5 is an empirically derived function. The latter, according to the authors, was developed without a constant term to logically force an estimate of zero speed when there is a value of zero for all covariates. Moreover, Equation 5-5 pertains to an 85-percentile circulating speed while Equation 5-4 is presumed to pertain to a predicted circulating design speed.

In summary, the international research suggests that speed can be related to the safety performance of roundabouts. However, there is a wide spectrum of definitions for speed variables, especially in the European literature, with no clear indication of the best variable specifications.

The following sub-sections aim, accordingly, to address this issue by investigating and comparing possible choices of speed variables, and to make a recommendation for the optimal choice, with design features as inputs.

These sub-sections then investigate the development of a roundabout SPF with predicted speed as the key input.

5.4.2 Selection and Estimation of Speed Prediction Model

For a speed measure to be representative of the design features in an approach-based crash prediction model, it must be reflective of the speeds in the vicinity of the approach. In earlier research (Chen, 2010; Chen et al., 2011), the authors tried to model individual speeds and speed differentials, but this proved to be fruitless.

The final determination of the most appropriate measure was achieved by running an "effect (variable) selection" procedure within the framework of generalized linear models (GLMSELECT Procedure) with the SAS software (SAS Institute Inc., 2012). Based on a pre-set group of variables, the procedure of "effect selection" iterates the entry or removal of effects until the selection stops at a minimum value of the model optimization criterion (the Schwarz Bayesian information criterion, SBC).

The effect selection procedures were conducted with AAS, SDSum and SDApproachAAS as the response variables, and six features as the impact factors, including country code (C_{CODE}, defined as a categorical variable for the U.S. or Italian data), inscribed circle diameter (D_{INS}), central island diameter (D_{INT}), entry width (W_{ENL}), circulating roadway width (W_{CR}), and exit width (W_{EXL}). This process first excluded SDSum and SDApproachAAS since even the optimized final models with these measures have very low coefficients of determination (R^2), as shown in Table 5-3. The results are more favorable with AAS as the response variable, as shown in Table 5-4.

Table 5-3 Speed Predictive Model with SDSum and SDApproachAAS as Responses

Items	SDSum Model	SDApproachAAS Model
Full collection of covariates	C_{CODE}, D_{INS}, D_{INT}, W_{ENL}, W_{CR}, and W_{EXL}	C_{CODE}, D_{INS}, D_{INT}, W_{ENL}, W_{CR}, and W_{EXL}
Covariates for selected model	C_{CODE}	C_{CODE}
Adjusted R^2	0.2559	0.2972

Table 5-4 Basic Parameters of Speed Predictive Models Considered with AAS as Response

Items	Effect Selection Run 1	Effect Selection Run 2
Full collection of covariates	C_{CODE}, D_{INS}, D_{INT}, W_{ENL}, W_{CR}, and W_{EXL}	C_{CODE}, D_{AV}, and W_{AV}
Covariates for selected model	C_{CODE},, D_{INT}, and W_{EXL}	C_{CODE}, D_{AV}, and W_{AV}
Adjusted R^2	0.8432	0.8297
Representativeness	Narrow	Wide
Final Recommendation		✓

As seen in Table 5-4, the results initially indicate that only central island diameter and exit width should be included (as indicated by the results for "Effect Selection Run 1"). However, a better model is obtained when the procedure was repeated with the average of the inscribed circle and central island diameters (D_{AV}), and the average of entry, circulating and exit widths as the derived variables (W_{AV}). The parameters of the final model are indicated by the column labeled "Effect Selection Run 2" in Table 5-4, while the model results are shown in Table 5-5.

While both runs in Table 5-4 have satisfactory statistical performance in terms of adjusted R^2, the covariates of "Effect Selection Run 2" provide greater representativeness. Consequently, the modeling of "Effect Selection Run 2" is considered to be preferable.

Table 5-5 Results for Recommended Speed Predictive Model

Model form: $AAS = a+b_0+b_1 \cdot D_{AV}+b_2 \cdot W_{AV}$				
Parameter	Estimate	Standard Error	t value	Pr > |t|
a	13.015958	0.775832	16.78	<0.0001
b_0: U.S.A.	-3.088964	0.620977	-4.97	<0.0001
: Italy	0			
b_1	0.034074	0.009867	3.45	0.0015
b_2	0.142936	0.053786	2.66	0.0118

The summary statistics of the predicted speeds with this model are shown in Table 5-6 for the larger dataset of 138 U.S. and 34 Italian roundabout approaches. The distribution of these predicted speeds is shown in Figures 5-3 and 5-4.

Table 5-6 Summary statistics for predicted speeds

Country	Variable	Unit	Min.	Max.	Mean	Standard Deviation	Frequency
U.S.A.	Predicted AAS	mph	14.2	24.5	17.2	2.5	138
Italy	Predicted AAS	mph	15.5	35.6	20.3	3.7	34

Figure 5-3 demonstrates how the model for predicting AAS can expand sample data size by allowing the inclusion of sites without observed speeds. Figure 5-4 only pertains to the 39 sites with both speeds, and is intended to visually illustrate the accuracy of the speed prediction model in Table 5-5.

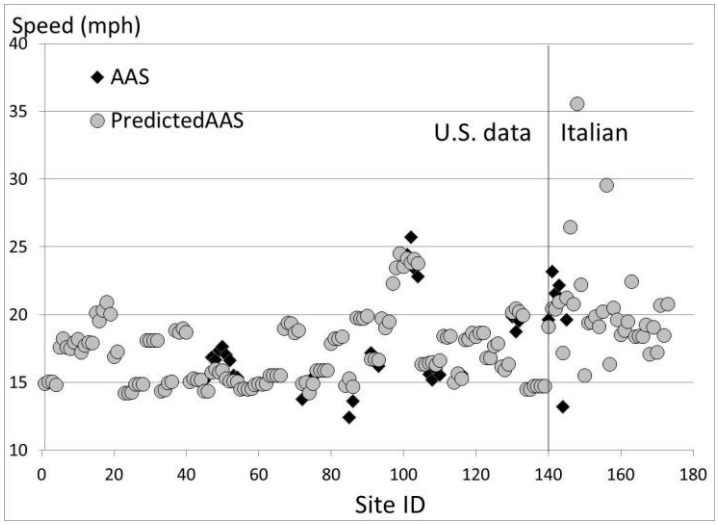


Figure 5-3 Predicted versus Observed AAS for All Sites

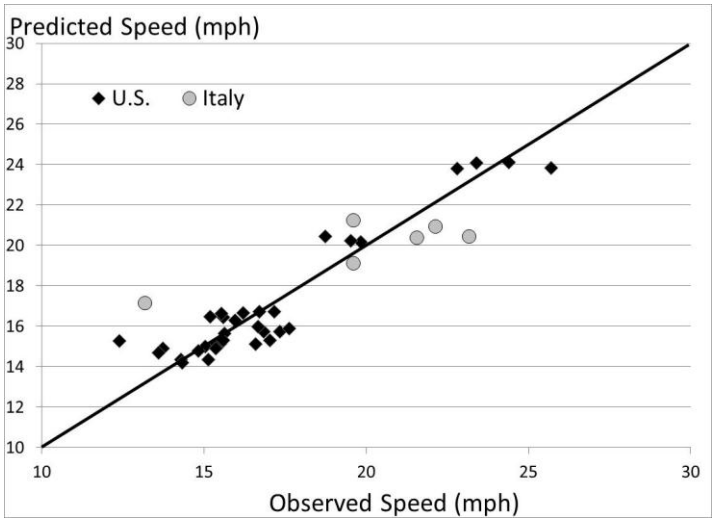


Figure 5-4 Predicted versus Observed AAS for 39 Sites with Both Speeds

5.4.3 Selection and Estimation of Safety Models Based on Predicted Speed

A. Model selection for SPF based on predicted speed

a. Model identification

Collision counts, which are non-negative integers with over-dispersion, are well described by a mixed Poisson distribution family (Chou and Steenhard, 2009). The special case of the Poisson-gamma distribution – an NB distribution model, which was introduced in Section 3.4.2, is addressed in this chapter (Chou and Steenhard, 2009; Lord et al., 2010*;* Persaud et al., 2010a).

b. Determination of functional form

The form of the roundabout speed-based SPF can be generally specified as follows:

$$Y_{it}|\mu_{it} \sim Po(\mu_{it})$$

$\mu_{it}=f(AADT, AAS_{predicted}; \beta_0, \beta_1, \beta_2)$

Dispersion parameter, $\alpha = \frac{k}{e^{(\gamma_0 \cdot AAS_{predicted})}}$ (5-6)

where

f is the functional form of the mean of the collisions, which is shown in Table 5-7, and *k* is the theoretical maximum threshold of the dispersion parameter.

Table 5-7 Possible Function Forms of μ_{it}

AAS / AADT	Power Function	Gamma Function	Exponential Function
Power Function	$\beta_0 \cdot \text{AADT}^{\beta_1} \cdot (\text{AAS}_{\text{predicted}})^{\beta_2}$	$\beta_0 \cdot \text{AADT}^{\beta_1} \cdot (\text{AAS}_{\text{predicted}})^{\beta_2} \cdot \exp(\beta_3 AAS_{predicted})$	$\beta_0 \cdot \text{AADT}^{\beta_1} \cdot \exp(\beta_2 AAS_{predicted})$

While a power function is universally used for AADT, the power, gamma or exponential functions could be considered for predicted AAS (Persaud et al., 2002; Rodegerdts et al., 2007; Chou and Steenhard, 2009), as depicted in Table 5-7. The dispersion parameter is specified as varying rather than constant to improve the fit, in accordance with current research thinking (Lord et al., 2010).

The empirical integral function (EIF) analysis (Hauer and Bamfo, 1997; Persaud et al., 2002) was applied to preliminarily identify the functional form of AAS. This approach plots a cumulative probability graph against one specific covariate, and then compares this empirical cumulative curve with standard curves of typical function forms in order to indicate the most appropriate relationship between the dependent variable and the candidate covariate.

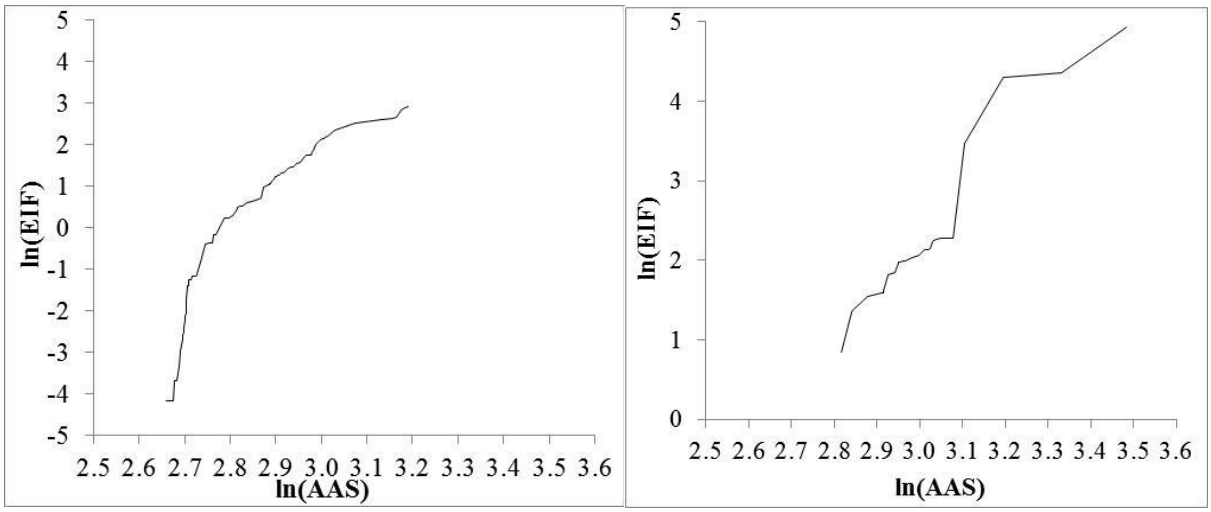
Figure 5-5 shows the results of the EIF analyses for the U.S. and Italian data. Here, the ordinate is the logarithm of the relevant original ordinate of the EIF curve, such that, if the plotted line is linear, the power function is suggested. For the U.S. data, a reasonably linear trend can be seen from ln(AAS) = 2.7 (i.e., for AAS approximately 24 km/h (14.9 mph) and above, which is consistent with the range of predicted AAS in the U.S. as shown in Table 5-6, suggesting that a power function for AAS may be appropriate. For the Italian data, the trend is globally linear, which indicates that a power function for AAS is not inappropriate for Italy.

Notwithstanding this preliminary indication, all of the forms in Table 5-7 were actually attempted, with the final recommendation based on a comparison of model performance.

c. Modeling approach and platform

A Bayesian modeling approach, which combines prior knowledge from a reference population and information from site-specific observations, is the most appropriate in maintaining consistency with this research. This is because model parameters in this approach are treated as unknown random variables with inferences based on the posterior distributions of the parameters. This gives more flexibility than fixed parameter estimations (Persaud et al., 2010a), a property that will increase the odds of attaining a successful model.

A general purpose Markov chain Monte Carlo (MCMC) simulation procedure was applied in the SAS (SAS Institute Inc., 2012) program for Bayesian model estimations. The optimization method and significant criteria are the same as those described at Section 4.4.2.



U.S. data Italian data

Figure 5-5 Empirical integral function analysis for AAS

The Bayesian Poisson-gamma approach was not successful for the Italian data so alternative approaches were considered instead.

B. Model estimation for SPFs based on predicted speed

a. Bayesian Poisson-gamma model for U.S. data

Prior distributions for all of the parameters (β_0, β_1, β_2, γ_0) are assumed to be non-informative N(0, 10^3) to reflect the lack of precise knowledge of the coefficient values (Persaud et al., 2010a).

In the case of multiple significant outputs, the preferred model is provided by a Bayesian model selection indicator, i.e., deviance information criterion (DIC) (Claeskens and Hjort, 2009; SAS Institute Inc., 2012; Spiegelhalter et al., 2002), It is frequently used in the Bayesian analysis of many parameters in complex models, and where its computation is typically an easy consequence of output from MCMC simulations (Claeskens and Hjort, 2009). The DIC, a Bayesian model selection indicator is defined as (SAS Institute Inc., 2012):

$$DIC = \overline{D(\theta)} + pD = D(\bar{\theta}) + 2pD \quad (5\text{-}7)$$

where

$\bar{\theta}$ is the posterior mean of parameter θ,

$\overline{D(\theta)}$ is the posterior mean of the deviance,

$D(\bar{\theta})$ is the deviance evaluated at $\bar{\theta}$, and

pD is the effective number of parameters.

In SAS, a smaller DIC indicates a better fit to the data set (SAS Institute Inc., 2012).

For the U.S. data, models with power and exponential forms for the AAS fulfilled the two above-mentioned criteria, while the model with the gamma form did not. On the basis of the DIC (652.641 for the power function, 652.923 for the exponential function), the model with the power form is the best, albeit by a small margin. As a result, the model with a power form for AAS is recommended. Details are listed in Table 5-8. These indicate that the mean absolute deviation (MAD) has the same order as the average observed collisions per year (=1.54), another indication that the model is basically adequate.

b. Alternative modeling for the Italian data

As noted above, the Bayesian Poisson-gamma modeling approach is not suitable for the Italian data, based on all assessment criteria, so an alternative model is required. After unsuccessful attempts with the usage of other Bayesian mixed Poisson family members, zero-inflated Poisson (ZIP) and zero-inflated negative-binomial (ZINB) models were considered. The ZIP model was found to be the most suitable:

$$P(y_{it}=0|x_i) = F_i + (1-F_i)\exp(-\mu_{it})$$
$$P(y_{it}|x_i) = (1-F_i)\frac{\exp(-\mu_{it})\cdot\mu_i^{y_{it}}}{y_{it}!}, \quad y_{it}>0 \qquad (5\text{-}8)$$

where

y_{it}, μ_{it}, x_i are the same as those in Equations 3-8, 3-9 and 3-10,

$\mu_{it} = \beta_0 \cdot \mathrm{AADT}^{\beta_1} \cdot (\mathrm{AAS}_{\mathrm{predicted}})^{\beta_2}$ or $\mu_{it} = \beta_0 \cdot \mathrm{AADT}^{\beta_1} \cdot \exp(\beta_2 \cdot \mathrm{AAS}_{\mathrm{predicted}})$, and

F_i is the probability that y_{it} =0.

Table 5-8 Estimation of recommended Bayesian Poisson-gamma SPF for U.S. data

Posterior Statistics, $\mu_{it} = \beta_0 \cdot \mathrm{AADT}^{\beta_1} \cdot (\mathrm{AAS}_{\mathrm{predicted}})^{\beta_2}$				
Parameter	Mean	Standard Deviation	Monte Carlo Standard Errors	Heidelberger-Welch Diagnostics
$\ln\beta_0$	-16.3755	2.1801	0.2991	PASS
β_1	0.5094	0.1245	0.0122	PASS
β_2	4.3314	0.7765	0.0899	PASS
Dispersion parameter, $\alpha = \frac{k}{e^{(\gamma_0 \cdot AAS_{predicted})}}$ =3/exp(0.0618*AAS$_{\mathrm{predicted}}$)				
DIC (smaller is better) = 652.641				
Mean Absolute Deviance, MAD [a] $= \frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$ =1.22				

Note: a. n is the sample size and y_i and $\hat{y}_i$ are the observed and estimated mean values at site i respectively (Lord et al., 2010).

ZIP modeling was accomplished with the "COUNTREG procedure" by using SAS (SAS Institute Inc., 2012). A model is considered significant when it required all parameter estimates

to be significant as determined by their p-values at the 10% level. Multiple significant models were ranked by Akaike's information criterion (AIC) (Claeskens and Hjort, 2009), defined as:

$$AIC(M) = 2\log\text{–}likelihood_{\max}(M) - 2\dim(M) \quad (5\text{-}9)$$

where *dim* is the number of estimated parameters for model *M*.

In SAS, a smaller AIC indicates a better model (SAS Institute Inc., 2012). Since the ZIP model with a power function for AAS had a smaller AIC (=782.5) than the model with the exponential function for AAS (=786.0), it is recommended as most suitable for the Italian data, and shown in Table 5-9. As seen from Table 5-9, the MAD has the same order as the average observed collisions per year (=3.90), another indication that the model is fundamentally adequate.

Table 5-9 Estimation of alternative ZIP SPF for Italian data

Parameter Estimates, $\mu_{it} = \beta_0 \cdot \text{AADT}^{\beta_1} \cdot (\text{AAS}_{\text{predicted}})^{\beta_2}$				
Parameter	Estimate	Standard error	t value	Pr > |t|
$\ln\beta_0$	-29.5239	1.2224	-24.15	<0.0001
β_1	2.8623	0.1302	21.98	<0.0001
β_2	0.6339	0.1628	3.89	<0.0001
AIC(smaller is better) = 782.5364				
Mean Absolute Deviance, MAD = 2.67				

5.4.4 Discussion of Model Results

a. Assessment of practical validity

The practical significance of the predicted speed-based SPF was further assessed by comparing the implied CMFs for changes in average speed with those recently derived and presented in NCHRP Report 617 (Harkey et al., 2008). The CMFs in the report were related to initial mean travel speed and change in mean travel speed. The speed related CMFs implied from the models in this dissertation are estimated from:

$$\text{CMF}_{\text{speed}} = (\text{AAS}_{\text{changed}} / AAS_{initial})^{\beta_2} \quad (5\text{-}10)$$

where the calibrated value of β_2 is given in Table 5-8 or 5-9.

Figure 5-6 plots CMFs of models found in Table 5-8 or 5-9, along with those found in NCHRP Report 617. This comparison shows that the CMFs provided by the U.S. SPF for total crashes are reasonably consistent with those in the NCHRP report, even though the latter pertains to FI crashes. On the other hand, CMFs provided by the Italian SPF are not as consistent, which is understandable, considering that the Italian driving environment is different from that in the U.S.

B. Model Comparison

Notwithstanding the differences in random structure, the model functional forms for the U.S. and Italy data are consistent. They each have power functions for both AADT and AAS. This will give practitioners a consistent structure for roundabout safety models, as well as a unique safety surrogate.

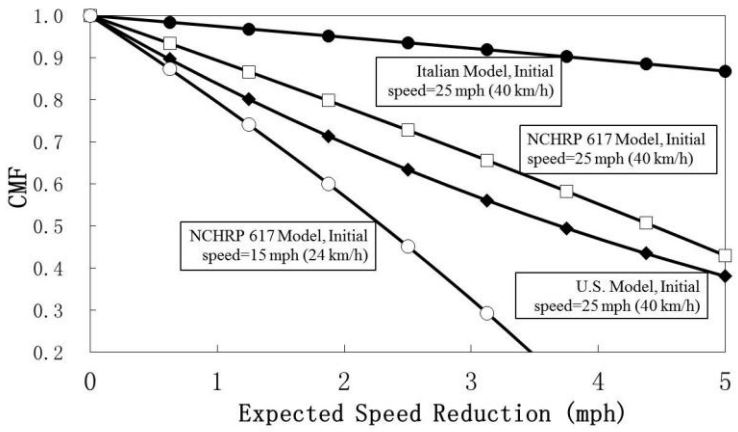


Figure 5-6 Comparisons of implied CMFs with CMFs in NCHRP Report 617

The difference between the U.S. and Italian models lies in the coefficients. Compared with the Italian SPF, the U.S. SPF has a smaller coefficient for the AADT but a larger coefficient for the AAS. This may indicate that the U.S. SPF shows more sensitivity for any changes in the AAS as opposed to the Italian model. This is also shown by Figure 5-7 which plots the curves of the predicted collisions versus the AAS at the level of the mean AADT for both countries.

