Improving Predictive Models for Road Safety Screening: Cost-Benefit Data Analytics and Deep Generative Models

by

Mohammad Zarei

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: **Dr. Karim Ismail**
Associate Professor, Dept. of Civil and Environmental Engineering,
Carleton University

Supervisor(s): **Dr. Bruce Hellinga**
Professor, Dept. of Civil and Environmental Eng,
University of Waterloo
Dr. Pedram Izadpanah
Adjunct Professor, Dept. of Civil and Environmental Eng,
University of Waterloo

Internal Member: **Dr. Liping Fu**
Professor, Dept. of Civil and Environmental Eng,
University of Waterloo

Internal Member: **Dr. Chris Bachmann**
Associate Professor, Dept. of Civil and Environmental Eng,
University of Waterloo

Internal-External Member: **Dr. Peter van Beek**
Professor, The Cheriton School of Computer Science,
University of Waterloo

Author's Declaration

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

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

Statement of Contributions

Several of the chapters of this thesis have already been published. Details of the publication and author contributions are provided at the beginning of each such chapter.

Abstract

The process of identifying the sites (e.g., road segments, intersections) within a road network that are most in need of detailed safety audits and interventions with the most potential for improvement is known as road safety network screening. This is often accomplished by developing crash predictive models or safety performance functions (i.e. statistical models that relate site characteristics to the number of crashes) and applying the empirical Bayes (EB) method to determine long-term crash risk and rank sites.

Developing SPFs can be costly and requires specialized skills. Jurisdictions must decide whether to use an outdated SPF or incur the costs of redevelopment but there are currently no methods to determine the impact of using an outdated SPF. In addition to being technically demanding and time-consuming, selecting the optimal parametric model and constructing a functional form that best matches the given data can have a considerable impact on the outcome of network screening. Exposure (usually expressed as the Annual Average Daily Traffic volume or AADT) is an important input for SPF develop, but AADT values are typically estimated from sparse measurements and contain error that can impact network screening outcomes and resource allocation for safety improvements. There is currently no way for road authorities to quantify the impact of these errors and the potential benefits from expending resources to make AADT estimates more accurate. Finally, when samples sizes are small, it is frequently not possible to develop statistically reliable SPFs.

This paper-based PhD research is dedicated to improve the crash predictive models and network screening process by addressing the mentioned gaps, in four main objectives. The first objective, fulfilled in Chapter 2, is to develop a benefit-cost analytic solution to determine the near-optimal time to redevelop SPFs considering the expected benefits of SPF redevelopment and the cost of redevelopment. Practitioners can use the proposed method using the data that is either already available in municipalities which have locally developed SPFs or can be easily estimated.

The second objective, addressed in Chapter 3, is to design a method to quantify the monetary benefit of improving AADT accuracy. A simulation based method is proposed and tested over different crash data conditions to evaluate the sensitivity of network screening outcomes to AADT error. The results showed that the potential consequences of using incorrect Average Annual Daily Traffic (AADT) values can vary greatly depending on factors such as error magnitude, sample size, SPF parameters, and sample mean. This means that the impact of AADT inaccuracy on network screening outcomes is not fixed but must be evaluated individually for each jurisdiction which shows the importance of the proposed method to determine how much effort should be put into improving the accuracy of AADT values.

The third objective, realized in Chapters 4 and 5, is to develop a non-parametric Empirical Bayes (EB) estimation method for modeling crash frequency data. To this end, CGAN-EB is formulated based on Conditional Generative Adversarial Networks (CGAN) which unlike parametric approaches, has no need for a pre-specified underlying relationship between dependent and independent variables in the proposed CGAN-EB and is able to model any type of distribution. CGAN-EB is applied to real-world and simulated crash data sets and outperformed the conventional approach (NB-EB) as a benchmark in terms of model fit, predictive performance and network screening outcomes.

Finally, the fourth objective, addressed in Chapter 6, is to investigate the application of deep generative models as a crash data augmentation method to improve SPFs when the size of the available observational dataset is insufficiently large to reliably develop an SPF. A different CGAN model form is used this time to fit the available observed crash data and to generate more (synthetic) crash data to address low sample size issue for SPF development. The method is evaluated using a real-world and simulated crash data. The results showed that data augmentation could improve the results in terms of reducing coefficient standard error and improving the accuracy of SPF predictions but that these improvements were only statistically significant for specific conditions.

Acknowledgements

I would like to thank my supervisors, Dr. Bruce Hellinga and Dr. Pedram Izadpanah, for their sincere support and encouragement. I also thank the committee members for reviewing this thesis.