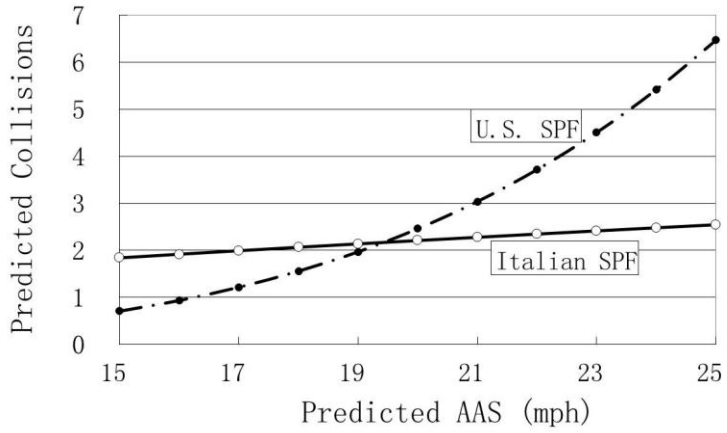


Figure 5-7 Curves of Predicted Collisions versus AAS (at the level of mean AADT)

5.5 SAFETY SURROGATE DEVELOPMENT VIA NON-MODELED INDIRECT CONVERSIONS

In terms of Scenario 2 described in Section 5.2, where a surrogate-based collision predictive model is not available, an alternative indirect method is needed to transfer the knowledge of surrogates into collision measures.

The basic idea is to estimate the rankings of collisions from those of surrogates, and then to convert the former into collision estimations, as shown in Figure 5-8.

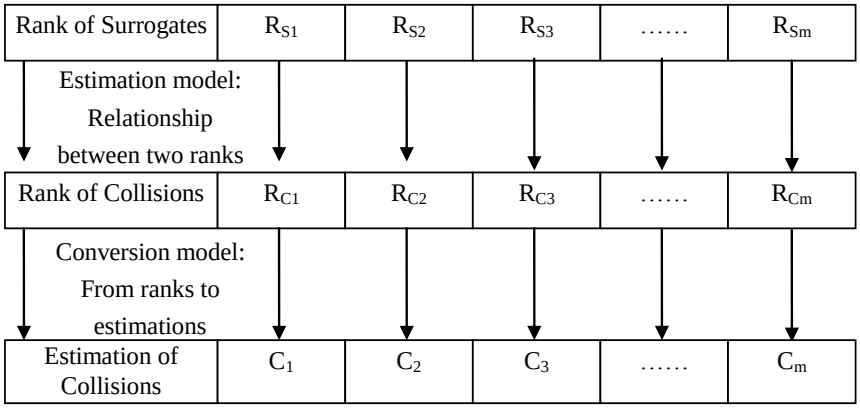


Figure 5-8 Process for Non-modeled Indirect Conversions of Collisions to Surrogates

To carry out this process, two steps are essential: the first step is to prove and establish a relation between the surrogates and collision rankings; the second step is to estimate the collisions from their rankings.

5.5.1 Evaluation of Correlation Strength between Surrogate and Collision Measures

Gettman et al. (2008) applied the "Spearman rank correlation coefficient" to evaluate the correlation strength of surrogate measures and corresponding predicted crash frequencies. This coefficient is estimated from the rankings of yoked pairs of predicted crash frequencies and surrogate measures under specific AADT levels by the following function (Gettman et al., 2008; WikiHow, 2012):

$$R_S = 1 - 6\frac{\sum d^2}{N(N^2-1)} \quad (5\text{-}11)$$

where,

d = rank of surrogate measure – rank of predicted crash frequency, and

N = number of paired ranks.

Then, the resulting correlation coefficient is compared with the critical coefficient value in an appropriate sample size and at an appropriate significance level. If the absolute value of the coefficient is greater than the critical value, then it can be concluded that there is a rank order relationship between these samples. If the R_s value is -1, then there is a perfect negative correlation between the two sets of data. If the R_s value is 1, then there is a perfect positive correlation between the two sets of data (Gettman et al., 2008).

In this dissertation, the correlation strength of the surrogate measure - predicted AAS - and observed annual crash rate is investigated by a Spearman rank correlation coefficient. The same U.S. and Italian datasets used in Section 5.4 are applied here, except that no speed-based safety model is established. While this does not produce the solid outcomes of Section 5.4, the intent is to merely explore the validity of this research.

The process in Figure 5-8 is accomplished by using 5 steps:

Step 1: 138 U.S. and 34 Italian sites with predicted AAS as described in Table 5-6 are sorted by their AADTs, from the lowest to the highest;

Step 2: the originally observed annual crash rates are ranked from lowest to highest and ordered from 1 to 172. If two or more records in the data have the same value, the mean of the rankings are to be determined in the way that these records had been originally ranked, and then the data are ranked with this mean (WikiHow, 2012);

Step 3: the original predicted AAS are ranked from lowest to highest and ordered from 1 to 172. If two or more records in the data have the same value, the mean of the rankings are to be determined in the way that these records had been originally ranked, and then the data are ranked with this mean (WiliHow, 2012);

Step 4: d = rank of surrogate measure – rank of predicted crash frequency is calculated; and

Step 5: the Spearman rank correlation coefficient is estimated with Equation 5-11.

After carrying out this 5-step process, the Spearman rank correlation coefficient for the predicted AAS and observed collisions is calculated as R_s=0.4894. According to Zar (1972), for N=172, the use of a Student's t-test provides an excellent way to test the significance of R_s by:

$$t = \frac{R_S}{\sqrt{\frac{(1-R_S^2)}{N-2}}} = 7.3172 \qquad (5\text{-}12)$$

For N=172, with a 95% confidence level (α=0.05), the critical t-test value is 1.974 (Dougherty, 2002; StatTools Home Page, 2012). The absolute t value is much greater than the critical t value, so it can be concluded that there exists a strong rank order relationship between surrogate - predicted AAS - and observed collision.

5.5.2 Model Applied to Convert Surrogate Rankings to Collision Rankings

Notwithstanding the proven correlation strength between the rankings of surrogates and collisions, no exact mathematical functions could be established for this correlation. Further regression needs to be accordingly conducted.

It was preliminarily indicated from the sample U.S and Italian data that the correlation modeling between surrogate and collision will not result in a simple linear or well-known non-linear model. As revealed in Figure 5-9, scatter plotting the surrogate and collision rankings does not reveal any pattern.

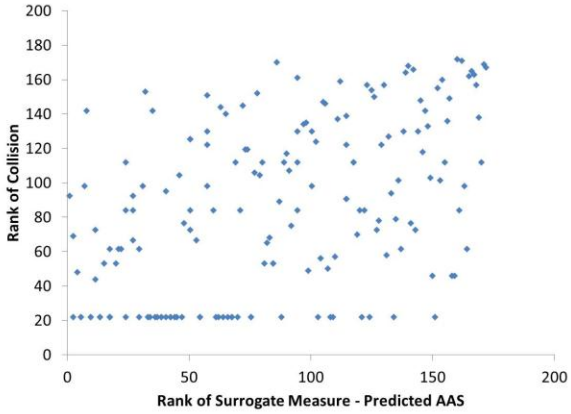


Figure 5-9 Scatter Plotting of Surrogate vs. Collision Rankings

Other than direct linear or non-linear regressions, other means to develop a model that converts surrogate rankings into collision rankings were attempted. A GLM model structure, similar to that used by Boonsiripant (2009) in attempting to estimate collisions from surrogates, was conducted in this study with an NB distribution. The model was developed with the statistical software of R (The R Foundation for Statistical Computing, 2012). Here, the independent variable is the ranking of the surrogate – the predicted AAS, which was generated in accordance with Step 3 in Section 5.5.1, whereas the response variable is the collision ranking, generated by Step 2 in Section 5.5.1.

The estimated model is shown in Table 5-10. Since both independent and response variables in this model are the "mean" ranks generated by the aforementioned five steps, these variables are not exactly ordinal integer variables anymore. So, a regression model for continuous variables was adopted in this case.

Table 5-10 Estimation of GLM Model Converting Rank of Surrogates to Rank of Collisions

Rank of Collisions, $R_{Ci} = \alpha(\text{Rank of Predicted AAS})^{\beta}$				
Parameter	Estimate	Standard error	z value	Pr >|z|
α	29.649	0.21058	16.096	< 2E-16
B	0.24970	0.04919	5.076	3.85E-07

5.5.3 Algorithm to Estimate Collision Values from Collision Rankings

For a local region, the rankings of collisions may be estimated by the model in Table 5-10, which is imported from external sources as prior knowledge. Then, these rankings need to be converted into collision values for the convenience of local practices. A special algorithm needs to be developed to achieve this conversion.

As a simplified attempt, the collision value, Ci, can be estimated from its ranking, R_{Ci} by:

$$C_i = C_{max} \times \frac{R_{Ci} - R_{Cmin}}{R_{Cmax} - R_{Cmin}} \quad (5\text{-}13)$$

Without specific statistical processing and testing involved in this simplified solution, the final estimated C_i is just an approximation rather than a statistically significant estimation. Notwithstanding the statistical inadequacy, this algorithm achieves four objectives, as follows.

1) The minimum estimated rank (R_{Cmin}) leads to the minimum number of collisions in a region (C_{min}), generally 0,
2) The maximum estimated ranking (R_{Cmax}) leads to the maximum number of collisions in a region (C_{max}),
3) Estimation of the same ranking leads to the same collision value, and
4) A greater estimation of ranking leads to a larger collision value, and vice versa.

That is to say, this algorithm is a sufficient solution for defining the "outer boundaries" and conveys appropriate scalar estimates of sequential collisions for a local region.

5.6 CHAPTER CONCLUSION

This chapter has contributed to research on ways to apply safety surrogates as alternative sources of prior knowledge for a local region. The sample facility used was the roundabout and the example surrogate is an operational speed measure – the predicted AAS.

The employment of surrogates as sources of prior knowledge comprises two scenarios: the availability of surrogate-based SPFs or the unavailability of such.

For the former, surrogate-based SPFs – roundabout SPFs based on predicted speed are developed for local regions. As for the sample facility – roundabouts, a speed predictive Bayesian Poisson-gamma SPF has been successfully developed for the sample U.S. data, while for the Italian data, an alternative ZIP model provides a better fit. Regardless of the structural differences, the developed U.S. and Italian SPFs both suggest that an approach that uses predicted speed seems to be promising for indirect estimations of roundabout safety performance from a model that first predicts speeds. Besides that, their functional forms for both AADT and speed variables are consistent. Hence, this research has successfully established a tool with "two-stages" that utilizes a roundabout safety surrogate: the first stage is to predict the AAS from geometric features and the second is to use the predicted AAS to estimate collisions.

At least three major advantages of this two-stage approach have been highlighted by the research documented in this chapter: first, as revealed during the statistical analyses, geometric features performed better in predicting speed than in predicting crashes directly; in addition, safety performance models involving only a speed measure rather than a variety of geometric variables were much more parsimonious and had higher statistical significance; finally, this approach is conceptually attractive in that predicted speed can be used rather than observed speed, thereby expanding the sample size and producing more robust models.

When a direct, surrogate-based SPF is not available, an indirect algorithm that converts surrogates into collision measures should be applied instead. The outcome obtained in this chapter can be summarized into a three-stage method:

Stage 1 –surrogate ranking – for the sample facility, it is based on predicted AAS – per Step 3 in Section 5.5.1,

Stage 2 – collision ranking estimated from surrogate rankings per the model described in Table 5-10, and

Stage 3 – collision ranking converted into collision measures by using Equation 5-13.

It should be specifically mentioned that in this chapter, the methodology for the first scenario (surrogate-based SPFs available) is more preferable than the methodology for the second scenario (surrogate-based SPFs unavailable). The former one satisfies scientific preciseness and statistical significance while the latter one doesn't. However, the latter algorithm is sufficient enough to define the "outer boundaries" and rank appropriate sequences of collisions for a local region. It is recommended that practitioners in local regions apply these two methodologies accordingly: if surrogate-based SPFs are available, then the relevant models are applied; otherwise, the algorithm is used as an alternative.

CHAPTER 6 KNOWLEDGE INTEGRATION BY BAYESIAN MODEL AVERAGING

Serial methodologies for conducting data sampling, importing external models, developing local collision predictive models, and applying safety surrogates have already been established in Chapters 2 to 5. Reference knowledge sources for the before-after evaluation (B/A) framework shown in Figure 1-2 have been constructed. The missing component of this framework is a way to integrate all of these knowledge sources, either imported or locally developed, to provide a unified reference for a B/A procedure.

Since the source of prior knowledge in a B/A process is the model, the integration of knowledge sources is practically the integration of models. This chapter will describe how model integration is accomplished through "model averaging" for application in a B/A scenario. A paper based on this research has already been published (Chen & Persaud, 2011).

This research documentation in this chapter is in six sections. Section 6.1 is for conceptual investigation of the basic ideas, key issues of model selection and model averaging, as well as past research reviews. Then, as presented in Section 6.2, methodologies of Bayesian Model Averaging (BMA) are investigated, including comparison between model selection and model averaging, theory and functions to be applied for next step model averaging applications. Section 6.3 is devoted to the selection of candidate models to be included into BMA process. Section 6.4 documents the conduct of a BMA process for all of the datasets described in Chapter 2. GOF tests are documented in Section 6.5 to support the BMA as an innovative application compared with conventional approaches for model selection. Finally, Section 6.6 summarizes this chapter.

6.1 BASIC CONCEPTS AND RELEVANT PAST RESEARCH

This section states the rationale for the research in this chapter and establishes a theoretical foundation for the following statistical analysis.

Two generations ago, setting up and analyzing a single model were challenging tasks by themselves, so researchers rarely went to the trouble of analyzing the same data via several alternative models (Claeskens and Hjort, 2009).

From the 1970s, the evolution of statistics, accompanied by the innovation of computing technology developments, resulted in a much longer list of candidate models that could be fitted to a data set. Since then, the need has been established for methods to integrate model fits (Claeskens and Hjort, 2009). However, in the road safety domain, model selection and synthesis practices are still much less prevalent compared to model development itself. As a result is is necessary to study theoretical basics and relevant past practices from other engineering professions to obtain fundamental ideas for the selection and synthesis of safety models, given a variety of candidate models.

6.1.1 Key Issues on Model Selection and Averaging

In the beginning of their book, Claeskens and Hjort (2009) provided the key general considerations involved in comparing, selecting or combining models as follows.

1) Models are approximations.
2) A balance between over-fitting and under-fitting.
3) Parsimony.
4) The context.
5) The focus.
6) Conflicting recommendations.
7) Model selection versus model averaging.

The text that follows provides elaboration on these issues.

1) Nature of approximation for models

When dealing with either model development or selection, modelers are not able to guess the "correct" or "true" model that is in the background and almost always unknown. Instead, modelers are working on an "almost-as-good, useful" model. This is the fundamental awareness from which most methods on model development and selection start.

2) Balance between over-fitting and under-fitting

This is about the trade-off between variance and bias. Fewer estimated parameters lead to lower variability but higher bias, while more estimated parameters result in higher variability but lower bias. A proper balance between over-fitting (too many parameters) and under-fitting (too few parameters) must be sought.

3) Parsimony

Parsimony is an important principle followed by most model development and selection methods. It requires that only parameters that really matter ought to be included in a model, such that inclusion of an extra term, or adding complexity, is only worthy if predictive ability is materially improved.

4) The context

Modeling and model selection are rooted in an appropriate scientific context and contexts could change from case to case. As an example, Breiman (2001) discussed "two cultures" of statistics: one is pro-prediction and classification (favoring even a "black box" model as long as it works well) while another is "deeper learning about models" (favoring discovery of a non-null parameter even if it might not help improve the inference precision).

5) The focus

In applied statistics pertinent to the road safety domain, it is common that some estimators in a model are more important than others. Hence, it is also normal that different model development and selection efforts with different criteria or aims lead to different recommendations. This is not paradoxical, since there are different preferences and loss functions. For example, some model development or selection methods focus on specific candidate estimators over other parameters.

6) Conflicting recommendations

This principle challenges efforts to seek the "best" model since different criteria or preferences may lead to different recommended best models.

7) Model selection versus model averaging

Model selection strategies work by assigning a certain score to each candidate model. In some cases, there might be a clear winner, but sometimes these scores might reveal several candidates that do almost as well as the winner. In such cases, there may be considerable advantages in combining inference output across these best models, rather than eliminating all but one of them.

8) Why these key issues matter to the before-after evaluation process

One important message that Claeskens and Hjort (2009) conveyed via the aforementioned seven key issues is that the question of which is the "true" or "best" model is not well posed. Model selection is a matter of choice, and pros and cons exist regardless of the chosen model. With the versatility of candidate models, any criterion for model comparison addresses some aspects while neglecting others. What is important is to find a balance among these seven key issues, rather than considering any one single issue.

The combining of inference outputs is more meaningful than a single inference source, especially for a B/A process. It stands to reason, then, that model averaging may be a viable alternative to the recommendation of a single preferred model.

6.1.2 Most Popular Criteria for Model Selection

The above seven key issues can be synthesized into two major aspects for scoring a model: one is accuracy of approximation, and another is the quality of the model itself. Regardless of the abundance of variants, the essence of model selection criteria is almost consistently constituted of two components. One component is rooted in the maximum log-likelihood, which is a gauge of model fit, while the other component is derived from the number of estimated parameters, which are the indicators of model size.

Most (but not all) selection methods are defined in terms of an appropriate "information criterion", a mechanism that uses data to give each candidate model a certain score. This then leads to a fully ranked list of candidate models, from the ostensibly best to the worst (Claeskens and Hjort, 2009). Below is a discussion of four commonly used information criteria.

A. Akaike's Information Criterion

Among these "information criteria," Akaike's information criterion (AIC) is one of the most popular and versatile strategies, as indicated in Section 5.4.3.

In some practices, like model fitting with SAS software (SAS Institute Inc., 2012), a finite-sample corrected version of AIC is:

$$AIC(M) = 2log - likelihood_{max}(M) - \frac{2ndim(M)}{n-dim(M)-1} \quad (6\text{-}2)$$

where
dim(M) is the length of the parameter vector of model M, and
n is the sample size of the data.

B. Bayesian Information Criterion

Schwarz (1978) and Akaike (1977, 1978) formulated the evolution of the AIC into the Bayesian information criterion (BIC) with a more severe penalty for model complexity. Its penalty is equal

to the logarithm of the sample size times the number of estimated parameters in the model. The BIC format is:

$$BIC(M) = 2log - likelihood_{max}(M) - (logn)dim(M) \quad (6\text{-}3)$$

where

dim(M) is the length of the parameter vector of model M, and

n is the sample size of the data.

C. Deviance Information Criterion

The deviance information criterion (DIC) (Claeskens and Hjort, 2009; SAS Institute Inc., 2012; Spiegelhalter et al., 2002) is used as a comparison with the BIC for a Bayesian solution. This has been introduced in Section 5.4.3.

D. Focused Information Criterion

The viewpoint for model selection expressed via the focused information criterion (FIC) is that the 'best model' should depend on the parameters under focus or of interest, such as the mean or the variance, or the particular covariate values, etc. Thus, the FIC allows and encourages different models to be selected for different parameters of interest (Claeskens and Hjort, 2009).

Consider a focus parameter, $\mu = \mu(\theta, \gamma)$, i.e. a parameter of direct interest, which is to be estimated with good precision. For the estimation, all components of θ are to be included in the model, but it is not clear which components of γ should be included when forming the final estimate. Perhaps all γ_j shall be included, perhaps none. This leads to the consideration of estimators in the form $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_0, S^C)$. The 'best' model for estimation of the focus parameter μ is the model for which the mean squared error of $\sqrt{n}(\hat{\mu}_S - \mu_{true})$ is the smallest (Claeskens and Hjort, 2009).

By adopting this convention, assume $(\hat{\theta}, \hat{\gamma})$ signals the maximum likelihood estimation in the full p + q-parameter model. Let $D_n = \sqrt{n}(\hat{\gamma}_S - \hat{\gamma}_0)$, the FIC score, for each of the sub models indexed by S, which may be defined as (Claeskens and Hjort, 2009):

$$FIC(S) = \hat{\omega}^t\big(I_q - \hat{G}_S\big)D_n D_n^t (I_q - \hat{G}_S)^t\hat{\omega} + 2\hat{\omega}^t\hat{Q}_S^0\hat{\omega} \quad (6\text{-}5)$$

where

ω is the vector of length q that appears in the asymptotic distribution of estimators under local misspecification,

Iq is the identity matrix with a size of q × q,

G_S is the matrix with a dimension of q × q, which is related to J, and the expected value of minus I(Y, θ0), often partitioned into four blocks. Here, I(Y, θ_0) is the second derivative of the log-likelihood with respect to θ, and

Q_S is the lower right-hand |S| × |S| matrix of the inverse information matrix J_S^{-1} of the sub-model S; here, |S| is the number of elements of S, then let $Q_S^0 = \pi_S^t Q_S \pi_S$, where π_S is the |S| × q projection matrix that maps a vector v = (v_1, . . . , v_q)^t to v_S = π S_v which contains only the v_j for which j ∈ S.

6.1.3 Relevant Past Research

Verkuilen (2009) briefly reviewed model selection and averaging, while pinpointing the "plethora" of the term "information criteria (IC)", including AIC, BIC, DIC, and FIC, as well as the corrected AIC (AICC). While the AIC chooses one model for efficiency, and the BIC chooses another for consistency, the FIC may be constructed to pick models that do well on a particular model output of importance to the investigator. Meanwhile, Verkuilen (2009) mentioned that model averaging also uses IC to make a better predictive model than simply choosing the best model according to one's favorite criterion.

Ando and Tsay (2010) specifically contributed to this body of research with a predictive likelihood to overcome the difficulty of conducting Bayesian model selection and averaging. They proposed an estimator of the expected log-predictive likelihood. The estimator is derived

by correcting the asymptotic bias of the log-likelihood of the predictive distribution as an estimate of its expected value.

Whitney and Ngo (2004) described a BMA procedure that used the BIC as the basis for the weights. Candidate models were fitted with the SAS GLIMMIX procedure and then each coefficient of the final model was averaged from the equivalent estimated coefficient of the candidate model with the BIC as the average weight.

In the traffic safety area, Li et al. (2008) noted that: "…with few exceptions, model selection in traffic safety studies does not receive as much attention as do the methods implemented to estimate the parameters in those models". They attempted to select the "best" model from several plausible model formulations by examining the fit of the models. As for a Bayesian model approach, the DIC was applied as a criterion of model comparison. The selected model was then diagnosed by "discrepancy statistics" such as:

$$T_1(y_i, \theta) = max(y_{it}),$$

$$T_2(y_i, \theta) = standard\ deviation\ of\ y_{it} \quad (6\text{-}6)$$

where

y_{it} is the observed number of crashes at site i during time t, and $y_{it} \sim \text{Poisson}(\theta_{it})$,and

θ is the mean of the Poisson distribution.

To carry out posterior predictive checks, the authors first generated replicate data sets y^{rep} from a fitted model. They then computed the value of the statistic $Tj(y^{rep}, \theta)$ by using each replicated dataset and compared the $Tj(y^{obs}, \theta)$ (where y^{obs} denotes the observed crash data) to the distribution of the $Tj(y^{rep}, \theta)$ over the replicates. This procedure simulates the classical approach in that it considers the behavior of the statistics under repeated sampling. Hence, Equation 6-6 defines the diagnostic statistics for the model fit.

Zou et al. (2012) applied BMA to predict motor vehicle crashes. Based on the observed data, Poisson and NB distribution models were originally developed. Candidate models with the

highest posterior model probabilities were selected. Then, the BMA procedure was conducted based on the candidate models to obtain coefficients of the final model by estimating the weighted average of the candidate model coefficients. The weight is the posterior model probability of every candidate model. Finally, the logarithm scores were used to evaluate the predictive performance of the BMA models. The evaluation results showed that the BMA is a promising approach to predict crashes. This paper, however, did not specify the function and estimation process for the posterior model probabilities.

In summary, the specific type of model smoothing by mixing, or in effect, model averaging, has been proven to be a sound alternative to selecting a single "best" model in many fields. However, in the road safety field, previous research has not sufficiently developed a methodology for safety model averaging. Research for this dissertation has further developed this methodology based on safety datasets, as described in Sections 6.2 and 6.3.

6.2 BAYESIAN MODEL AVERAGING (BMA) METHODOLOGY

Since all of the estimated models investigated in Chapter 4 have already passed the statistical significance criteria, they are all considered. This section deals with a post processing method of coefficient estimation. Where multiple choices are present, there could be two different post-estimation solutions: one is the exclusive model selection approach, while the other is a non-exclusive model averaging approach. Section 6.2 begins with a conceptual comparison between these two approaches, in which the superiority of model averaging is highlighted. The following sub-section introduces the BMA methodology. After the preliminary selection of candidate models, a BMA is conducted for datasets described in Chapter 2. Finally, GOF tests will establish the legitimacy and advantage of the BMA.

6.2.1 Model Selection versus Model Averaging

The models presented in both Chapters 3 and 4 all have pros and cons. No single model is clearly preferable. The traditional model selection process abandons all but one model, subsequently

leading to the neglect of uncertainty of the selected model when it is given blind faith. Moreover, these models were developed by different approaches, which apply different measures for GOF testing. Without a common set of measures, model selection cannot be confidently accomplished.

The model averaging method can overcome these drawbacks in that it requires the estimation of the average of a parameter from a set of candidate models. By doing so, model selection uncertainty is included in the estimate of the precision of the parameters, and thus unconditional estimates of variances and standard errors are produced (Warner College of Natural Resources, 2011).

In the context of before-after evaluations, model averaging is conceptually a better way to seek reference information than selection of the sole "best" model since it integrates different knowledge sources, while selection of best model can neglect vast knowledge sources.

Finally, the model development in this dissertation serves as referential information for before-after evaluations, and for this reason, this research is aimed at integrating all established statistical significant models, instead of ranking and finally eliminating all but one.

6.2.2 Bayesian Model Averaging Theory and Functions

Either a frequentist or a Bayesian approach can accomplish model averaging once relevant criteria for model selection, as stated in Section 6.1.2, are considered when assigning weights in the averaging (Claeskens and Hjort, 2009). The weights used in frequentist model averaging are not available for Bayesian models. Yet BMA weights are adaptable for both frequentist and Bayesian models. Accordingly, the BMA is preferred. A summary of this approach, based on Claeskens and Hjort (2009), follows.

Suppose that there is a set of reasonable models for estimating a parameter of interest μ from a dataset of y. Here, μ is defined and has a common interpretation for all of the considered models. Define:

- Prior probabilities $P(M_j)$ for all considered models, labeled as M_1, M_2, …, M_k
- Prior distribution $\pi(\theta_j|M_j)$ for the parameters θ_j of the M_j model

From model fitting, an integrated likelihood is obtained for model M_j, which is denoted as $\lambda_{n,j}(y)$, and is also the marginal likelihood. Its value can be found in the output by SAS. By using Bayes' theorem, the posterior density of the model is obtained as:

$$P(M_j|y) = \frac{P(M_j)\cdot\lambda_{n,j}(y)}{\sum_{j'=1}^{k} P(M_{j'})\cdot\lambda_{n,j'}(y)} \quad (6\text{-}7)$$

Only log-likelihood values, not absolute likelihood itself, were directly output by SAS. Given that the log-likelihood values in this dissertation were all negative and their values were very small, computations of absolute likelihood values from log-likelihood values yielded no meaningful result. For example, the absolute likelihood values calculated from log-likelihood values in Table 6-1 were all "0". For purpose of model averaging, in this dissertation, log-likelihood values were applied instead of absolute likelihood values. Moreover, the final value of $\lambda_{n,j}(y)$ was estimated as (-1/log-likelihood) instead, in accordance with the logic that a larger likelihood value leads to greater weight.

If the posterior density of μ for model M_j is $\pi(\mu|M_j, y)$, then the integrated posterior density of μ can be expressed as:

$$\pi(\mu|y) = \sum_{j=1}^{k} P(M_j|y)\cdot\pi(\mu|M_j, y) \quad (6\text{-}8)$$

Finally, the posterior mean of the parameter is, likewise, a weighted average of the posterior means in the separate models, estimated as:

$$E(\mu|y) = \sum_{j=1}^{k} P(M_j|y) \cdot E(\mu|M_j, y) \quad (6\text{-}9)$$

6.3 SELECTION OF CANDIDATE MODELS

Since the theme of this dissertation is integration of prior model knowledge, all of the established SS models, which are either calibrated HSM models or locally developed full and multi-level models, were considered in the model averaging process. The candidate models are denoted by a checkmark (✓) as shown in Tables 4-4 to 4-9 (see more details in Chapter 4).

Based on each of six different databases (two populations and four sample groups), the calibrated HSM models are averaged with SS locally developed "single-level" full models developed with the same database. The averaging procedure is repeated for the calibrated HSM and multi-level models. Finally, there are in total 12 BMA procedures conducted in this dissertation study. The number of candidate models changes from one procedure to another, depending on the number of models that are SS.

Details on the candidate models are provided in Tables 6-1 to 6-6.

Table 6-1 Estimates of Posterior Model Probabilities (Toronto, population)

No. of BMA Procedure	Model Hierarchy	Approach	Procedure	Log-likelihood	Model Probability	
					Prior	Posterior
No.1	Calibrated HSM Model	Calibrated	HSM type calibration	-7874.5	1/7	18.593%
	Full Model	Frequentist	GENMOD NB	-7241.7	1/7	20.217%
			COUNTREG ZIP	-1440722	1/7	0.102%
			NLMIXED	-7277.5	1/7	20.118%
		Bayesian	Poisson -gamma	-7694.19	1/7	19.028%
			Poisson -lognormal	-13741.7	1/7	10.654%
			Poisson -Weibull	-12971	1/7	11.287%
No.2	Calibrated HSM Model	Calibrated	HSM type calibration	-7874.5	1/3	32.226%
	Multi-level Model	Frequentist	NLMIXED	-7283.5	1/3	34.841%
		Bayesian	Poisson -gamma	-7705.22	1/3	32.934%
			Poisson -lognormal	-13089.7	-	-
			Poisson -Weibull	-	-	-

Table 6-2 Estimates of Posterior Model Probabilities (Toronto, sample size=680)

No. of BMA Procedure	Model Hierarchy	Approach	Procedure	Log-likelihood	Model Probability	
					Prior	Posterior
No.3	Calibrated HSM Model	Calibrated	HSM type calibration	-3261.47	1/6	21.267%
	Full Model	Frequentist	GENMOD NB	-3027.7	1/6	22.909%
			COUNTREG ZIP	-599533	1/6	0.116%
			NLMIXED	-3035.85	1/6	22.847%
		Bayesian	Poisson -gamma	-3226.6515	1/6	21.496%
			Poisson -lognormal	-6102.63	1/6	11.366%
			Poisson -Weibull	-6102.59	-	-
No.4	Calibrated HSM Model	Calibrated	HSM type calibration	-3261.47	1/4	27.552%
	Multi-level Model	Frequentist	NLMIXED	-3053.45	1/4	29.429%
		Bayesian	Poisson -gamma	-3229.14	1/4	27.828%
			Poisson -lognormal	-5915.09	1/4	15.192%
			Poisson -Weibull	-	-	-

Table 6-3 Estimates of Posterior Model Probabilities (Toronto, sample size=588)

No. of BMA Procedure	Model Hierarchy	Approach	Procedure	Log-likelihood	Model Probability	
					Prior	Posterior
No.5	Calibrated HSM Model	Calibrated	HSM type calibration	-2827.884	1/6	21.166%
	Full Model	Frequentist	GENMOD NB	-2602.2	1/6	23.001%
			COUNTREG ZIP	-516522	1/6	0.116%
			NLMIXED	-2633.45	1/6	22.728%
		Bayesian	Poisson -gamma	-2782.8785	1/6	21.508%
			Poisson -lognormal	-5213.89	-	-
			Poisson -Weibull	-5213.26	1/6	11.481%
No.6	Calibrated HSM Model	Calibrated	HSM type calibration	-2827.884	1/3	33.892%
	Multi-level Model	Frequentist	NLMIXED	-2631.1	1/3	36.427%
		Bayesian	Poisson -gamma	-3229.04	1/3	29.681%
			Poisson -lognormal	-12972.9	-	-
			Poisson -Weibull	-	-	-

Table 6-4 Estimates of Posterior Model Probabilities (Edmonton, population)

No. of BMA Procedure	Model Hierarchy	Approach	Procedure	Log-likelihood	Model Probability	
					Prior	Posterior
No.7	Calibrated HSM Model	Calibrated	HSM type calibration	-2670.63	1/5	22.017%
	Full Model	Frequentist	GENMOD NB	-2481.4	1/5	23.696%
			COUNTREG ZIP	-		
			NLMIXED	-2484	1/5	23.671%
		Bayesian	Poisson -gamma	-2556.54	1/5	22.999%
			Poisson -lognormal	-7717.55	-	-
			Poisson -Weibull	-7718.12	1/5	7.618%
No.8	Calibrated HSM Model	Calibrated	HSM type calibration	-2670.63	1/3	32.030%
	Multi-level Model	Frequentist	NLMIXED	-2480.4	1/3	34.487%
		Bayesian	Poisson -gamma	-2554.78	1/3	33.483%
			Poisson -lognormal	-7531.52	-	-
			Poisson -Weibull	-	-	-

Table 6-5 Estimates of Posterior Model Probabilities (Edmonton, sample size=400)

No. of BMA Procedure	Model Hierarchy	Approach	Procedure	Log-likelihood	Model Probability	
					Prior	Posterior
No.9	Calibrated HSM Model	Calibrated	HSM type calibration	-2063.518	1/5	22.435%
	Full Model	Frequentist	GENMOD NB	-1925.6	1/5	24.042%
			COUNTREG ZIP	-	-	-
			NLMIXED	-1911.4	1/5	24.220%
		Bayesian	Poisson -gamma	-1981.68	1/5	23.362%
			Poisson -lognormal	-7792.27	1/5	5.941%
			Poisson -Weibull	-6462.66	-	-
No.10	Calibrated HSM Model	Calibrated	HSM type calibration	-2063.518	1/3	32.082%
	Multi-level Model	Frequentist	NLMIXED	-1922.15	1/3	34.441%
		Bayesian	Poisson -gamma	-1977.54	1/3	33.477%
			Poisson -lognormal	-6215.65	-	-
			Poisson -Weibull	-	-	-

Table 6-6 Estimates of Posterior Model Probabilities (Edmonton, sample size=300)

No. of BMA Procedure	Model Hierarchy	Approach	Procedure	Log-likelihood	Model Probability	
					Prior	Posterior
No.11	Calibrated HSM Model	Calibrated	HSM type calibration	-1549.476	1/5	22.427%
	Full Model	Frequentist	GENMOD NB	-1447.4	1/5	24.009%
			COUNTREG ZIP	-	-	-
			NLMIXED	-1459.25	1/5	23.814%
		Bayesian	Poisson -gamma	-1494.03	1/5	23.260%
			Poisson -lognormal	-5354.47	1/5	6.490%
			Poisson -Weibull	-4491.9	-	-
No.12	Calibrated HSM Model	Calibrated	HSM type calibration	-1549.476	1/3	32.186%
	Multi-level Model	Frequentist	NLMIXED	-1448.4	1/3	34.432%
		Bayesian	Poisson -gamma	-1493.92	1/3	33.383%
			Poisson -lognormal	-4360.08	-	-
			Poisson -Weibull	-	-	-

6.4 APPLICATION OF BAYESIAN MODEL AVERAGING

It has been rationally assumed that the prior probabilities of the models are equally distributed among the candidate models (Claeskens and Hjort, 2009). Thus, for each of the model averaging procedures with N candidate models, prior probabilities $P(M_j)$ are assumed as 1/N for each model.

Based on the log-likelihood values listed in Tables 4-4 to 4-9 for the relevant models, the posterior probabilities of the models were calculated by Equation 6-7, and are shown in Tables 6-1 to 6-6.

The posterior probabilities of the models are in effect the weights applied in averaging the parameter means. The final model parameters were obtained from Equation 6-9.

For the BMA full model, the final functional form is:

$$\begin{aligned}\mu_{it} = {} & \beta_0 \cdot (majorAADT)^{\beta_1} \cdot (\min orAADT)^{\beta_2} \cdot \exp(\beta_3 \cdot I_{1lt}) \cdot \exp(\beta_4 \cdot I_{2lt}) \cdot \\ & \exp(\beta_5 \cdot I_{3lt}) \cdot \exp(\beta_6 \cdot I_{4lt}) \cdot \exp(\beta_7 \cdot I_{1rt}) \cdot \exp(\beta_8 \cdot I_{2rt}) \cdot \exp(\beta_9 \cdot I_{3rt}) \\ & \cdot \exp(\beta_{10} \cdot I_{4rt}) \cdot \exp(\beta_{11} \cdot class)\end{aligned} \quad (6\text{-}10)$$

where

variable meanings are explained in Table 4-1, and

β_0, β_1,….β_{11} are the estimated coefficients.

For BMA full models, the estimation results of all candidate and final BMA models are presented in Tables 6-7 to 6-12. Since most of the model variables have point-estimated numbers as coefficients, the final BMA models yield mostly coefficients with fixed figures. As for the β_{11} of Equation 6-10, since some candidate models have categorical variables with regards to the level of type, in this case, the BMA model cannot yield a constant β_{11}; instead, it remains as a "categorical" variable that is assigned values by specific level of type.

For the BMA multi-level model, the final functional form is the same as Equation 3-18 for consistency. However, since coefficients "A", "B", and "C" (Equation 3-18) for candidate multi-level models are variables other than constants, the coefficients of the final BMA multi-level models cannot be represented by constant values. Instead, the BMA model can only be expressed by functions as:

$$N_{mv} = \tilde{A}(AADT_{maj})^{\tilde{B}}(AADT_{\min})^{\tilde{C}} \quad (6\text{-}11)$$

where

N_{mv} = predicted average crash frequency for multiple-vehicle collisions,

$\tilde{A}, \tilde{B}, \tilde{C}$ are parameter coefficients estimated for the first-level BMA model, and

$$\tilde{A} = \sum_{j=1}^{k} P(M_j|y) \cdot E(A_j|M_j, y) \quad (6\text{-}12)$$

$$\tilde{B} = \sum_{j=1}^{k} P(M_j|y) \cdot E(B_j|M_j, y) \quad (6\text{-}13)$$

$$\tilde{C} = \sum_{j=1}^{k} P(M_j|y) \cdot E(C_j|M_j, y) \quad (6\text{-}14)$$

where

$p((M_j|y)$ is the posterior probability of candidate model j, and

A_j, B_j and C_j are the coefficients of candidate model j, and estimated by Equation 3-18.

To be specific, the relevant coefficients and collision modification factors (CMFs) of the calibrated HSM models are reshuffled in Equation 3-19 to match the requirements of model averaging.

Hence, in Tables 6-13 to 6-18, only sub-level coefficients are presented for individual candidate models. The final BMA multi-level models do not have fixed coefficient values; instead, their coefficients are variables that can only be estimated respectively from Equations 6-12 to 6-14.

Table 6-7 Parameter Estimates for Full BMA Models (Toronto, Population)

Parameter[a]	Calibrated HSM Model	Frequentist			Bayesian			BMA Model	P[b]
		GENMOD NB	COUNTREG ZIP	NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull		
BMA weights	18.593%	20.217%	0.102%	20.118%	19.028%	10.654%	11.287%	100.000%	
constant (β0)	7.878E-05	7.431E-04	1.750E-03	1.164E-03	1.823E-03	3.915E-04	4.135E-04	8.359E-04	1.000
major AADT (β1)	1.070	0.518	0.513	0.642	0.652	0.498	0.494	0.666	1.000
minor AADT (β2)	0.230	0.454	0.469	0.453	0.404	0.668	0.666	0.449	1.000
I_{1lt} (β3)	-0.105	-0.133	-0.157	-0.148	-0.190			-0.113	0.781
I_{2lt} (β4)	-0.211	-0.193	-0.232	-0.214	-0.245			-0.168	0.781
I_{3lt} (β5)	-0.315	-0.271	-0.339	-0.311	-0.336			-0.240	0.781
I_{4lt} (β6)	-0.416	-0.202	-0.206	-0.225	-0.233			-0.208	0.781
I_{1rt} (β7)	-0.041	-0.049	-0.108	-0.059	-0.084			-0.046	0.781
I_{2rt} (β8)	-0.083	-0.070	-0.125	-0.089	-0.104			-0.068	0.781
I_{3rt} (β9)	-0.128		-0.035					-0.024	0.187
I_{4rt} (β10)	-0.163		0.035					-0.030	0.187
class (β11)		categorical	categorical	-0.074	-0.089			categorical	0.595

Note: a. Please see variable definitions in Table 4-1. This adapts Table 6-8 to 6-18 as well.

Note: b. P-posterior effect probability. This is the sum of all probabilities from all models in which the coefficient exists (Whitney and Ngo, 2004). P indicates the strength of the effect that a variable has on the model. This adapts Table 6-8 to 6-18 as well.