Dedication

To my beautiful wife, Raziyeh (Rosie), for her unwavering support, encouragement, and love throughout the duration of my studies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AADT Average Annual Daily Traffic. 4, 5

CGAN Conditional Generative Adversarial Network. 8

COV Coefficient of Variation. 19

CURE Cumulative Residuals. 25

EB Empirical Bayes. 3, 5

EPDO Equivalent Property Damage Only. 38

FI False Identification. 62

GANs Generative Adversarial Networks. 56

GLM Generalized Linear Model. 64

HSM Highway Safety Manual. 3

MAE Mean Absolute Error. 60

MAPE Mean Absolute Percentage Error. 60

MCT Method Consistency Test. 60

NB Negative Binomial. 12

NS Network Screening. 3, 5

PDO Property Damage Only. 38

PDT Prediction Difference Test. 60

PMD Poisson Mean Difference. 62

PPB Percentage PSI Benefit. 38

PSI Potential for Safety Improvement. 5

RDT Rank Difference Test. 60

SCT Site Consistency Test. 60

SPF Safety Performance Function. 3

SPFs Safety Performance Functions. 4–6

TPG Traffic Pattern Groups. 13

Chapter 1

Introduction

In this chapter, a high-level overview of road safety network screening and its challenges are presented. The motivations, objectives, and methodologies used in the research are then discussed.

1.1 An Introduction to Road Safety Network Screening

Traffic crashes pose a significant health and social policy issue in most countries. Apart from grief and suffering, traffic crashes cause enormous social and economic losses, accounting for 1-3% of the gross domestic product of most countries [1]. In Canada in 2018, for example, the annual count of fatal and injury crashes are about 1,700 and 150,000 respectively and the levied societal costs is about $20 billion (1.2% of Canadian GDP) [2]. Taking into account the aforementioned societal cost pertaining to traffic crashes, authorities annually monitor and take actions to reduce severity and frequency of crashes in their road network. Roadway Safety Management Process (RSMP) is a recommended procedure for that reason proposed by Highway Safety Manual [3]. RSMP aims to monitor and reduce crash frequency and severity on the existing road network by following 6 steps: network screening, diagnosis, countermeasure selection, economic appraisal, prioritize projects, and finally safety effectiveness evaluation. Among these steps, this thesis focuses on network screening in which road network locations (sites) are ranked on the basis of the potential for reducing their average crash frequency.

Detecting hotspots[1] in a road network, referred to as NS in the HSM [3], is a statistical analysis procedure that helps determine how to appropriately allocate limited safety improvement resources. The output of NS is a ranked (prioritized) list of hotspots representing locations at which performing further safety reviews and implementing countermeasures are expected to provide the greatest safety improvements. Consequently, the proper prioritization of sites is critical for improving safety and for ensuring that safety improvement resources are effectively used.

Although there are several NS methods in the literature [4, 3], the ones that use crash predictive models (e.g. SPF) with EB method have been shown to provide better results [5, 6, 7] and have become more common in practice. Three major steps of the EB method including their input, outputs, and the corresponding challenges that are investigated in this research are shown in Figure 1.1.