Table 6-8 Parameter Estimates for Full BMA Models (Toronto, sample size=680)

Parameter	Calibrated HSM Model	Frequentist			Bayesian			BMA Model	P
		GENMOD NB	COUNTREG ZIP	NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull		
BMA weights	21.267%	22.909%	0.116%	22.847%	21.496%	11.366%	0.000%	100.000%	
constant (β0)	7.878E-05	7.431E-04	1.750E-03	1.164E-03	1.823E-03	3.915E-04	4.135E-04	8.911E-04	1.000
major AADT (β1)	1.070	0.534	0.543	0.668	0.748	0.557	0.562	0.727	1.000
minor AADT (β2)	0.230	0.455	0.469	0.488	0.398	0.659	0.658	0.426	1.000
I_{1lt} (β3)	-0.105	-0.106	-0.120	-0.110	-0.225			-0.120	0.886
I_{2lt} (β4)	-0.211	-0.156	-0.189	-0.250	-0.271			-0.196	0.886
I_{3lt} (β5)	-0.315	-0.271	-0.383	-0.276	-0.426			-0.284	0.886
I_{4lt} (β6)	-0.416	-0.212	-0.229	-0.214	-0.336			-0.258	0.886
I_{1rt} (β7)	-0.041		-0.061					-0.009	0.214
I_{2rt} (β8)	-0.083		-0.090	-0.126	-0.110			-0.070	0.657
I_{3rt} (β9)	-0.128	-0.073	-0.128					-0.044	0.443
I_{4rt} (β10)	-0.163		0.145					-0.034	0.214
class (β11)		categorical	categorical	-0.063	-0.089			categorical	0.674

Table 6-9 Parameter Estimates for Full BMA Models (Toronto, sample size=588)

Parameter	Calibrated HSM Model	Frequentist			Bayesian			BMA Model	P
		GENMOD NB	COUNTREG ZIP	NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull		
BMA weights	21.166%	23.001%	0.116%	22.728%	21.508%	0.000%	11.481%	100.000%	
constant (β0)	7.878E-05	8.273E-04	1.454E-03	1.039E-03	1.487E-03	2.810E-04	2.615E-04	7.947E-04	1.000
major AADT (β1)	1.070	0.481	0.512	0.625	0.677	0.510	0.514	0.684	1.000
minor AADT (β2)	0.230	0.472	0.498	0.467	0.393	0.694	0.694	0.428	1.000
I_{1lt} (β3)	-0.105		-0.096		-0.150			-0.055	0.428
I_{2lt} (β4)	-0.211	-0.108	-0.181	-0.074	-0.218			-0.133	0.885
I_{3lt} (β5)	-0.315		-0.064		-0.229			-0.116	0.428
I_{4lt} (β6)	-0.416	-0.139	-0.174		-0.263			-0.177	0.658
I_{1rt} (β7)	-0.041	-0.078	-0.134	-0.156				-0.062	0.670
I_{2rt} (β8)	-0.083		-0.138	-0.218	-0.036			-0.075	0.655
I_{3rt} (β9)	-0.128		-0.131	-0.263				-0.087	0.440
I_{4rt} (β10)	-0.163		0.040					-0.034	0.213
class (β11)		categorical	categorical	-0.067	-0.091			categorical	0.674

Table 6-10 Parameter Estimates for Full BMA Models (Edmonton, Population)

Parameter	Calibrated HSM Model	Frequentist			Bayesian			BMA Model	P
		GENMOD NB	COUNTREG ZIP	NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull		
BMA weights	22.017%	23.696%	0.000%	23.671%	22.999%	0.000%	7.618%	100.000%	
constant (β0)	7.878E-05	3.332E-05	1.178E-03	1.502E-05	2.351E-05	1.828E-05	2.789E-05	3.633E-05	1.000
major AADT (β1)	1.070	0.721		0.740	0.680	0.573	0.574	0.782	1.000
minor AADT (β2)	0.230	0.610		0.601	0.623	0.752	0.752	0.538	1.000
I_{1lt} (β3)	-0.105	0.216						0.028	0.457
I_{2lt} (β4)	-0.211							-0.046	0.220
I_{3lt} (β5)	-0.315							-0.069	0.220
I_{4lt} (β6)	-0.416							-0.091	0.220
I_{1rt} (β7)	-0.041							-0.009	0.220
I_{2rt} (β8)	-0.083							-0.018	0.220
I_{3rt} (β9)	-0.128							-0.028	0.220
I_{4rt} (β10)	-0.163	0.368		0.327	0.377	0.232	0.238	0.234	1.000
class (β11)		-0.360		0.366	0.339	0.444	0.444	0.113	0.780

Table 6-11 Parameter Estimates for Full BMA Models (Edmonton, sample size=400)

Parameter	Calibrated HSM Model	Frequentist			Bayesian			BMA Model	P
		GENMOD NB	COUNTREG ZIP	NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull		
BMA weights	22.435%	24.042%	0.000%	24.220%	23.362%	5.941%	0.000%	100.000%	
constant (β0)	7.878E-05	3.140E-05	1.589E-03	9.801E-06	2.289E-05	1.105E-05	4.023E-05	3.360E-05	1.000
major AADT (β1)	1.070	0.715		0.783	0.664	0.651	0.536	0.795	1.000
minor AADT (β2)	0.230	0.625		0.596	0.641	0.800	0.776	0.543	1.000
I_{1lt} (β3)	-0.105	0.263		0.311				0.115	0.707
I_{2lt} (β4)	-0.211							-0.047	0.224
I_{3lt} (β5)	-0.315							-0.071	0.224
I_{4lt} (β6)	-0.416							-0.093	0.224
I_{1rt} (β7)	-0.041							-0.009	0.224
I_{2rt} (β8)	-0.083							-0.019	0.224
I_{3rt} (β9)	-0.128							-0.029	0.224
I_{4rt} (β10)	-0.163	0.335		0.414	0.329		0.167	0.221	0.941
class (β11)		-0.371		0.327	0.358		0.481	0.074	0.716

Table 6-12 Parameter Estimates for Full BMA Models (Edmonton, sample size=300)

Parameter	Calibrated HSM Model	Frequentist			Bayesian			BMA Model	P
		GENMOD NB	COUNTREG ZIP	NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull		
BMA weights	22.427%	24.009%	0.000%	23.814%	23.260%	6.490%	0.000%	100.000%	
constant (β0)	7.878E-05	2.597E-05	3.927E-04	7.927E-06	1.660E-05	6.221E-06		3.006E-05	1.000
major AADT (β1)	1.070	0.777		0.832	0.755	0.755		0.849	1.000
minor AADT (β2)	0.230	0.568		0.569	0.573	0.750		0.505	1.000
I_{1lt} (β3)	-0.105	0.328						0.055	0.464
I_{2lt} (β4)	-0.211							-0.047	0.224
I_{3lt} (β5)	-0.315							-0.071	0.224
I_{4lt} (β6)	-0.416							-0.093	0.224
I_{1rt} (β7)	-0.041							-0.009	0.224
I_{2rt} (β8)	-0.083							-0.019	0.224
I_{3rt} (β9)	-0.128							-0.029	0.224
I_{4rt} (β10)	-0.163	0.443		0.297	0.386			0.230	0.935
class (β11)		-0.3336		0.3665	0.354			0.090	0.711

Table 6-13 Parameter Estimates for Candidate[a] Multi-level Models (Toronto, Population)

Parameter	Calibrated HSM Model	Frequentist	Bayesian			P
		NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull	
BMA weights	32.226%	34.841%	32.934%	0.000%	0.000%	
α0	CMF	-7.3137	-7.1157	-6.4058		1.000
LT (α1)		-0.063	-0.0837	-0.037		1.000
RT (α2)		-0.02344	-0.0167	-0.0157		1.000
β0	1.07	0.5668	0.6033	0.5796		1.000
class at B (β1)		0.005428	0.00166	-0.00575		1.000
γ0	0.23	0.6005	0.5528	0.5487		1.000
class at C (γ1)		-0.01568	-0.0131	-		1.000

Note: a. Final multi-level BMA Model Coefficients are estimated by Equation 6-12 to 6-14. This note adapts Table 6-14 to 6-18 as well.

Table 6-14 Parameter Estimates for Candidate Multi-level Models (Toronto, sample size=680)

Parameter	Calibrated HSM Model	Frequentist	Bayesian			P
		NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull	
BMA weights	27.552%	29.429%	27.828%	15.192%	0.000%	
α0	CMF	-8.064	-8.0765	-7.9852		1.000
LT (α1)		-0.06924	-0.1069	-0.036		1.000
RT (α2)		-0.03515	-0.0149	-0.0155		1.000
β0	1.07	0.6068	0.6417	0.6182		1.000
class at B (β1)		0.008632	0.00804	-0.00413		1.000
γ0	0.23	0.6388	0.6265	0.5835		1.000
class at C (γ1)		-0.01845	-0.0201	-		0.848

Table 6-15 Parameter Estimates for Candidate Multi-level Models (Toronto, sample size=588)

Parameter	Calibrated HSM Model	Frequentist	Bayesian			P
		NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull	
BMA weights	33.892%	36.427%	29.681%	0.000%	0.000%	
α0	CMF	-7.7264	-8.1939	-6.4827		1.000
LT (α1)		-0.03671	-0.1038	-0.0398		1.000
RT (α2)		-0.05399	-0.0248	-0.0224		1.000
β0	1.07	0.4881	0.6363	0.4576		1.000
class at B (β1)		0.0153	0.00932	0.0103		1.000
γ0	0.23	0.7229	0.6429	0.6796		1.000
class at C (γ1)		-0.02561	-0.0212	-0.0189		1.000

Table 6-16 Parameter Estimates for Candidate Multi-level Models (Edmonton, population)

Parameter	Calibrated HSM Model	Frequentist	Bayesian			P
		NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull	
BMA weights	32.030%	34.487%	33.483%	0.000%	0.000%	
$\alpha 0$	CMF	-10.6091	-9.9176	-10.9095		1.000
LT ($\alpha 1$)		-0.04553	-0.0584	0.5734		1.000
RT ($\alpha 2$)		0.104	0.1228	0.7523		1.000
$\beta 0$	1.07	0.9238	0.8553	0.2319		1.000
class at B ($\beta 1$)		-0.1324	-0.1225	0.4441		1.000
$\gamma 0$	0.23	0.3399	0.3409	2.957997		1.000
class at C ($\gamma 1$)		0.1881	0.1752	-		1.000

Table 6-17 Parameter Estimates for Candidate Multi-level Models (Edmonton, sample size=400)

Parameter	Calibrated HSM Model	Frequentist	Bayesian			P
		NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull	
BMA weights	32.082%	34.441%	33.477%	0.000%	0.000%	
$\alpha 0$	CMF	-10.9139	-10.4615	-9.9686		1.000
LT ($\alpha 1$)		-0.05555	-0.0545	-0.0237		1.000
RT ($\alpha 2$)		0.09313	0.1124	0.0821		1.000
$\beta 0$	1.07	1.0701	0.9886	0.9736		1.000
class at B ($\beta 1$)		-0.239	-0.2074	-0.2772		1.000
$\gamma 0$	0.23	0.2111	0.2492	0.2149		1.000
class at C ($\gamma 1$)		0.3099	0.2741	0.3504		1.000

Table 6-18 Parameter Estimates for Candidate Multi-level Models (Edmonton, sample size=300)

Parameter	Calibrated HSM Model	Frequentist	Bayesian			P
		NLMIXED	Poisson - gamma	Poisson - lognormal	Poisson - Weibull	
BMA weights	32.186%	34.432%	33.383%	0.000%	0.000%	
α_0	CMF	-11.0388	-10.4376			1.000
LT (α_1)		-0.05947	-0.0732			1.000
RT (α_2)		0.1163	0.1535			1.000
β_0	1.07	1.0287	0.6947			1.000
class at B (β_1)		-0.1643	0.0306			1.000
γ_0	0.23	0.2727	0.5831			1.000
class at C (γ_1)		0.2212	-			0.666

6.5 STATISTICAL DIAGNOSTICS OF BMA MODELS

Individual full and multi-level models developed in this dissertation were evaluated with respect to two aspects: statistical diagnostics to test whether models were statistically valid (convergence as a whole, and significance of individual variable coefficients) and goodness-of-fit tests to measure a model's prediction performance (GOF, including overall measures, and CURE plots). (See definitions and more details in Chapter 3 and 4 for all these measures.)

Among these two aspects, statistical diagnostics were conducted along with model fitting and coefficient estimation by SAS; hence they were all directly derived from software outputs, as shown as those model estimation results in Chapters 3-5. In contrast, GOF tests were further calculated after model fitting by applying the model to test datasets.

BMA models in this dissertation were averaged from candidate models manually through the BMA approach introduced earlier this chapter, and were not exported directly from statistical software, so there is no direct convergence measure for the BMA model. Its convergence is rooted in the convergence of individual candidate models. The same approach was used in past research (i.e., Zou et al., 2012).

The statistical significance of each parameter of BMA models also relies on that of each candidate model. Besides, there is another measure - posterior effect probability (denoted as P in Table 6-7 to Table 6-18; see note of Table 6-7). P indicates the strength of the effect that a variable has on the final BMA model. A value of P closer to 1.0 is interpreted as a stronger effect for the variable.

By this means, estimation results of Table 6-7 to Table 6-18 showed that parameters of multi-level BMA models have generally stronger effects than those of full BMA models. With significantly smaller Ps, many parameters of Table 6-7 to Table 6-12 have only weak effects in the full BMA models.

The BMA model evaluation was mainly focused on its GOF tests, as introduced in Section 6.6.

6.6 GOODNESS OF FIT TESTS FOR BMA MODELS

Tables 6-19 and 6-20 present the comparisons of the overall GOF measures (explained in Chapter 3) for the calibrated HSM, BMA full and BMA multi-level models, respectively, for Toronto and Edmonton.

Table 6-19 GOF Test Measures for BMA Models for Toronto Datasets

Model Hierarchy	Database	Average Observed Collision	Pearson coefficient r	MPB	MAD	MSPE
Calibrated HSM Model	Population	10.34	0.61	-8.12	8.23	146.66
	Sample (size=680)	10.14	0.63	-7.99	8.11	148.16
	Sample (size=588)	10.17	0.63	-8.07	8.16	149.47
Full BMA model	Population	10.34	0.87	5.91	6.38	71.82
	Sample (size=680)	10.14	0.91	5.27	12.67	119.03
	Sample (size=588)	10.17	0.91	1.19	4.41	13.70
Multi-level BMA model	Population	10.34	0.84	**-0.22**	**3.57**	**27.97**
	Sample (size=680)	10.14	0.89	**0.78**	**4.02**	**13.67**
	Sample (size=588)	10.17	0.88	**0.81**	**4.27**	**13.02**

Note: bold numbers indicate that the Multi-level BMA models are better than the corresponding Full BMA models.

Table 6-20 GOF Test Measures of BMA Models for Edmonton Datasets

Model Hierarchy	Database	Average Observed Collision	Pearson coefficient r	MPB	MAD	MSPE
Calibrated HSM Model	Population	12.59	0.49	-7.49	8.75	215.22
	Sample (size=680)	12.31	0.48	-7.19	8.43	192.37
	Sample (size=588)	12.76	0.54	-7.68	8.92	229.78
Full BMA model	Population	12.59	0.82	0.99	6.67	29.30
	Sample (size=680)	12.31	0.83	0.90	6.78	22.20
	Sample (size=588)	12.76	0.85	0.75	7.12	19.22
Multi-level BMA model	Population	12.59	**0.83**	**-0.19**	**5.60**	**27.68**
	Sample (size=680)	12.31	**0.84**	**-0.0003**	**5.43**	**18.40**
	Sample (size=588)	12.76	**0.85**	**-0.09**	**5.52**	**16.55**

Note: bold numbers indicate that the Multi-level BMA models are better than the corresponding Full BMA models.

All four GOF test measures in Table 6-20 and three of four GOF measures in Table 6-19 (those in bold numbers) indicate two important conclusions:

- BMA models fit better than calibrated HSM models
- BMA multi-level models fit better than BMA full models

From the standards set in Chapter 3, it is clear that the calibrated HSM models result in much higher bias than both types of BMA models. For example, the MSPE measure, which highlights especially sensitive differences thanks to its quadratic leveling, reveals that both types of BMA models have a lower order of magnitude of predictive errors compared to the calibrated HSM models.

Furthermore, GOF measures show preference for BMA multi-level models over their full model counterparts. Most GOF test measures (denoted by the bold numbers in Tables 6-19 and 6-20) indicate that multi-level models are better than full models. As an example, the MSPE measures underline the superiority of the multi-level models, based on either population or sample.

The complementary assessment of model fit over the entire range of covariates was accomplished with CURE plots, as preliminarily discussed in Chapter 3. These plots are shown as Figures 6-1 to 6-6. The cumulative residuals are shown for the calibrated HSM models with grey color, thin solid lines, for the BMA full models with blue-color, horizontally dotted, intermediate-width dashed lines, and for the BMA multi-level models with red-color, vertically dotted, bold and wide dashed lines. The 2-standard deviation (2SD) boundaries in the CURE plots are those of calibrated HSM models, which resulted in the widest possible boundaries. This means that once a model's cumulative residual curve is outside the 2SD boundaries of calibrated HSM models, it can be deemed as having a bias that is much too large.

The CURE plots reveal that the BMA models provide improvements in two aspects. First, the cumulative residual curves of the BMA models are closer to the x-axis than those of the calibrated HSM models, indicating less bias for the BMA models. Second, the BMA models markedly diminished the predictive bias over ranges of AADTs for which the calibrated HSM

models for the Toronto population showed a high bias. The most significant improvement was for the Toronto population, for which the cumulative residual curve for the calibrated models was outside two standard deviation boundaries at a lower AADT level, while that for the BMA full model was mostly inside, and that for the BMA multi-level model was completely within these boundaries.

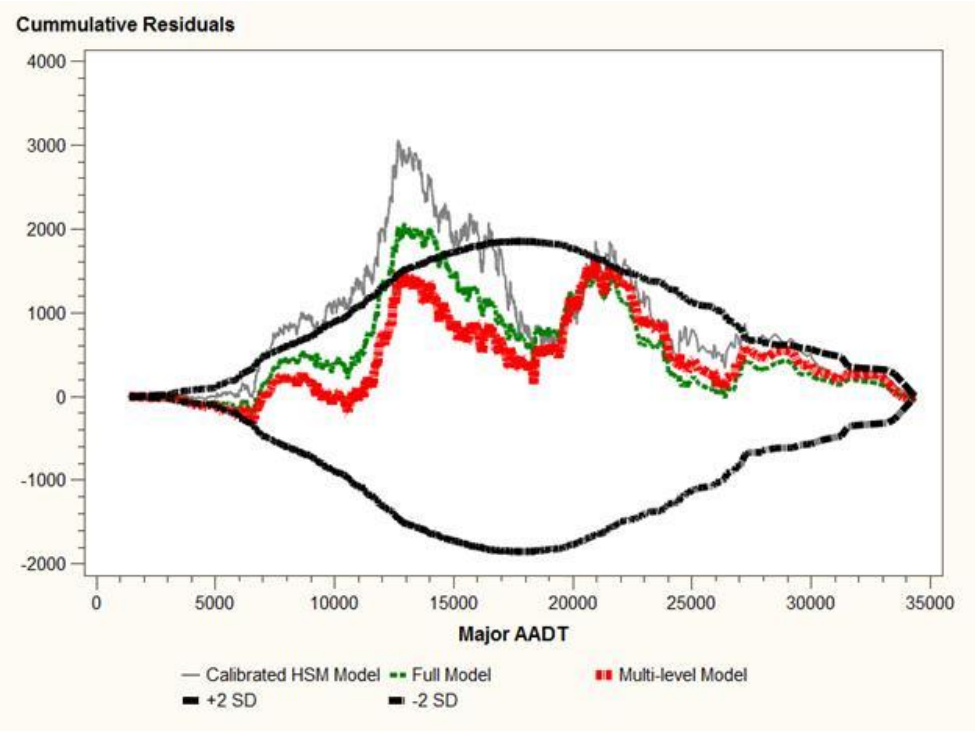


Figure 6-1 CURE Plots (Toronto, population)

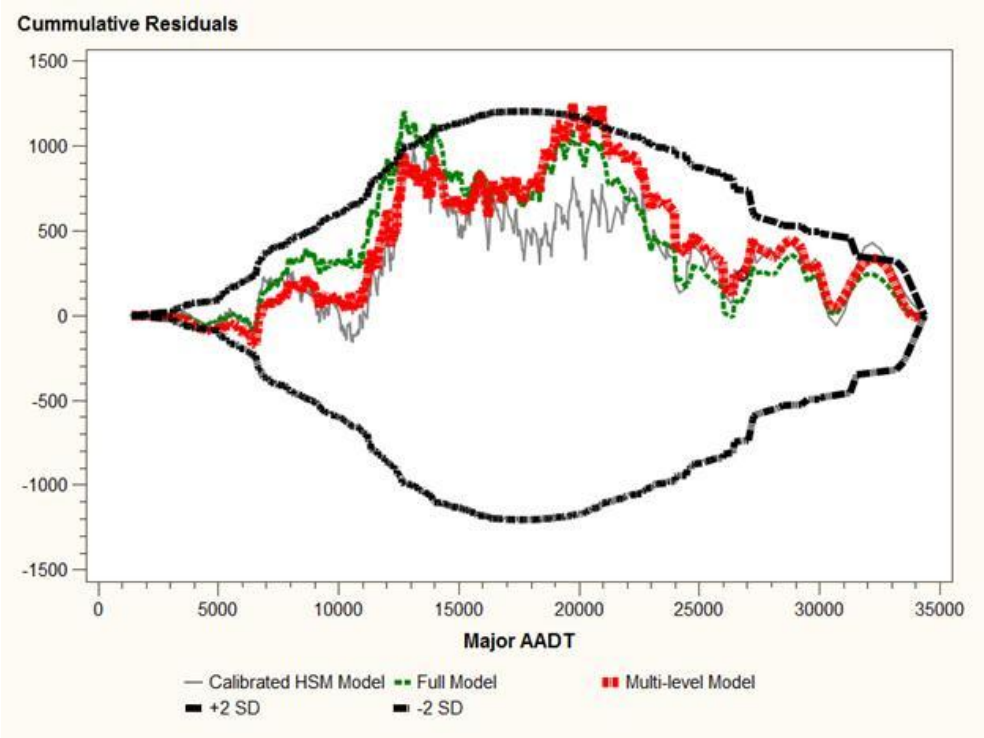


Figure 6-2 CURE Plots (Toronto, sample size=680)

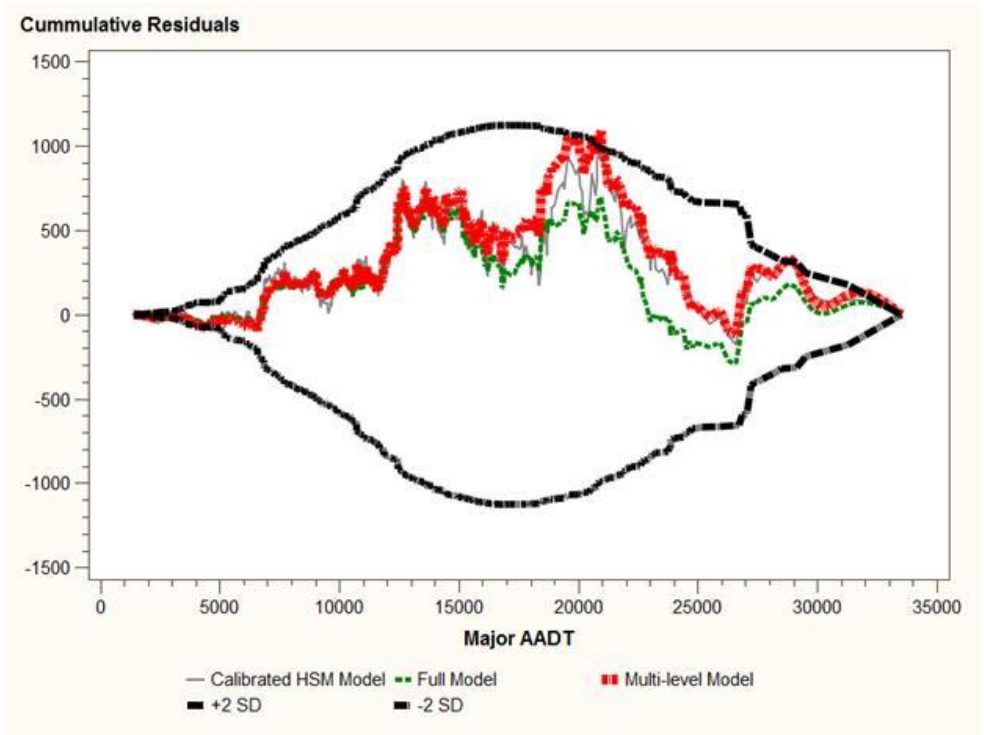


Figure 6-3 CURE Plots (Toronto, sample size=588)

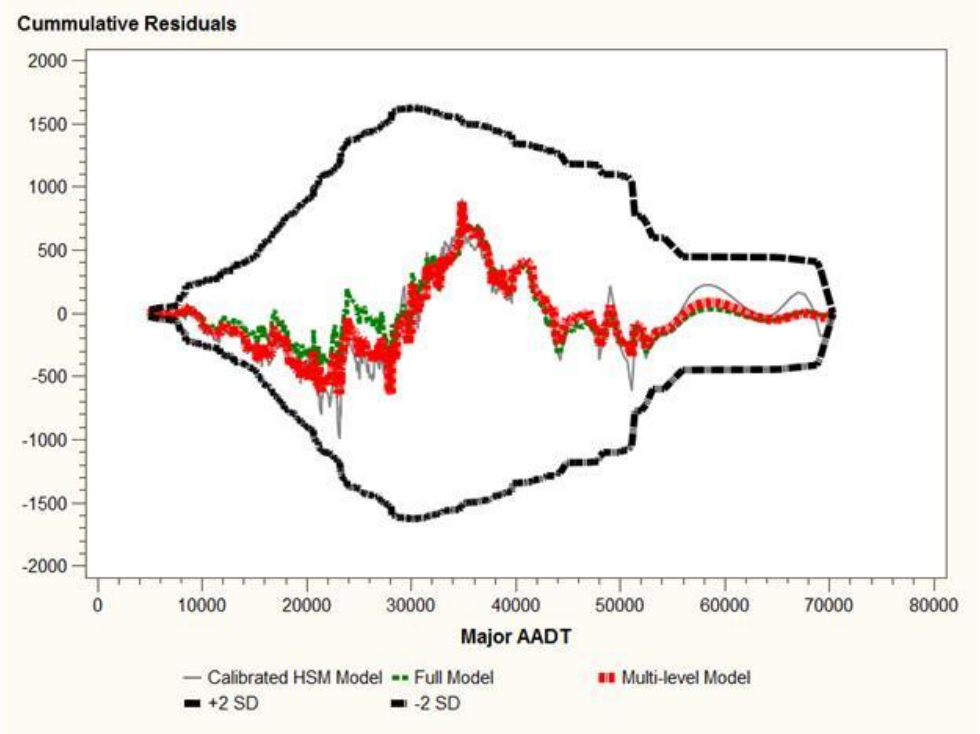


Figure 6-4 CURE Plots (Edmonton, population)

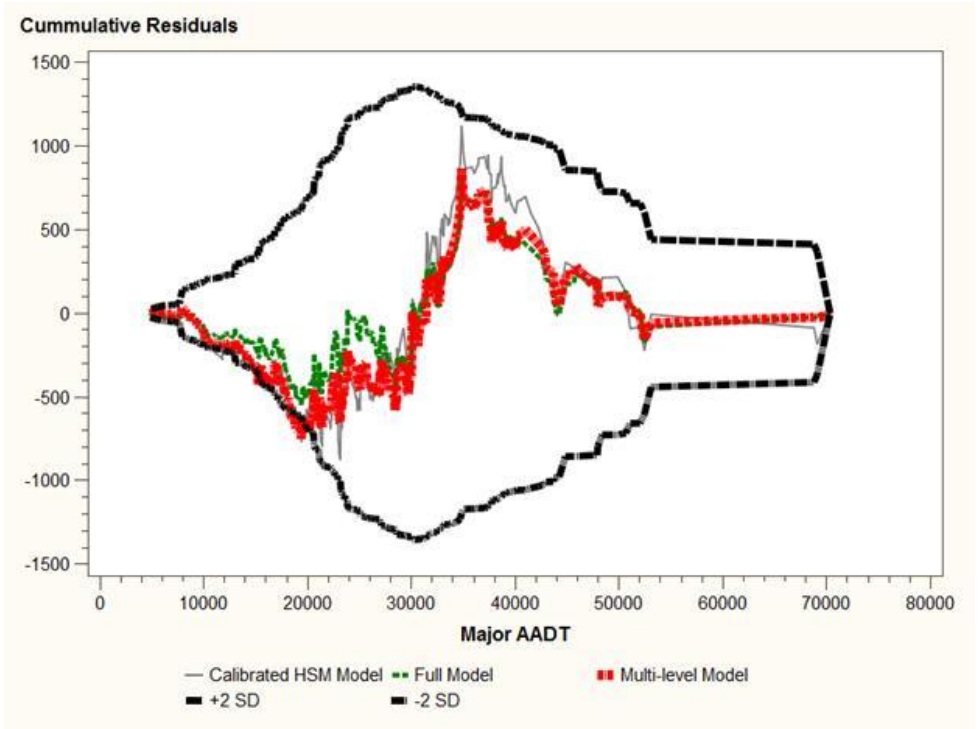


Figure 6-5 CURE Plots (Edmonton, sample size=400)

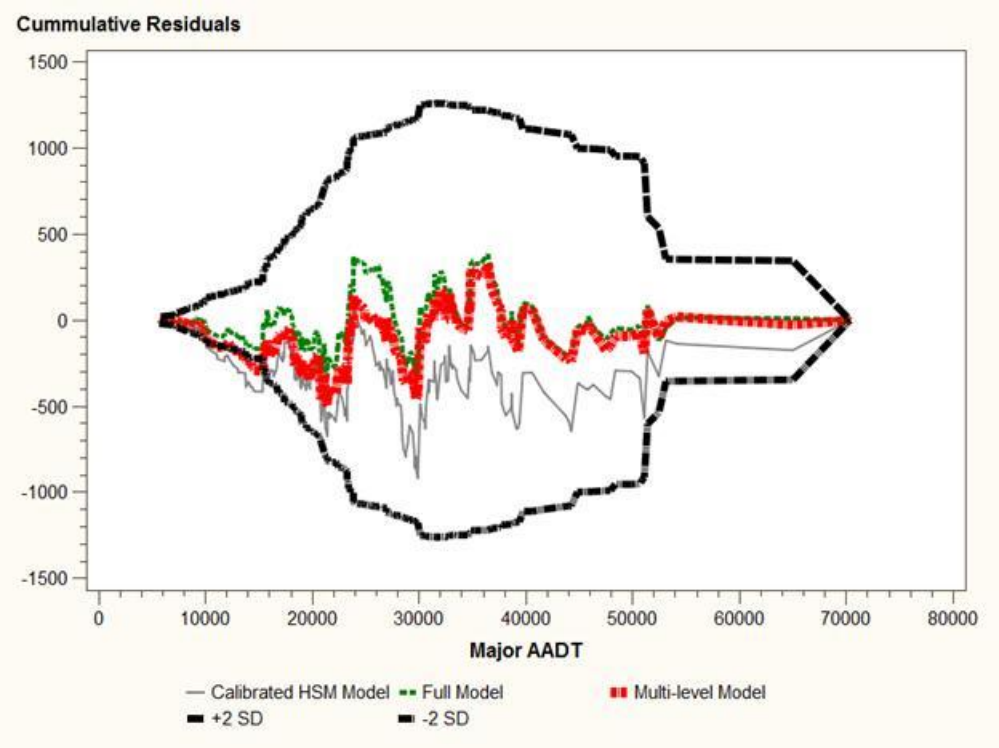


Figure 6-6 CURE Plots (Edmonton, sample size=300)

The CURE plots also suggest that BMA multi-level models fit either better than, or equally as well, as their full model counterparts. For the Toronto population, which has the largest data set, only the BMA multi-level models have CURE plots that entirely resided inside the 2SD boundaries. This phenomenon was also seen for the Toronto sample with a size of 680. For other databases, similarly, the CURE plots for the BMA multi-level models were also seen to be predominantly inside the 2SD boundaries.

In summary, after the investigation of the BMA procedures, two BMA models were obtained: a BMA full model with a single-level structure, and a BMA multi-level model. The overall GOF measures and CURE plots all proved the superiority of the BMA multi-level models compared to the BMA full models in terms of fitting. This provides evidence of the success of a multi-level hierarchy as well as the Bayesian model averaging approach.

6.7 CHAPTER CONCLUSIONS

This chapter has focused on the methodology to merge various referential sources in the context of before-after evaluations, which is a model selection and averaging process.

Following on a series of calibrations of HSM models and locally developed models in Chapter 3 and Chapter 4, this chapter has further explored the diversity of model development with the frequentist and Bayesian approaches, analytical MLE and simulation methods, as well as a variety of random structures, various function forms and diversified statistical procedures.

Given the diversity of model development, a method that integrates these models together is essential to proceed in the B/A process. In this chapter, a BMA process has been investigated by using model integration to forge a unified knowledge source for the subsequent B/A steps.

GOF overall measures and CURE plots were established to evaluate the performance of three types of models: calibrated HSM, BMA full and BMA multi-level models. These tests predominantly lead to two conclusions: BMA solutions, either as full or multi-level models, incur less bias than calibrated HSM models; and BMA multi-level models fit better than BMA full models.

The BMA process is conceptually attractive in that it imports local knowledge into an alien calibrated model to enhance its transferability, and in that it addresses the uncertainty of data in overcoming the potential limitations of a single "best" model. Furthermore, the BMA multi-level model is conceptually superior to its full model counterpart since it retains the homogeneity of the model while still being capable enough of addressing model individuality for a specific local region.

In conclusion, the success in estimating BMA multi-level models is sufficiently evident from a conceptual perspective and from the statistical features seen in both the GOF overall measures and CURE plots. The results of the exploratory study in this chapter are promising enough to suggest that the BMA, especially in the multi-level model form, can be applied as a unified

model for all possible knowledge sources, either imported or locally developed, for B/A reference groups.

This success will pave the way for a specific local jurisdiction to proceed to the next step of the B/A based on a well-established referential knowledge base. The focus is poised to now switch from the reference to the treated group. The next chapter will contribute to the methodological exploration for assessing similarity of reference and treated groups, which is the last ring in the overall methodological chain of before-after evaluations described by Figure 1-2.

CHAPTER 7 POST-ASSIGNMENT MATCHING BETWEEN COMPARISON AND TREATED GROUPS

Chapters 2 to 6 have focused on reference groups for before-after evaluation (B/A). The purpose of the research presented in those chapters was to address "external validity" (Cook and Campbell, 1979) of before-after evaluations through appropriate data sampling (Chapter 2), to investigate the estimation of unbiased, widely-representative referential information (Chapters 3 and 4), to examine the use of safety surrogates in the event that direct collision estimation is not available (Chapter 5), and lastly to develop a Bayesian Model Averaging process to merge all knowledge sources for SPFs used in before-after evaluations (Chapter 6). The outcome of these chapters is crucial for before-after evaluations since it facilitates the attainment of "generalizability" (Godwin et al., 2003) in safety performance estimation from reference groups so as to minimize the effect of "regression to the mean".

Nevertheless, the work presented in those chapters is not the entirely scope of the dissertation research. In this chapter, the focus of the B/A process turns from the reference group to the treated group. The final measure of a B/A process, the safety improvement resulting from a certain type of treatment, comes from collision comparisons of treated groups, with and without the treatment.

The appropriate assignment or post-assignment treatment of the treated group secures the "internal validity" (Cook and Campbell, 1979), which enhances the final performance of a B/A process, in addition to the "external validity" obtained from procedures developed in the previous chapters. The primary source of the "internal validity" is the appropriate assignment to the treated group from the population. In the traffic safety domain, this is accomplished with the aid of "network screening", one of two major tasks in safety management. As described in Chapter 1, network screening has been a popular topic in traffic safety research with the result that there are well-established methodologies. Accordingly, the assignment of the treated group with the aid of network screening is not the topic of this dissertation research. The focus is on

another major task of traffic safety management – the B/A process, which was also preliminarily explained in Chapter 1. To this end, this Chapter is poised to contribute to the post-assignment handling of the treated group, which is another viable approach to gain "internal validity".

The foundation of a valid B/A is that the reference group shares systematic similarities with the treated group, so as to secure the safety improvement that comes only from treatment, not from any other factors (Dattalo, 2010). Subsequently, the statistical solution of post-assignment handling in this chapter is a "propensity" matching process, which investigates the "comparability" between the reference and treated groups. When a difference is detected, an equivalent adjustment will be made to calibrate the referential estimation, in order to minimize bias.

After carrying out theoretical investigations in Section 7.1, Section 7.2 will utilize the treated group data described in Chapter 1 and, accordingly, a propensity matching process will be conducted between the treated and reference groups. An indicator of "comparability", defined as the "propensity score", will be estimated. Section 7.3 will apply the computed propensity score to conduct a matching process and select the comparison group. Section 7.4 investigates optional applications of the outcomes of propensity score matching. Finally, Section 7.5 will give a brief conclusion for this chapter.

7.1 BASIC CONCEPTS AND PAST RESEARCH

One major goal of statistical research is to identify the causal relationship between variables. The term "research validity" is used to evaluate the merit of a research study, and research validity can be divided into 4 components (Cook and Campbell, 1979; Morgan et al., 2000): (1) measurement reliability and statistics, (2) internal validity, (3) measurement validity and generalizability of the constructs, and (4) external validity.

B/A studies can also be evaluated by these four components of "research validity". Among them, Items (1) and (3) come from measurements of variables, while Items (2) and (4) are taken from

the whole research. That is to say, for the whole research process of a B/A evaluation, major element of success would arise from internal and external validity.

7.1.1 Internal and External Validity

Cook and Campbell (1979) defined internal validity as "the approximate validity with which we can infer that a relationship is causal." Internal validity depends on the strength or soundness of a design, and influences whether one can conclude that an independent variable or intervention had caused the dependent variable to change (Cook and Campbell, 1979).

External validity, in statistics, is defined as "inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments and outcomes" (Glasgow et al., 2007; Shadish et al., 2002). With external validity, causal relationships can be generalized into different measures, persons, settings, and times (Steckler and McLeroy, 2008).

While external validity gains "generalizability", internal validity brings about "reliability and accuracy" (Godwin et al., 2003) for a research study.

For a treatment-centered B/A evaluation, external validity is related to random sampling (RS), i.e., how the sample is drawn from a population. Internal validity, in contrast, is related to random assignment (RA), i.e., how the participants are allocated to the treated and reference groups (Dattalo, 2010). RS helps to approximate results from studying the entire population, and consequently, minimizes the sampling bias and maintains "generalizability". By applying RA, the treated and reference groups share entirely the same characteristics except for the treatment itself, and consequently, maximize the "accuracy" of the treatment evaluation.

"External validity" has been addressed and determined by the methodologies provided in Chapters 2 to 6. "Internal validity", on the contrary, will be further discussed in this chapter.

As mentioned above, internal validity is related to the RA of treated and reference groups. However, one of the pre-conditions of B/A evaluation studied in this dissertation is that the treated group has been selected and the treatment has been conducted in advance.

There are, therefore, alternative ways to obtain internal validity when RA is not available, as detailed in Section 7.1.2.

7.1.2 Concepts and Methodologies of Post-assignment Matching or Adjustment

A. Basic Concepts

The objective of randomization in statistics is to obtain groups that are comparable in terms of both observed and unobserved characteristics. When randomization is not possible, causal inference is complicated by the fact that a group that received a treatment or experienced an event may be very different from another group that did not experience the event or receive the treatment. Thus, it is not clear whether a difference in a certain outcome of interest is due to the treatment or the product of prior differences among groups. There are two ways of overcoming this problem (Coca-Perraillon, 2007):

- adjust the estimates of the treatment effect by using the measured characteristics of each group as covariates in a model, and

- select (or re-select) groups that are similar in terms of observed characteristics before making a comparison, which may still involve some type of model adjustment.

The propensity score method is most often used to facilitate this purpose. In short, propensity score methods have been developed to facilitate the similarities of comparison groups. “Similar”, in this sense, refers to the distribution of observed characteristics (Rubin, 2007). A thorough propensity score matching process comprises two steps, including:

- estimating the propensity score, and

- grouping observations that are similar, or, sample matching.

The first step of propensity score matching is to estimate the likelihood that a sample is assigned to the treatment group given certain characteristics, i.e., the propensity score. More formally, the propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates (D'Agostino, 1998).

The second step, after estimating the propensity scores, uses the scores to group observations that are similar to each other. One way of accomplishing this is to classify treated and untreated observations into subgroups and then separately compare the outcome for each subgroup. This method is usually referred to as "sub-classification" on the propensity scores (Rosenbaum and Rubin, 1984). The other way is to match one treated unit to one or more untreated controls, which is usually referred to as "matching" on the propensity score (Rosenbaum and Rubin, 1983). To implement the second step, two different algorithms are taken into consideration: local and global optimal algorithms. Local optimal algorithms are used to make optimal decisions at each step without attempting to make the best overall (global) decision (Coca-Perraillon, 2007). In contrast, global optimal algorithms borrow from the vast literature on network flows so that the matching problem is to find the path that minimizes the total distance between treated and untreated groups (Rosenbaum, 1989).

B. Methodologies and Processes to Compute Propensity Scores

Propensity scores are the predicted probabilities from a logistic model which models the probabilities of being at the various levels of a predictor of primary interest as a function of a set of secondary variables. In the SAS, this is implemented by the "PROC LOGISTIC" procedure (Leslie and Thiebaud, 2007; Hebert, 2009; SAS Institute Inc., 2012). Here, the propensity score is the conditional probability of each sample receiving a particular treatment based on pre-treatment variables. The logistic model is developed as (Leslie and Thiebaud, 2007) (adjusted to match the context of road safety before-after evaluations):

$$P(Y=1)=\frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}} \quad (7\text{-}1)$$

where

Y=1 if the sample is assigned to the treated group; Y=0 if the sample is assigned to the reference group

X_i are the independent variables included into the logistic modeling

α , β_i are the estimated coefficients

7.2 PROPENSITY SCORE COMPUTATION

7.2.1 Data Applied for Computing Propensity Scores

The data used in computing propensity scores include two types: treated and reference groups.

As previously discussed in Chapter 1, this dissertation has selected urban 4SG intersections as the sample facility, and left turn protection of signalized intersections (also called exclusive left turn signals) as the sample treatment. The sample data are from the City of Toronto, Ontario, Canada. For the treatment of "left turn protection", Toronto has a treated group of 61 intersections. In addition, reference groups include the reference population and two reference groups acquired from a data sampling process provided in Chapter 2. The data for all the groups are compared in Table 7-1.

Table 7-1 actually reveals that there is significant heterogeneity between the treated and reference groups. The mean collisions of the treated group are higher than those of the reference groups, while the AADT levels of the treated group are also higher than those of the reference groups. Besides that, the majority of treated sites have 4 approaches with a left turn lane, which is a different mix from the reference groups. It is expected that this heterogeneity will be further reflected in the propensity scores estimated between each group.

7.2.2 Propensity Score Estimations between Treated and Reference Groups

As mentioned in Section 7.1, propensity scores are computed by the "PROC LOGISTIC" procedure in the SAS software (Leslie and Thiebaud, 2007; Hebert, 2009; SAS Institute Inc., 2012). The reference population and two reference groups were each paired with 61 treated sites, resulting in three datasets to be applied for propensity score computation.

Table 7-1 Measurement Comparisons of Treated and Reference Groups

Measurements	Treated Group (61 sites)	Reference Population (1629 sites)	Reference Group (680 sites)	Reference Group (588 sites)
Mean Multi-vehicle total collisions (SD[a])	79.9(66.6)	62.0 (57.5)	60.8 (59.3)	61.0 (59.0)
Mean Multi-vehicle injury collisions (SD)	35.8(27.6)	16.8 (17.0)	16.5 (17.5)	16.7 (17.5)
Mean Years (SD)	4.0(1.9)	6 (0)	6 (0)	6 (0)
Mean Major AADT (SD)	35267(11719)	13822 (5657)	13213 (6094)	13005 (5861)
Mean Minor AADT (SD)	18096(9729)	3914 (3930)	4117 (4104)	4040 (3897)
No. of Approaches with Left-turn Lanes (%)	0-13.1%; 1-6.6%; 2-9.8%; 3-8.2%; 4-62.3%	0-24.3%; 1-14.9%; 2-29.8%; 3-10.9%; 4-20.1%	0-25.0%; 1-14.6%; 2-29.4%; 3-10.7%; 4-20.3%	0-25.0%; 1-14.3%; 2-29.9%; 3-10.5%; 4-20.2%
No. of Approaches with Right-turn Lanes (%)	0-41.0%; 1-21.3%; 2-21.3%; 3-8.2%; 4-8.2%	0-56.4%; 1-23.3%; 2-13.8%; 3-3.6%; 4-2.8%	0-55.4%; 1-23.1%; 2-13.7%; 3-4.4%; 4-3.4%	0-55.3%; 1-23.5%; 2-13.9%; 3-4.1%; 4-3.2%

Note: a. SD – Standard Deviation

For each dataset, except for the response variable Y (Y=1 for treated group and Y=0 for comparison group), six independent variables (referred to as the "effect" in model estimation) (SAS Institute Inc., 2012) were applied for logistic modeling to estimate propensity scores: annual multi-vehicle total collisions, major AADTs, minor AADTs, number of approaches with left-turn lanes, number of approaches with right-turn lanes, and intersection class.

In the SAS software, the "PROC LOGISTIC" procedure to estimate propensity scores is an iterative process with each step:

- keeping the model convergent as a whole (output of iteration history finally displayed as "convergence criterion satisfied", that is, the relative gradient convergence, GCONV<= 1E-8),
- entering additional effects that meets a significance level of 5% for entry into the model as determined by their p-values from chi-square testing, and
- removing effects that are no longer significant at the 5% level as determined by their p-values from chi-square testing.

The iterative process terminates when there are no more effects that are significant at 5% for entry into the model. Then, the parameters for all of the significant effects are estimated and the propensity scores are directly computed as the fitted values, i.e., the conditional probability for a sample to be assigned into a treated group.

Since the key issue of this part of the study is propensity itself, the following analysis will directly focus on the estimated propensity scores.

In terms of the three pairs of datasets mentioned above, the summarized statistics of their propensity scores are listed in Table 7-2. This Table reveals that the propensity scores of the reference groups are clearly different from those of the treated group, which means that the reference and treated groups do not have secondary variables with the same characteristics (compared to the treatment itself, the primary variable) as the number of collisions, AADT and geometric features (SAS Institute Inc., 2012).

An ideal B/A process relies on one important pre-condition: the comparison group has the same secondary variables as the treated group, so that the differences between these two groups are only from the treatment itself, not from other secondary factors (SAS Institute Inc., 2012). However, in this case, the reference groups have different propensity scores in comparison with the treated group. In this circumstance, a sample matching procedure should be conducted to

select appropriate comparison groups for treated objects based on similarity of their propensity scores (Leslie and Thiebaud, 2007; Hebert, 2009; SAS Institute Inc., 2012).

Table 7-2 Summarized Statistics of Propensity Scores between Treated and Reference Groups (Before Sample Matching)

Datasets	Group	Statistics of Propensity Scores			
		Mean	Standard Deviation	Min	Max
Treated Group + Reference Population	Reference	0.0055	0.0490	2.614E-08	0.9487
	Treated	0.8525	0.2610	1.160E-02	1.0000
Treated Group + Reference Group (sample size=680)	Reference	0.0121	0.0756	1.584E-06	0.9519
	Treated	0.8657	0.2420	2.656E-02	1.0000
Treated Group + Reference Group (sample size=588)	Reference	0.0082	0.0534	1.767E-09	0.6978
	Treated	0.9210	0.2137	1.731E-02	1.0000

7.3 COMPARISON GROUP MATCHING BY PROPENSITY SCORES

Once propensity scores are calculated, there are three common methods that utilize them for the next step in a process (Rosenbaum and Rubin, 1983). These are regression adjustment stratification (sub-classification), and matching.

The first method, regression adjustment, uses propensity scores as additional factors for a GLM or logistic model. The second method involves grouping subjects into strata based on their observed characteristics. The third method matches treated and comparison subjects by their propensity scores (Leslie, 2011).

As for this dissertation research, the models have been well established, so the first method will not be helpful anymore. For the same reason, sub-classification of observations is not needed. Under the current stage of the B/A process, only the third method is applicable, that is, the determination of comparison sites which have similar propensity scores to the treated sites.

For an individually treated site, this procedure involves simply finding another comparison site with the closest propensity score. For a group of treated sites, like the sample treated group with 61 intersections in this dissertation research, there is actually no built-in function in statistical software such as SAS to carry out this procedure (SAS Institute Inc., 2012), despite the many efforts to develop specific macros to implement case-control matching for a certain application (Parsons, 2004).

With respect to individual application cases, the sample matching method could be very flexible and individualized, which is why there is no a built-in function to do so. In other words, users can select their own appropriate method to accomplish matching as long as the proximity of the propensity scores is improved between the treated and comparison groups.

In terms of the research in this chapter, it is clear that the treated group has much higher propensity scores compared to the reference population or reference groups. Hence, comparison group matching can be simplified into a propensity score ranking procedure. This results in the selection of the comparison group with an adequately high enough propensity score that is closest to that of a treated group, in order to significantly decrease systematic heterogeneity between these two groups.

Table 7-3 shows the effects after comparison group matching based on the propensity scores. This Table clearly illustrates that the treated group has systematic heterogeneity with any possible comparison group. However, there are also some patterns that appear in Table 7-3, as well as some that are found when comparisons are made between Tables 7-2 and 7-3. The key observations are as follows:

- Comparison groups have a similarly wide range as the treated group, which means that the former covers the sites that match the treated sites.
- After matching, compared to the scores before the matching is carried out per Table 7-2, the scores of the selected comparison groups in Table 7-3 significantly increase. Through the matching process, the gap between the treated and comparison groups is indeed reduced.
- The treated group has much higher mean propensity scores than any other comparison groups. Even when the comparison group was narrowed down to the smallest group with the closest

characteristics to the treated group, for example, the matched comparison group with 75 sites in Table 7-3 still has much smaller mean propensity scores than its treated group.

- The comparison groups have much higher standard deviations than the treated group, which means that sites in one comparison group have different characteristics while sites in the treated group are relatively more similar.

Table 7-3 Summarized Statistics of Propensity Scores between Treated and Comparison Groups (After Sample Matching)

Datasets	Group	Statistics of Propensity Score			
		Mean	Standard Deviation	Min	Max
Treated Group + Reference Population	Matched Comparison (75 sites)	0.1039	0.2064	9.661E-03	0.9487
	Matched Comparison (212 sites)	0.0398	0.1312	2.618E-03	0.9487
	Treated	0.8525	0.2610	1.160E-02	1.0000
Treated Group + Reference Group (sample size=680)	Matched Comparison (62 sites)	0.1172	0.2263	1.059E-02	0.9519
	Matched Comparison (211 sites)	0.0378	0.1327	2.019E-03	0.9519
	Treated	0.8657	0.2420	2.656E-02	1.0000
Treated Group + Reference Group (sample size=588)	Matched Comparison (65 sites)	0.0702	0.1476	5.033E-03	0.6978
	Matched Comparison (203 sites)	0.0236	0.0891	4.057E-04	0.6978
	Treated	0.9210	0.2137	1.731E-02	1.0000

In this section, comparison groups are generated through a propensity score matching process. This methodology has been found to be functional in making comparison groups more similar with a treated group, so as to decrease comparison bias. However, for the case of the treatment sites with left turn protection in Toronto, after matching, the selected comparison groups still have much lower mean scores than the treated group, which means that they are still systematically heterogeneous. Although this is beyond the scope of this dissertation study, it is worthwhile here to review the conventional mechanism of “network screening” in traffic safety

applications. This is a mechanism that deliberately ranks and chooses sites with the highest expected collision frequency; as a result, the treatment sites have dramatic and systematic heterogeneity compared to other unselected sites. In the Toronto example, the treatment sites with left turn protection are typically intersections with high incidents of collisions, high traffic volumes, greater dimensions and more exclusive left turn lanes. These features clearly set these 61 sites apart from the other Toronto intersections applied as the reference group in this study.

In short, at least for the left turn protection treatment in the Toronto case, attempts to locate ideally matching comparison groups is fruitless due to the network screening mechanism when the treated sites were selected. Notwithstanding this reality, the propensity score matching process is indeed useful to generate comparison groups that have clearly closer characteristics to the treated groups.

7.4 APPLICATIONS OF PROPENSITY SCORE MATCHING

Depending on the purpose, local peculiarities, features of the data, and other possible differences, the application of propensity score matching could be carried out in two ways, as described in detail in the following subsections.

7.4.1 Option 1: Propensity Score Matching for Model Ranking and Selection

The first option is a simple and straightforward application of propensity score matching, without requiring any further data processing. The idea is simply to rank models developed with different reference groups (or samples) with their propensity scores, and recommend the model that is associated with the closest propensity score to that of the treated group. The recommended model would be applied in the final B/A comparison stage, without any further model calibration or adjustments.

In the case of this dissertation study, Table 7-3 shows that the reference group with 688 sample sites clearly obtains higher propensity scores than the reference group with 580 sample sites, i.e.,

the former is much closer to the treated group score, and, as a result, a model to be developed based on the reference group should be recommended for the final B/A comparison if selection is to be made between these reference groups.

7.4.2 Option 2: Propensity Score Matching for Model Adjustment

The second option is to go through model adjustment processing in order to reduce B/A comparison bias due to heterogeneity between the comparison and treated groups.

Once the comparison groups are selected, there are two different ways to utilize the comparison groups, depending on the method applied in the B/A process (see details in Chapter 1).

- For a two-step EB method, comparison groups are physically involved in the B/A process to integrate local referential knowledge into an exported SPF.
- For a one-step FB method, the reference groups are used in the earlier stages of the development of the FB models, rather than in the B/A comparison stage. There is actually no one physical comparison group (subset of reference population or sampled reference groups) that can be used in B/A comparisons, and B/A comparisons are thus completed in a one-step procedure.

In this dissertation study, not only is the FB method applied to develop SPFs, but BMA algorithms are also applied to enhance integration between local knowledge and external SPFs. This was demonstrated in the research work in Chapters 2 to 6. Thanks to these studies, the last stage of a B/A comparison can be completed in one-step, that is to say, there is no need to involve a physical comparison group.

However, this does not mean that the comparison groups generated through propensity score matching are useless. Previous studies have proven that comparison groups have more similar characteristics to treated groups than the whole reference population or groups, and also, this similarity would decrease comparison bias in the estimation of safety benefits which are derived only from the treatment itself, and not from other secondary factors.

So, these comparison groups will finally be used to calibrate the BMA models developed in Chapter 6, which are the models utilized in the final B/A estimation. After this calibration, models used in the B/A comparison would reflect more information from the comparison groups, so as to enhance the comparison accuracy.

The calibration method of the HSM has been described in Chapter 3. Hence, the intermediate steps of calibration are omitted in this chapter, and the final results are shown in Table 7-4.

Table 7-4 Summarized Statistics of Propensity Scores between Treated and Comparison Groups (After Sample Matching)

Datasets	Comparison Groups	Model	Σ (Observations)/Σ(Predictions)		Adjustment Factor
			After Matching	Before Matching	
Treated Group + Reference Population	75 Matched Comparison Sites	Multi-Level	0.5728	1.0216	0.5607
		Full	0.3704	0.6362	0.5822
	212 Matched Comparison Sites	Multi-Level	0.6910	1.0216	0.6764
		Full	0.4374	0.6362	0.6875
Treated Group + Reference Group (sample size=680)	62 Matched Comparison Sites	Multi-Level	0.4999	0.8442	0.5921
		Full	0.2777	0.4457	0.6232
	210 Matched Comparison Sites	Multi-Level	0.6179	0.8442	0.7319
		Full	0.3329	0.4457	0.7469
Treated Group + Reference Group (sample size=588)	65 Matched Comparison Sites	Multi-Level	0.4495	0.8184	0.5492
		Full	0.4674	0.7560	0.6182
	211 Matched Comparison Sites	Multi-Level	0.5452	0.8184	0.6661
		Full	0.5490	0.7560	0.7262

The final output of this calibration is the "adjustment factor" shown in the last column of Table 7-4, which is the ratio of Σ(Observed collisions)/Σ(Predicted collisions) of the comparison group divided by the ratio of Σ(Observed collisions)/Σ(Predicted collisions) of the reference population or reference group. It reflects the differential ratio between the subsets of the comparison groups and their whole reference groups.

Finally, the BMA models established in Chapter 6 will be further adjusted as:

$$\text{Final model} = \text{BMA model} \times \text{Adjustment Factor} \quad (7\text{-}2)$$

The BMA models for applying Equation 7-2 have been discussed in detail in Chapter 6. Among the six Toronto area BMA models, the functional form of three BMA full models was established as Equation 6-10, and their parameter estimates were included in Tables 6-7 to 6-9; the functional form of the three BMA multi-level models was explained by Equation 6-11, and their parameter estimates were included in Tables 6-13 to 6-15. The adjustment factors are listed in Table 7-4.

Finally, the series of models adjusted by Equation 7-2 can be used in B/A comparisons, which will be illustrated by one application example in Chapter 8.

7.4.3 Discussion on Propensity Score Applications

The final adjustments for the left turn treatment of signalized intersections from the Toronto area are shown in Table 7-4.

With reference to Table 7-1, it can be seen that the treated group has higher collisions and traffic volumes, and the majority of the treated sites have approaches with left turn lanes. These features are all clearly different from the reference groups. For the matched comparison groups which have similar characteristics, the estimations proved that the SPF models tended to "overestimate" collisions, i.e., predictions exceeded the actual collisions.

One of the most popular indicators of treatment effects is the collision reduction rate (CRR). It is calculated as (Lan, 2010):

$$CRR = 1 - \frac{\sum_{i=1}^{N_T} \sum_{t=t_Y+1}^{t_Y+t_Z} Y_{i,t}}{\sum_{i=1}^{N_T} \sum_{t=t_Y+1}^{t_Y+t_Z} \lambda_{i,t}} \quad (7\text{-}3)$$

where

$Y_{i,t}$ = observed collisions at site i in year t,

$\lambda_{i,t}$ = expected collisions without treatment for site i in year t in the after period,

t_Y = treatment implementation year

t_Z = the number of years after treatment

N_T = number of treated sites

Given that, after adjustments these models are applied in the final B/A comparison to estimate the CRR, it stands to reason that the CRR would be moderated by those adjustment factors as shown in Table 7-4.

Suppose that there is no adjustment or consideration for propensity, it could then be foreseen that the SPFs without adjustments would overestimate the "after period postulated collisions expected without treatment" at the treated sites. As a result, models without adjustment would tend to exaggerate the safety treatment effects through a deliberately enlarged CRR. This can be seen in Table 7-4 in that observed collisions for comparison groups are much smaller than the predicted collisions. It stands to reason that, because the treated group is similar to comparison groups, the original BMA models would similarly over-predict the treated group collisions without treatment. Thus, it is rational to further adjust the BMA models to lower the predicted collisions, and finally obtain a smaller and more rational CRR.

With rationality, and in principle, the estimation of treatment effects for sites with high collisions, high traffic volumes and wider roadways should be based on similar sites. If comparison groups are also selected with high collisions, high traffic volumes and wider roadways, then it is actually not surprising if a lower CRR is observed rather than estimations without adjustments.

7.5 CHAPTER CONCLUSIONS

While the previous chapters have focused on reference groups, this chapter explores the treated group of a B/A process. The basic principle of this chapter is that accurate estimation of treatment safety benefits can only be done between treated and comparison groups that have "propensity". Otherwise, there will be comparison bias. To reduce this bias, the concept of "propensity scoring" has been applied in this chapter, and the following steps have been carried out to achieve better comparison accuracy:

- propensity score computation,
- comparison group matching through propensity scores, and
- further processing of safety models based on propensity score matching.

There are two options for the last step: models can be simply ranked and selected without adjustments based on their associated propensity scores; or they can be adjusted by using the procedure introduced in Section 7.4.2.

The BMA models described in Chapter 6 would go through optional processing as suggested in this chapter, and the final models would be applied in the final stage of a B/A process; that is, the estimation of a safety benefit indicator – the CRR - for the target treatment.

The framework, procedures and theoretical methodology for the calculation of CRRs have been described in detail in Chapter 1. This process would be different depending on the choice of two B/A approaches, i.e., two-step EB or one-step FB procedure. Based on previous model developments, this dissertation research has applied the latter. In this case, the last stage of a B/A process is a one-step CRR computation as shown in Equation 7-3. This is straightforward and well established. As a result, this dissertation will not go into depth on this stage. Instead, the details will be included as an application example in the next chapter.

CHAPTER 8 APPLICATION EXAMPLE AND DISCUSSION OF DISSERTATION RESEARCH

Chapters 2 to 7 have thoroughly reviewed all of the major research work on before-after evaluations (B/A). The primary purpose of that research sequence was to seek a better methodology that will provide more accurate estimation for the effect of a safety treatment. This goal is ultimately achieved by the outcomes from the Chapters 2 to 7 studies including:

- sample size estimation and data sampling, which obtained valid sample reference groups from a reference population (Chapter 2);
- multi-level SPF development to address local factors (Chapter 3 and Chapter 4);
- use of safety surrogates in the event that direct collision estimation is unavailable (Chapter 5);
- BMA to integrate knowledge sources in order to maximize the representativeness of the final BMA model (Chapter 6); and
- propensity score matching to reduce systematic heterogeneity between the comparison and treated groups in order to minimize B/A comparison bias (Chapter 7).

From a methodological perspective, all of the logic steps for B/A research in this dissertation study have been completed. However, from a practical perspective, there is one remaining item, and that is to apply these methodologies to complete a real B/A process, i.e., to genuinely compute the treatment effect. This chapter serves this purpose in demonstrating the utility of the outcomes from the previous chapters with an application example. Specifically, an attempt is made to finalize the whole B/A process by achieving the final step of calculating the safety benefit.

Following this application example, the second part of this chapter will be devoted to a discussion on the whole methodological framework established in this dissertation with a comparison of this framework with conventional approaches.

8.1 APPLICATION EXAMPLE

The key indicator of treatment effect is the CRR, which is calculated as (Lan, 2010):

$$CRR = 1 - \frac{\sum_{i=1}^{N_T} \sum_{t=t_Y+1}^{t_Y+t_Z} Y_{i,t}}{\sum_{i=1}^{N_T} \sum_{t=t_Y+1}^{t_Y+t_Z} \lambda_{i,t}} \quad (8\text{-}1)$$

where

$Y_{i,t}$ = observed collisions at site i in year t,

$\lambda_{i,t}$ = expected collisions without treatment for site i in year t in the after period,

t_Y = treatment implementation year

t_Z = number of years after the treatment

N_T = number of treated sites

In Equation 8-1, $Y_{i,t}$ is simply the actual observed collisions after the treatment. However, the generation of $\lambda_{i,t}$ could be different depending on which of two different approaches is pursued.

- For the EB approach, the safety model used to estimate "expected collisions" in Equation 8-1 is entirely imported from other sources without local information, so the CRR must be estimated with a two-step procedure. As Step 1, a weighted average estimation (EB) of the "expected collisions" between the observed collisions of the treated group and the predicted collisions from a SPF is made. Then in Step 2, the estimated EB collisions are applied into Equation 8-1 as the "expected collisions without treatment" to calculate the CRR.

- For the FB approach, the safety model itself has already been developed with the integration of local information. Hence, the safety model can be directly applied to estimate "expected

collisions without treatment" in Equation 8-1 so that the CRR is obtained in a one-step procedure.

In this dissertation study, in addition to the FB approach, one additional innovative step is the use of the BMA, which integrates diversified local knowledge into one model, so that the CRR can be directly calculated with a one-step approach.

All of the model developments in this dissertation are based on three different databases: the reference population (same type of intersections/segments in the entire city) and two reference groups with different sample sizes obtained by the sampling approaches developed in Chapter 2. For each database, two different types of BMA models were developed: BMA full model and BMA multi-level model (see more details in Chapter 4). In addition to the BMA approach, Chapter 7 investigated the propensity score matching method to further adjust the BMA models so that there is less comparison bias. Finally, there are different adjustment factors obtained based on the different comparison groups.

As a result, there are a variety of choices for CRR calculations here. As an application example, the case that was applied in Chapters 2 through to 7 will also be used here. That is, urban 4SG intersections are the sample facility, and signalized intersections with protected left turn (also called exclusive left turn signals) installations are used for the treatment. The sample data are from the City of Toronto. For the "left turn protection" treatment, the treated group is comprised of of 61 intersections. (See data on summarized statistics in Chapter 1; in addition, more comparative statistics were included in Chapter 7)

Based on the methodologies developed throughout Chapters 2 to 7, the treatment effect of this application example, with the CRR as the indicator, was calculated based on the above-mentioned different reference groups, models and adjustment factors. Table 8-1 shows a summary of the results.

The statistical analysis in Chapter 3, 4 and 6 suggested that, for Toronto and Edmonton sample data, BMA multi-level models have better fitting performance than BMA full models. In this

dissertation, model performance was evaluated by two types of GOF tests: calculated overall measures and CURE plots (originally introduced in Chapter 3). First, the four calculated overall measures in Tables 6-20 and 6-21 showed that BMA multi-level models fitted better than BMA full models for all Toronto and Edmonton datasets. Besides, the six CURE plots from Figures 6-1 to 6-6 suggest that BMA multi-level models fitted either better than, or equally as well as their full model counterparts.

However, the results indicating better of performance of BMA multi-level models based on Toronto and Edmonton sample data cannot be generalized for applicability to other places without validation. The results cannot be taken as the reason to exclude BMA full models neither. On the contrary, in this dissertation both BMA multi-level and full models were proven as statistically significant. That is why both types of models are retained for the final stage.

The calculations in Table 8-1 based on Toronto sample data suggest that both BMA multi-level and full models tend to "over-estimate" the safety treatment, more so for BMA full models than for their multi-level counterparts. Given that both types of models overestimate, it is safer in practice to select the relatively moderate type. Based on this "principle of caution", the CRRs estimated from the multi-level models should be preferred as estimates of the effect of protected left turning treatments in Toronto area than those estimated from full models.

For the sample treatment and sample area applied in this dissertation, if there is no adjustment, the safety models will tend to "over-estimate" the collisions for the treated group. As a result, after adjustment, in principle the CRRs should be more robust than those based on unadjusted models.

Again, these are conclusions and indications based just one application example and they cannot be generalized for all treatments of all regions without additional supporting evidence.

Table 8-1 CRR Calculations Based on Different Reference Groups, Models and Adjustment Factors

Datasets	Observed Collisions after Treatment	BMA Model	Postulated Collisions without Treatment	Propensity Score Matching Adjustment		CRR Estimated	CRR Recommended
				# of Sites of Comparison Group [1]	Adjustment Factor [a]		
Treated Group + Reference Population	4537	Multi-Level	10263	N/A	No Adjustment	0.5579	0.2116-0.3465
				75	0.5607	0.2116	
				212	0.6764	0.3465	
		Full	21429	N/A	No Adjustment	0.7883	
				75	0.5822	0.6363	
				212	0.6875	0.6920	
Treated Group + Reference Group (sample size=680)	4537	Multi-Level	12832	N/A	No Adjustment	0.6464	0.4029-0.5169
				62	0.5921	0.4029	
				210	0.7319	0.5169	
		Full	31173	N/A	No Adjustment	0.8545	
				62	0.6232	0.7665	
				210	0.7469	0.8051	
Treated Group + Reference Group (sample size=588)	4537	Multi-Level	13527	N/A	No Adjustment	0.6646	0.3893-0.4965
				65	0.5492	0.3893	
				211	0.6661	0.4965	
		Full	17169	N/A	No Adjustment	0.7357	
				65	0.6182	0.5725	
				211	0.7262	0.6361	

Note: a. find more methodological details in Chapter 7.

In terms of the adjustment itself, there are two different types of comparison groups examined: one is smaller and the other is a larger group. The former has systematic features that are more in sync with the treated group (see Chapter 7), but probably higher randomness, while the larger groups have lower randomness but fewer "similarities" with the treated group, so CRR intervals are recommended rather than point estimations, which are bounded respectively by adjustment

factors from the smaller and larger comparison groups based on BMA multi-level models, as shown in Table 8-1.

The recommended CRR is the final indicator of the treatment effect for the example treatment: protected left turns (exclusive left turn signals) for urban 4SG intersections in the City of Toronto.

Theoretically the estimation of the variance (or dispersion) of CRR could act as an important addition to the CRR itself. For example a small variance shows accurate estimation. However, this dissertation achieved this goal in different ways, as follows:

First, the CRR was estimated through referential models. In investigating BMA models this dissertation addressed already the value of referential models, which ultimately was reflected in CRR in that this led to a CRR with less bias, or less dispersion.

Additionally, the recommended CRR had intervals as shown as Table 8-1, in addition to point estimations. This way, the degree of dispersion of the CRR has been taken into consideration.

In conclusion, this dissertation did address the reduction of the variance of the CRR by its focus on model validity and by it application of interval estimation for the CRR.

8.2 DISCUSSION ON METHODOLOGIES DEVELOPED IN THIS DISSERTATION

Analyses, tests and verifications conducted in Chapters 2 to 7 have individually demonstrated that each section in this dissertation study is necessary and beneficial for the B/A process. Compared to previous conventional practices, all of these datasets and methodologies help to gain better treatment effect estimation with less bias and higher representativeness. The following sub-sections summarize discussions of the specific issues addressed.

8.2.1 Datasets and Their Utilities for Before-after Evaluations

As a whole, there are four different categories of datasets developed and applied in this dissertation study: reference population, and reference, comparison and treated groups. Each dataset is essential and useful for the B/A process.

The reference population, in this dissertation, indicates the entire set intersections or road segments (with the treated group excluded) for a jurisdiction with specific characteristics, e.g., all 4SG intersections in a city, all multi-lane highway segments in a province/state, etc. All of the other groups mentioned above are selected from the reference population. For the purpose of the investigation, and supported by relevant research projects, this dissertation research was able to obtain all necessary data items for the two reference populations: 4SG intersections in Toronto and Edmonton (with the treated group excluded). These two reference populations were used to calibrate safety models exported from other regions to a local area and to develop local safety models. Generally speaking, thanks to the large number of samples, the reference population is an ideal database for safety model calibration and development.

However, from a practical perspective, putting together all of the data items for a reference population, especially filling in data items that require manual input or field surveys, is usually unrealistic and also extremely ineffective from a cost perspective. Under this circumstance, an alternative and smaller subset of the reference population is necessary. In this dissertation, the subsets are called reference groups. The reference groups are extracted from the reference population with data sampling approaches (Chapter 2) to ensure that they are appropriate representations of the reference population and at the same time, have an adequate number of sites to calibrate and develop safety models. In Chapters 3 and 4, a variety of models were developed based on the reference groups, and all are statistically significant. This further proves that reference groups can replace reference populations for real world application.

The treated group in this dissertation comprises 61 4SG intersections in Toronto, for which the treatment was protected left turns (exclusive left turn signals). The selection of these 61 sites was conducted before this dissertation study. So from the beginning of this dissertation study, the

assignment of the treated group was already carried out. However, the assignment of the treated group, which typically emerges from a network screening process, is not within the scope of this dissertation study.

Due to the evolution of the treated group from a conventional network screening approach that tends to identify high collision frequency locations for consideration for treatment, this systematic heterogeneity between the treated and reference groups is widespread in road safety practice. If this is the case, another dataset, the comparison group, is essential to reducing comparison bias due to systematic heterogeneity. Comparison groups were extracted from the reference population or groups and had similar characteristics as the treated group. The comparison groups in this dissertation did not have a physical presence in the final B/A comparison since it is not necessary for the one-step comparison approach applied in this dissertation. However, comparison groups in this dissertation are used to adjust the BMA models applied in the final B/A comparison, and there is evidence that these adjustments are necessary. Otherwise, the treatment effects would be distorted. (See Chapter 7 for more details.)

8.2.2 Methodologies and Their Utilities for Before-after Evaluations

All of the methodologies developed in Chapters 2 to 7 are necessary and useful to the B/A process.

Chapter 2 highlighted the methodologies used to identify sample size, and targeted a specific modeling power and data sampling approach to extract reference groups from the population to secure higher feasibility and efficiency for the data collection. This secured the reference group, which, as a subset of the population, maintains statistical consistency with the population.

A variety of methods on developing local models, either analytical or simulated, with either a single or multiple hierarchies, were investigated and estimated in Chapters 3 and 4. They include different model structures, estimation approaches and collection of variables. Instead of developing a single model and ascertaining that it is most suitable, this dissertation is open to all

possible models and took into consideration various information and knowledge sources from different angles.

This dissertation has also investigated safety surrogates in order to utilize them in cases where direct collision measures are not available or are inadequate. This situation is quite prevalent in safety research, so this part of the research is essential.

More importantly, this dissertation has explored an innovative method that integrates all eligible models to form one model. This is the BMA approach. By means of the BMA, no eligible models would be neglected, that is to say, no knowledge source is omitted. The final BMA models take into consideration all viable knowledge sources, either imported or locally developed, to secure a much wider representativeness for the model used to estimate the referential knowledge for a B/A process.

Yet the further adjustment of BMA models based on the propensity score matching approach in Chapter 7 is not carried out in vain. The sample data analysis showed that, without this approach, even if the BMA models were directly applied to the final B/A comparison, i.e., to calculate the CRR, they could still distort the treatment effect. The statistical examples conducted in this dissertation suggest that even after network screening, such an adjustment introduced in Chapter 7 could still further improve the homogeneity between the treated and reference groups. Besides, the approach developed in Chapter 7 is also a sound alternative for those cases without a sufficiently scientific network screening procedure conducted in advance.

In short, the methodologies developed in this dissertation will perform better than the conventional HSM approach in local B/A practices. Practitioners will utilize more locally information for valid treatment effect estimations. As one comparison example, most local applications, due to data and technical difficulty, can only calibrate the HSM model as a whole to yield one single calibration factor, without any locally-specific CMFs. In contrast, methodologies in this dissertation embed local data and specifics into each step of the B/A process so that effects are all locally customized.

8.3 CHAPTER CONCLUSIONS

This chapter has featured one application example to finalize the process of a thorough B/A by completing the estimation of a treatment effect. It has also been used to further demonstrate the rationality and utility of the investigated methodologies and procedures in this dissertation.

The second part of this chapter has further discussed the rationality and utility of all the datasets, methodologies developed and applied in Chapters 2 to 7. Although each individual dataset and method has proven to be essential and useful to the B/A process in each chapter, they are summarized as one in this chapter, in reinforcing the rationality and utility of the entire dissertation study.

In summary, all of the datasets and methodologies developed and applied in Chapters 2 to 7 can be essential and beneficial to the B/A process.

CHAPTER 9 ACCOMPLISHMENTS, CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER STUDY

Chapters 2 to 7 have contributed to filling in the gaps of conventional methodologies for the before-after evaluation (B/A process). Chapter 8 concluded the dissertation research work with an application example and a discussion of the entire research in that context. This chapter is the conclusive summary of the contents in Chapters 2 to 8 and consists of three sections: accomplishments, conclusions and recommendations for future studies.

9.1 ACCOMPLISHMENTS

B/A methodologies are not a new topic in the traffic safety domain. The basic framework for processing before-after evaluations has been well-established and widely applied in practices in the real world for a lengthy amount of time. Both traditional EB and the more recent FB approaches have fueled a substantial amount of research. Nevertheless, there are still some methodological issues that are causing uncertainty and bias in current treatment effect estimations.

In particular, previous research may have concentrated on safety model development itself while paying insufficient attention to the stages before and after model development. This dissertation has sought to balance the research in all of the stages for a thorough processing of the before-after evaluation in order to address all major methodological issues in current applications. In so doing, the following accomplishments have resulted from this dissertation study.

1. Conducting pre-test data sampling to select appropriate local reference groups

If treatment effect estimation is seen as one test, pre-test data sampling is an important and inevitable step to assemble adequate, sufficient and appropriate reference groups for the next step

in model development. Regardless of its importance, few researchers in the road safety field have carried out data sampling prior to modeling. Most research immediately commences from the modeling procedure itself, which is based on existing, arbitrarily selected or readily available reference group, without applying a statistical data sampling process.

This dissertation has challenged this status quo by investigating a data sampling approach to select appropriate reference groups through two consecutive steps: Step 1, which estimates the appropriate sample size to seek a specific level of modeling power (1-B, see details in Chapter 2); and Step 2, in which, in accordance with the pre-estimated sample size, stratified sequential probability proportional to size (PPS) data sampling is conducted to select appropriate reference groups that are adequate for model development, while maintaining statistical consistency with the whole reference population.

2. Development of local safety models with multiple hierarchies, various random distributions and with different approaches

Unlike most safety modeling efforts that only concentrate on one single model, and thus neglecting other choices, this dissertation has developed a variety of local safety models. First of all, model structures have been identified with multiple hierarchies, including single-level full models and multi-level (hierarchical) structures. While multi-level safety models have key merits, such as addressing local specifics while maintaining structural consistency, this dissertation research did not reject the full model.

Traditionally, SPF developments favored NB distribution, e.g., a special case of Poisson-gamma distribution. This dissertation also took NB distribution into consideration, but at the same time, included others from the mixed Poisson family, such as Poisson-lognormal and Poisson-Weibull distributions.

Moreover, this dissertation has applied both the “Frequentist” and “Bayesian” approaches to develop a local model. The former uses the MLE process and yielded fixed model parameters. The latter uses a simulation process and treated parameters as random variables.

The advantage of development via multiple models is clear: any statistically significant (SS_ model would have useful information and different SS models provide different knowledge sources from different perspectives. To keep all of these SS models, means that no useful information and knowledge sources are excluded. The dissertation research has contributed to advancing this philosophy.

3. Converting knowledge from safety surrogates into collision measures

Collision measures are the most favored indicator for safety performance. However, the estimation of collisions relies on adequate historical collision data and these are not always available and sufficient for such estimation. If this is the case, indirect safety measures are to be applied, i.e., safety surrogates. This dissertation has selected the predicted speed of modern roundabouts as the sample and proved its connection with both collision measures and design features, which confirmed that predicted speed can be used as a safety surrogate in the event that collision measures are absent. In so doing, the dissertation has contributed to knowledge on the validity of using safety surrogates.

4. Exploration of Bayesian model averaging to integrate different knowledge sources

In the B/A context, safety models are applied to calculate "postulated collisions without treatments", i.e., referential information. So the most important characteristic of these models is that they must have wide representation. Single models, regardless of their positive attributes, have difficulties in providing widespread enough information.

This dissertation has investigated an innovative approach to integrating all eligible models together without exclusions. This is achieved by the use of the BMA, an approach that takes many eligible models and merges them into one, by averaging their parameters which are weighted by their posterior model probabilities.

This is one of the most important achievements of this dissertation study. It has developed a viable option for traditional model comparison and selection in traffic safety practices that

usually end with a single recommended model that and may be inappropriate for before-after evaluations, as was demonstrated.

The final BMA models developed in this dissertation were tested and found to display better application performance in addition to conceptual superiority.

5. Refining validity of treatment effect estimation by propensity score matching and applying comparison groups to adjusted BMA models

Due to the conventional assignment method of treated groups, they are usually created with very high heterogeneity compared to the reference group. This has led to the observation from the dissertation data that, when BMA modeling is directly applied to compute the "postulated collisions without treatment", there is still the tendency to exaggerate the treatment effect.

In order to solve this problem, this dissertation has investigated a propensity score matching approach that is carried out post-assignment, to generate comparison groups that are relatively more similar to the treated groups. Then, these comparison groups are applied into the calibration process to further adjust the BMA models and secure higher validity for the final treatment effect estimation.

9.2 CONCLUSIONS

After this series of systematic investigations on the B/A methodologies, the following conclusions can be drawn.

- It is important to address, as this dissertation has done, not one, but five different B/A methodologies that comprise the whole process. The fundamental reason is that all of these methodologies are essential for a valid before-after treatment effect analysis and none of them are already well established through previous research. With any one of the methodologies deficient, the treatment effect would be distorted. Also, all five

methodologies are sequentially followed until the treatment effect estimation is optimized when comparison bias is minimal and internal and external validities are maximized. In addition, none of the datasets developed in this dissertation study are redundant. Although the final treatment effect was estimated through comparisons of the treated group itself, and with and without treatments, other datasets including reference population, reference groups and comparison groups were also used to model the "postulated collisions without treatment" of the treated group so they all played their own roles in the treatment effect analysis.

- For the sake of a valid analysis on the treatment effect, to determine all eligible models through different channels, and then finally combine them all together, is more promising than recommending a single model. To serve this purpose, this dissertation has developed models by many means and finally applied the BMA approach to integrate all significant models together. The way that they were developed is not important here; they could be either imported from an external source or locally developed.

- Pre-phase data sampling and post-stage adjustment are as equally important as the model development itself. From a practical perspective, data sampling before modeling means that the data are more accessible, and post-assignment sample matching and model adjustments enhance the internal validity when assignment of a treated group is already carried out beforehand and beyond control. From a theoretical perspective, both pre-phase data sampling and post-assignment sample matching in this dissertation require advanced statistical methods and considerable analytical processing. They are all worthy research topics, but were neglected in previous road safety studies. This dissertation has thus enhanced these two aspects.

- Most significantly, this dissertation has explored and applied alternative model selection and an averaging mechanism that are unique in comparison to conventional model recommendation practices. Rather than recommending one candidate while neglecting all others, this dissertation integrates all eligible models by means of the BMA without exclusion. In the before-after evaluation scenario, this has contributed to less comparison bias and higher external validity.

9.3 FUTURE STUDIES

This dissertation research can be further enhanced and extended as follows.

1. A breakthrough strategy can be explored for the assignment of a treated group which can achieve lower "innate" systematic heterogeneity between treated and reference groups.

2. Innovative safety model paradigms can be further explored, e.g., multi-level model applied in the dissertation can include more dynamic safety attributes in traffic operations and better address local specifics with more flexible parameters.

3. The mechanism to integrate knowledge sources from safety surrogates and collision prediction models can be further investigated. Within the scope of this part of the dissertation investigation, the studied safety surrogate is still an independent referential knowledge that is not merged into the final model. In the future, this should be carried out and knowledge sources from safety surrogates and collision prediction models should be integrated.

4. Further research on other before-after evaluation control factors are beyond scope of this dissertation. One example is, where appropriate data are available, investigation on how the expected treatment effect, or expected collision reduction, could be taken into consideration for reference group size determination based on the new approach recommended by this dissertation.

5. Application products that are friendly to all front-line engineers can be researched and developed in order to translate theoretical research, as was accomplished in this dissertation, into real life engineering practice, through, e.g., incorporation of the research results as additional information in the current HSM and other application tools. One example is the development of applicative tools, e.g., spreadsheet macros or other software that can automatically conduct the before-after evaluation procedures based on methodologies established by this dissertation.

Another example is the construction of CMFs or CM-Functions more accessible to general practitioners. The dissertation focused on models, so development of new CMFs or CM-Functions can follow in future studies, based on components developed in this dissertation.

REFERENCES

1. Aguero-Valverde, J. (2013). Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. Accident Analysis and Prevention 50, pp. 289-297.

2. Ando, T. and Tsayb, R. (2010). Predictive Likelihood for Bayesian Model Selection and Averaging. International Journal of Forecasting 26, pp. 744–763.

3. Akaike, H. (1977). On entropy maximization principle. In Applications of Statistics (Proceedings of Symposium, Wright State University, Dayton, Ohio, 1976), pages 27–41. North-Holland, Amsterdam.

4. Akaike, H. (1978). A new look at the Bayes procedure. Biometrika, 65:53–59.

5. American Association of State Highway and Transportation Officials (AASHTO) (2010). Highway Safety Manual. Washington, DC, United States of America.

6. Amundsen, F. and Hyden, C. (1977). Proceedings of First Workshop on Traffic Conflicts, Oslo, Institute of Transport Economics.

7. Ando, T. and Tsay, R. (2010). Predictive Likelihood for Bayesian Model Selection and Averaging. International Journal of Forecasting 26, pp. 744-763.

8. Bassani, M., Sacchi, E. (2011). Experimental Investigation into Speed Performance and Consistency of Urban Roundabouts: An Italian Case Study. Presented at the 3rd International Conference on Roundabouts, Transportation Research Board, Carmel, Indiana (US), May 18–20, 2011.

9. Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis, Second Edition,

New York: Springer-Verlag.

10. Berger, J. O. and Wolpert, R. (1988). The Likelihood Principle. Second Edition, Hayward, California: Institute of Mathematical Statistics, monograph series.

11. Bernardo, J. M. and Smith, A. F. M. (1994). Bayesian Theory. New York: John Wiley & Sons.

12. Boonsiripant, S. (2009). Speed Profile Variation as a Surrogate Measure of Road Safety Based on GPS-Equipped Vehicle Data. Ph.D. Dissertation presented to the Academic Faculty, Georgia Institute of Technology.

13. Breiman, L. (2001). Statistical Modeling: the Two Cultures. Statistical Science, 16: 199-231.

14. Brude, U. and Larsson, J. (2000). What Roundabout Design Provides The Highest Possible Safety? Nordic Road and Transport Research, No.2.

15. Carlin, B. P. and Louis, T. A. (2000). Bayes and Empirical Bayes Methods for Data Analysis. Second Edition, London: Chapman & Hall.

16. Carriquiry, A., Pawlovich, M.D., 2005. From empirical Bayes to Full Bayes: Methods For Analyzing Traffic Safety Data. http://www.dot.state.ia.us/crashanalysis/pdfs/ eb fb comparison whitepaper october2004.pdf. Accessed on November 19, 2012.

17. Chen, Y. (2010). Interaction of Design Features, Speed and Safety of Roundabouts. Published in Proceedings of the 20th Canadian Multidisciplinary Road Safety Conference, Niagara Falls, Ontario, Canada.

18. Chen, Y., Persaud, B. and Lyon, C. (2011). Effect of Speed on Roundabout Safety Performance – Implications for Use Of Speed As A Surrogate Measure. Peer-reviewed Conference Paper, Transportation Research Board 90th Annual Meeting, Washington D.C., USA.

19. Chen, Y., Persaud, B., Sacchi, E., and Bassani, M. (2012) Investigation of models to relate roundabout safety to predicted speed. Accident Analysis and Prevention (In-Press and published online at: http://dx.doi.org/10.1016/j.aap.2012.04.011).

20. Chen, Y., Persaud, B., and Sacchi, E. (2012) Improving transferability of safety performance functions by Bayesian model averaging. Journal of Transportation Research Record (TRR, accepted and In-Press).

21. Chin, H. C. and Huang, H. (2008). Modeling Multilevel Data in Traffic Safety: A Bayesian Hierarchical Approach. Transportation Accident Analysis and Prevention. Nova Science Publishers, Inc., New York, pp. 53-106.

22. Chou, N. T., Steenhard, D. (2009). A Flexible Count Data Regression Model Using SAS® PROC NLMIXED. SAS Language Reference: Dictionary, Fourth Edition. http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#titlepage.htm. Accessed on June 22, 2011.

23. Chromy, J. R. (1979). Sequential Sample Selection Methods. Proceedings of the American Statistical Association, Survey Research Methods Section, 401–406.

24. Claeskens, G. and Hjort, N. L. (2009). Model Selection and Model Averaging. Cambridge University Press.

25. Coca-Perraillon, M. (2007). Local and Global Optimal Propensity Score Matching. Paper 185-2007, Statistics and Data Analysis, Proceedings of SAS Global Forum 2007, Orlando, Florida, U.S.A.

26. Congdon, P. (2001). Bayesian statistical modeling. Wiley, Chichester. Wiley series in probability and statistics.

27. Cook, T.D., Campbell, D.T. (1979). Quasi-Experimentation: Design and Analysis for Field Settings. Boston: Houghton Mifflin.

28. Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association,* Vol. 91, No. 434, pp. 883-904

29. D'Agostino, R.H. (1998). Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Nonrandomized Control Group, Statistics in Medicine, 17, pp. 2265-2281.

30. Dattalo, P. (2010). Strategies to Approximate Random Sampling and Assignment. Published by Oxford University Press (New York), Inc., 198 Madison Avenue, New York City, New York, U.S.A.

31. Davis, G. A. (2004). Possible Aggregation Biases in Road Safety Research And A Mechanism Approach To Accident Modeling. Accident Analysis and Prevention 36, pp. 1119–1127.

32. Dmitrienko, A., Chuang-Stein, C. and D'Agostino, R. B (2007). Pharmaceutical Statistics Using SAS: A Practical Guide (Chapter 14: Decision Analysis in Drug Development). SAS Institute Inc., Cary, NC, USA.

33. Dougherty, C. (2002). Introduction to Econometrics: Statistical Tables, Second edition, Oxford University Press, Oxford, United Kingdom.

34. El-Basyouny, K. and Sayed, T. (2010). Full Bayes Approach to Before-and-After Safety Evaluation with Matched Comparisons: Case Study of Stop-Sign In-Fill Program, Transportation Research Record: Journal of the Transportation Research Board, No. 2148, Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 1–8.

35. Elvik, R. (2009). Developing Accident Modification Functions - Exploratory Study. Transportation Research Record: Journal of the Transportation Research Board, No. 2103, Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 18-

24.

36. Everitt, B.S. (2002) The Cambridge Dictionary of Statistics, CUP ISBN 0-521-81099-X.

37. Federal Highway Administration (FHWA) (2000). Roundabouts: A Informative Guide (Chapter 5: Safety). FHWA-RD-00-67, Federal Highway Administration of U.S. Department of Transportation, Washington, D.C., USA.

38. Federal Highway Administration (FHWA) (2008). Roundabouts: A Safer Choice. Federal Highway Administration (FHWA) of U.S. Department of Transportation, Washington, D.C., USA. http://safety.fhwa.dot.gov/intersection/roundabouts/fhwasa08006/. Accessed on April 20, 2010.

39. Federal Highway Administration (FHWA) (2009a). Collision Modification Factors Clearinghouse, maintained by the University of North Carolina Highway Safety Research Center, and funded by the Federal Highway Administration (FHWA) of U.S. Department of Transportation.) http://www.cmfclearinghouse.org. Accessed on November 30, 2011.

40. Federal Highway Administration (FHWA) (2009b). Surrogate Safety Assessment Model (SSAM). Publication No. FHWA-HRT-10-020, Turner-Fairbank Highway Research Center of the Federal Highway Administration (FHWA) of U.S. Department of Transportation, 6300 Georgetown Pike McLean, VA 22101-2296.

41. Fitzpatrick, K., Harwood, D. W., Anderson, I. B. and Balke, K. (1999). Accident Mitigation Guide for Congested Rural Two-Lane Highways. National Cooperative Highway Research Program Project 440. National Academy Press, Washington, D.C., USA.

42. Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004). Bayesian Data Analysis. Second Edition, London: Chapman & Hall.

43. Gelman, A. and Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models (Chapter 20: Sample Size and Power Calculation). Cambridge University Press.

44. Gettman, D. and Head, L. (2001). Surrogate Safety Measures from Traffic Simulation

Models. Final Report, FHWA-RD-03-050, Office of Safety Research and Development, Turner-Fairbank Highway Research Center of the Federal Highway Administration (FHWA) of U.S. Department of Transportation, 6300 Georgetown Pike McLean, VA 22101-2296.

45. Gettman, D., Pu, L., Sayed, T. and Shelby, S. (2008). Surrogate Safety Assessment Model and Validation: Final Report. Report No. FHWA-HRT-08-051, Turner-Fairbank Highway Research Center of the Federal Highway Administration (FHWA) of U.S. Department of Transportation, 6300 Georgetown Pike McLean, VA 22101-2296.

46. Glasgow, R. E., Green, L. W. and Ammerman, A. (2007). A Focus on External Validity. Evaluation & the Health Professions, Volume 30, pp. 115-117.

47. Godwin, M., Ruhland, L., Casson, I., MacDonald, S., Delva, D., Birtwhistle, R., Lam, M. and Seguin, R. (2003). Pragmatic Controlled Clinical Trials in Primary Care: the Struggle between External and Internal Validity. BMC Medical Research Methodology, doi:10.1186/1471-2288-3-28. http://www.biomedcentral.com/1471-2288/3/28. Accessed on June 21, 2012.

48. Goldstein, H. (1999). Multilevel Statistical Models. Institute of Education, Multilevel Models Project, London, U.K.

49. Gross, F., Lyon, C., Persaud, B., and Srinivasan, R. (2012a). Safety-Effectiveness of Converting Signalized Intersections to Roundabouts. Peer-reviewed Conference Paper, Transportation Research Board 91st Annual Meeting, Washington D.C., USA.

50. Gross, F., Hamidi, A., Scurry, K. (2012b). Issues Related to Combination of Multiple Crash Modification Factors. Peer-reviewed Conference Paper, Transportation Research Board 91st Annual Meeting, Washington D.C., USA.

51. Hanurav, T. V. (1967). Optimum Utilization of Auxiliary Information: π_{ps} Sampling of Two Units from a Stratum. Journal of the Royal Statistical Society, Series B, 29, 374–391.

52. Harkey, D. L., Srinivasan, R., Baek, J. and others. (2008). Accident Modification Factors for Traffic Engineering and ITS Improvements. National Cooperative Highway Research Program (NCHRP) Report 617, Transportation Research Board of the National Academies, Washington, D.C., USA.

53. Harwood, D.W., Council, F.M., Hauer, E., Hughes, W.E. and Vogt, A. (2000). Prediction of the Expected Safety Performance of Rural Two-Lane Highways. Report No. FHWA-RD-99-207, Federal Highway Administration, Washington, D.C., USA.

54. Heidelberger, P. and Welch, P.D. (1983). Simulation Run Length Control in the Presence of an Initial Transient. Operations Research, 31:1109-1144.

55. Hauer, E. (1985). Observational Before-after Studies in Road Safety. Pergamon Press.

56. Hauer, E., and Bamfo, J. (1997). Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables. Published in Proceedings of International Co-operation on Theories and Concepts in Traffic Safety (ICTCT) 97 Conference, Lund, Sweden, November 5-7.

57. Hebert, P. L. (2009). A Practical Guide to Propensity Score Models. http://www.academyhealth.org/files/2009/monday/HebertP.pdf. Access on October 6, 2012.

58. Ivan, J. N., Ravishanker, N., Jackson, E. and Aronov, B. (2010). Incorporating Wet Pavement Friction into Traffic Safety Analysis. University of Connecticut, Technical Report for Connecticut Department of Transportation under Connecticut Cooperative Transportation Research Program.

59. Joseph, L., Wolfson, D. B. and Du Berger, R. (1995). Some comments on Bayesian Sample Size Determination. The statistician, 44, No. 2, pp. 167-171.

60. Lan, B., B. Persaud, C. Lyon and Bhim, R. (2009). Validation of A Full Bayes Methodology for Observational Before–After Road Safety Studies And Application To Evaluation Of

Rural Signal Conversions. Accident Analysis and Prevention 41, pp. 574–580.

61. Lan, B. (2010). Exploration of Theoretical and Application Issues in Using Fully Bayes Methods for Road Safety Analysis. PhD Degree Dissertation.

62. Lee, P.M. (1997). Bayesian statistics: an introduction. 2nd edition. Arnold, London.

63. Lee, J. Y., Chung, J. H. and Son, B. S. (2008). Analysis of traffic accident size for Korean Highway Using Structural Equation Models. Accident Analysis and Prevention 40, pp. 1955–1963.

64. Leslie, S. and Thiebaud, P. (2007). Using Propensity Scores to Adjust For Treatment Selection Bias. Paper 184-2007, Statistics and Data Analysis, SAS Global Forum 2007, Orlando, Florida, U.S.A.

65. Leslie R. S. (2011). Propensity Score Methods Using SAS. http://www.basug.org/downloads/2011q3/Scott.pdf. Accessed on June 25, 2012.

66. Li, W., Carriquiry, A., Pawlovich, M. and Welch, T. (2008). The Choice of Statistical Models in Road Safety Countermeasure Effectiveness Studies in Iowa. Accident Analysis and Prevention 40, pp. 1531–1542.

67. Lohr, S. L. (2009). Sampling: Design and Analysis. Second Edition, Duxbury Press, Pacific Grove, California, U.S.A.

68. Lord, D., Washington, S. P. and Ivan, J. N. (2005). Poisson, Poisson-gamma and Zero-inflated Regression Models of Motor Vehicle Crashes: Balancing Statistics Fit and Theory. Accident Analysis and Prevention 37, pp. 35-46.

69. Lord, D. and Miranda-Moreno, L.F. (2008a). Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson gamma models for modeling motor vehicle crashes: a Bayesian perspective. Safety Science 46, 751–770.

70. Lord, D., Persaud, B., Washington, S. P., Ivan, J. N., Lyon, C. and Jonsson, T. (2008b). Methodology to Predict the Safety Performance of Rural Multilane Highways. NCHRP Web-only Document 126. http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w126.pdf. Accessed on June 29, 2011, pp. 53-59.

71. Lord, D., Kuo, P. and Geedipally, S. R. (2010). Comparison of Application of Product of Baseline Models and Accident-Modification Factors and Models with Covariates - Predicted Mean Values and Variance. In Transportation Research Record: Journal of the Transportation Research Board, No. 2147, Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 113–122.

72. Luo, Y., Guo, X., Li, H. and Zhu, X. (2008). The Traffic Safety Study Based on Cluster Analysis and Sampling Theory. Proceedings of 2008 International Conference on Intelligent Computation Technology and Automation, 20-22 October, Changsha, Hunan, China.

73. Maycock, G. and Hall, R. D. (1984). Accident at 4-arm roundabouts, Transport and Road Research Laboratory Report, n.1120, Crowthorne, U.K.

74. Miaou, S.P. and Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. Transportation Research Record 1840, 31–40.

75. Morgan, G. A., Gliner, J. A., and Harmon, R. J. (2000). Internal Validity. clinicians' guide to research methods and statistics, pp. 529-531.

76. Muller, K. E. and Peterson, B. L. (1984). Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis. Computational Statistics & Data Analysis, 2, 143–158.

77. Muller, K. E. and Benignus, V.A. (1992a). Increasing Scientific Power with Statistical Power. Neurotoxicology and Teratology, 14, 211–219.

78. Muller, K. E., LaVange, L. M., Ramey, S. L., and Ramey, C. T. (1992b). Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications. Journal of the American Statistical Association, 87 (420), 1209–1226.

79. Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Philosophical Transactions of the Royal Society of London A, 236, pp. 333–380.

80. Oh, J., Lyon, C., Washington, S., Persaud, B. and Bared, J. (2003). Validation of FHWA Crash Models for Rural Intersections. In Transportation Research Record: Journal of the Transportation Research Board, No. 1840, Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 41–49.

81. Parsons L.S. (2004). Performing a 1:N Case-Control Match on Propensity Score. Paper 165-29，, SUGI (SAS User Group International) 29, Montréal, Québec, CANADA.

82. Perkins, D. and Bowman, B. (1986). Effectiveness Evaluation by Using Non-Accident Measures Of Effectiveness. Transportation Research Record, Vol. 905.

83. Persaud, B., Lord, D. and Palmisano, J. (2002). Calibration and Transferability of Accident Prediction Models for Urban Intersections. In Transportation Research Record: Journal of the Transportation Research Board, No. 1784, Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 57–64.

84. Persaud, B., Lan, B., Lyon, C. and Bhim, C. (2010a). Comparison of Empirical Bayes and Full Bayes Approaches for Before-after Road Safety Evaluations. Accident Analysis and Prevention 42, pp. 38-43.

85. Persaud, B., Lyon, C. and Chen, Y. (2010b). Warrants of Roundabouts. Traffic Safety Management Section Report for Ministry of Transportation Ontario under Highway Infrastructure Innovation Funding Program.

86. Persaud, B., Lyon, C., S. Faisal, S. and Chen, Y. (2011a). Adoption of Highway Safety Manual Methodologies for Safety Assessment of Canadian Roads. Final Report for Transport Canada under Canada's National Road Safety Vision (CNRSV) Class Contribution Program.

87. Persaud, B., Lyon, C. and Chen, Y. (2011b). Adoption of Highway Safety Manual and SafetyAnalyst Methodologies for Safety Performance Assessment of Ontario Highways. Traffic Safety Management Section Report for Ministry of Transportation Ontario under Highway Infrastructure Innovation Funding Program.

88. Persaud, B., Saleem, T., Lyon, C. and Chen, Y. (2012a). Safety Performance Functions For Estimating the Safety Benefits of Proposed or Implemented Countermeasures. Draft Report for Transport Canada under Canada's National Road Safety Research and Outreach Program.

89. Persaud, B., Saleem, T., Chen, Y. and Lyon, C. (2012b). Development of Safety Performance Functions to Apply SafetyAnalyst Methodologies for Development and Evaluation of Countermeasures on Ontario Highways. Provincial Highways Management Division Report for Ministry of Transportation Ontario under Highway Infrastructure Innovation Funding Program.

90. Robert, C. P. (2001). The Bayesian Choice. Second Edition, New York: Springer-Verlag.

91. Rodegerdts, L., Blogg, M., Wemple, E., Myers, E., and others (2007). Roundabouts in the United States. National Cooperative Highway Research Program (NCHRP) Report 572, Transportation Research Board of the National Academies, Washington, D.C., USA.

92. Rosenbaum, P.R., and Rubin, D.R. (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika, 70, No. 1, pp. 41-55.

93. Rosenbaum, P.R., and Rubin, D.R. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, Journal of the American Statistical Association, 79, No. 387, pp. 516-524.

94. Rosenbaum, P.R. (1989). Optimal Matching for Observational Studies, Journal of the American Statistical Association, 84, No. 408, pp. 1024-1032.

95. Rubin, D.B. (2007). The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials, Statistics in Medicine, 26, pp. 20-36.

96. SAS Institute Inc. (2012). SAS Help and Documentation. Software of SAS 9.2, Copyright (c) 2002-2008. Cary, NC, USA.

97. Sawalha, Z. and Sayed, T. (2006). Transferability of accident prediction models. Safety Science, No. 44. pp. 209–219. http://journals2.scholarsportal.info/tmp/3944295495924913249.pdf. Accessed on June 27, 2011.

98. Sayed, T. and Zein, S. (1999). Traffic Conflict Standards for Intersections. Transportation Planning and Technology, Vol. 22, pp. 309-323.

99. Sayed, T. and de Leur, P. (2008). Collision Modification Factors for British Columbia. BC Ministry of Transportation & Infrastructure, 2008. http://www.th.gov.bc.ca/publications/eng_publications/safety/CMFs_for_BC_2008.pdf. Accessed on September 30, 2011.

100. Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6:461–464.

101. Shadish, W. R., Cook, T. D. and Campbel, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, Boston, New York, U.S.A.

102. Shankar, V.N., Milton, J.C., Mannering, F. (1997). Modeling Accident Frequencies as Zero-altered Probability Processes: an Empirical Inquiry. Accident Analysis and Prevention

29 (6), pp. 829 – 837.

103. Signorini, D. F. (1991). Sample Size for Poisson Regression. Biometrika (1991), 78, 2, pp. 446-50.

104. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B, 64:583–639.

105. StatTools Home Page (2012). T Test : Table of Critical T Values. http://www.stattools.net/tTest_Tab.php. Accessed on May 7, 2012.

106. Steckler, A. and McLeroy, K. R. (2008). The importance of External Validity. American Journal of Public Health, Vol 98, No. 1, pp. 9-10.

107. Stephens, K. M. (2001). The Handbook of Applied Acceptance Sampling. Milwaukee, Wisconsin: ASQ Quality Press.

108. Stevanovic, A., Kergaye, C. and Haigwood, J. (2011). Assessment of Surrogate Safety Benefits of an Adaptive Traffic Control System. 3rd International Conference on Road Safety and Simulation, September 14-16, 2011, Indianapolis, USA.

109. Tarko, A. P. (2006). Calibration of Safety Prediction Models for Planning Transportation Networks. In Transportation Research Record: Journal of the Transportation Research Board, No. 1950, Transportation Research Board of the National Academies, Washington, D.C. USA, pp. 83–91.

110. Tarko, A., Davis, G., Saunier, N., Sayed, T. and Washington, S. (2009). White Paper - Surrogate Measures of Safety. ANB20(3) Subcommittee on Surrogate Measures of Safety, ANB20 Committee on Safety Data Evaluation and Analysis, Transportation Research Board of the National Academies, Washington, D.C., USA.

111. The R Foundation for Statistical Computing (2012). Software of R version 2.14.2 (2012-02-29), Copyright © 2012.

112. Thompson, H. and Perkins, D. (1983). Surrogate Measures for Accident Experience at Rural Isolated Horizontal Curves. Paper presented at the Annual Meeting of the Transportation Research Board, Washington, DC.

113. Turner, S. A., Roozenburg, A.P. and Smith, A.W. (2006). Roundabout Crash Prediction Models. New Zealand Transport Agency research report 386, New Zealand Transport Agency.

114. Ullah, S., Finch, C. F., and Day, L.(2010). Statistical modelling for falls count data. Accident Analysis and Prevention 42, pp. 384-392.

115. Verkuilen, J. (2009). A Review of Model Selection and Model Averaging. Journal of Educational and Behavioral Statistics, Vol. 34, No. 4, pp. 561–562.

116. Vijayan, K. (1968). An Exact π_{ps} Sampling Scheme: Generalization of a Method of Hanurav. Journal of the Royal Statistical Society, Series B, 30, 556–566.

117. Wang, F. and Gelfand, A. E. (2002). A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. Statistical Science, Vol. 17, No. 2, 193–208.

118. Wasserman, L. (2004). All of Statistics: A Concise Course in Statistical Inference. New York: Springer-Verlag.

119. Warner College of Natural Resources (2011). MARK program help file, software of MARK. Colorado State University, USA. http://warnercnr.colostate.edu/~gwhite/mark/markhelp. Accessed on July 28, 2011.

120. Whitney, M. and Ngo, L. (2004). Bayesian Model Averaging Using SAS@ Software, paper 203-29, SUGI 29 Proceedings, Montreal, Quebec, Canada.

121. WikiHow (2012). How to Calculate Spearman's Rank Correlation Coefficient. http://www.wikihow.com/Calculate-Spearman's-Rank-Correlation-Coefficient. Accessed on May 7, 2012.

122. Wikipedia (2011). Frequentist Inference. http://en.wikipedia.org/wiki/Frequentist_inference. Accessed on July 10, 2011.

123. Williams, R. L. and Chromy, J. R. (1980). SAS Sample Selection Macros. Proceedings of the Fifth Annual SAS Users Group International Conference, 5, 392–396.

124. Yanmaz-Tuzel, O. and Ozbay, K. (2010). A comparative Full Bayesian Before-and-after Analysis and Application to Urban Road Safety Countermeasures in New Jersey. Accident Analysis and Prevention 42, pp. 2099–2107.

125. Ye, Z., Zhang, Y. and Lord, D. (2011). Goodness-Of-Fit Testing For Accident Models with Low Means. Presentation at the 3rd International Conference on Road Safety and Simulation, September 14-16, 2011, Indianapolis, U.S.A.

126. Ye, Fan and Lord, D. (2011). Comparing Three Commonly Used Crash Severity Models on Sample Size Requirements: Multinomial Logit, Ordered Probit and Mixed Logit Models. Presented at the 90th Annual Meeting of the Transportation Research Board, Washington D.C., U.S.A.

127. Zar, J. H. (1972). Significance testing of the Spearman rank correlation. Journal of the American Statistical Association, Vol. 67, No. 339, pp. 578-580.

128. Zou, Y., Lord, D., Zhang, Y. and Peng, Y. (2012). Application of the Bayesian Model Averaging in Predicting Motor Vehicle Crashes. Transportation Research Board 91st Annual Meeting, National Academy of Sciences, Washington D.C., USA.