[1] high crash concentration sites

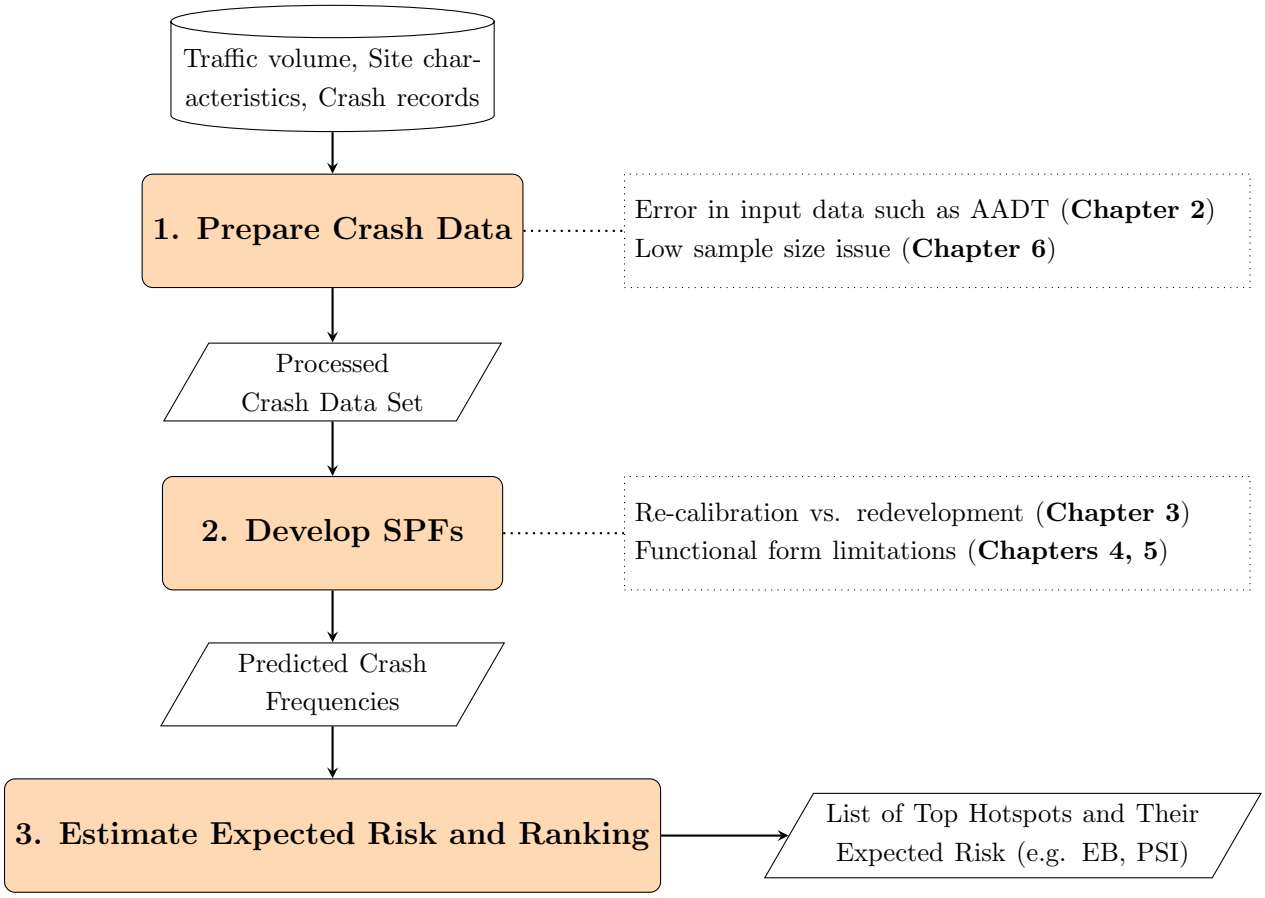


Figure 1.1: Schematic of the components of NS and their challenges

NS begins with collecting and processing geometric and operational characteristics of all sites in the network (traffic volume data, crash records and site characteristics such as traffic control type, number of lanes, lane width, shoulder width, horizonal curvature, speed limit, area (urban vs rural), etc.). The processed data sets are then used for developing SPFs which are usually a set of generalized linear models to predict the frequency of crashes of different severity levels or impact types on the basis of geometric and operational characteristics of each site. Among the explanatory variables, traffic volume data, which is typically reported as AADT, is typically not measured directly (due to the cost of deploying traffic volume sensors) but is estimated through one of several available methods. As a result, AADT values contain uncertainty as well as estimation errors. In addition the the size of the available data set is often less than the recommended sample size for developing reliable SPFs ([8]).

Regardless of the specific SPF-based method [3] used, in order for SPFs to be applied within the NS process, the SPFs must be determined through a development (i.e. fit a model using local data) or re-calibration (i.e. adjust the output of a pre-existing SPFs that has been developed using data from another region or from the local region using older data) process [9]. After developing (or re-calibrating) SPFs, the predicted crash frequencies are used to estimate the expected risks. This is commonly done using EB method to estimate the expected crash frequency (or EB estimates), which is the weighted average of the observed and predicted crash frequency [10]. Finally, sites are ranked based on the EB estimates or Potential for Safety Improvement (PSI) values (i.e. EB estimate minus predicted crash frequency).

In the next section, motivations for this research are presented. Also, note that in this thesis, a literature review is provided in each chapter to give a thorough understanding of the related research and the research gap being addressed in that chapter. This structure also helps to avoid confusion and redundancy.

1.2 Motivations

In light of the NS steps stated in the preceding section, the following issues serve as the motivating factors behind this PhD research:

- One of the inputs for the NS process is a measure of exposure which is typically expressed in term of the annual average daily traffic volume (e.g. AADT). Most jurisdictions operate a traffic data collection program that consists of the deployment of a relatively small number of permanent count stations and the annual collection of short-term counts (typically 8 or 12 hour counts from a single day) at a subset of network locations. These short-term counts may be obtained through manual counts or dedicated sensors temporarily deployed at the location. Various methods [11, 12] can then be used to expand the short term counts to estimate AADT. Given the cost of collecting short term counts, counts are not obtained from all sites and even for those sites at which counts are obtained, counts may not be collected each year. As a result, the available AADT data often suffers from a considerable amount of error and uncertainty [13, 11]. This raises the important question of how sensitive are Network Screening outcomes to AADT accuracy? If outcomes are quite sensitive, then it may be optimal for road authorities to invest resources to improve AADT accuracy as a means to improving NS outcomes.
- Jurisdictions typically conduct NS on an annual basis for the purposes of planning and implementing safety interventions. It is costly for jurisdictions to develop SPFs

every year. While SPF calibration could be an acceptable option, it is known that the reliability of SPFs decreases as time passes due to changes over time in factors such as network structure, enforcement, land use, driving behavior, vehicle types and technologies, demographics of travellers, traffic volumes, and pedestrian and cyclist volumes. Changes in both observed and unobserved factors can affect the relationship between crash frequency and the explanatory variables used in the SPF, thus reducing its predictive accuracy. For observed variables such as traffic volume, changes may occur at a faster rate or in a different manner than initially anticipated when the SPF was developed, leading to discrepancies between the modeled and actual crash frequencies. Unobserved or unobservable factors can also impact the relationship between crash frequency and explanatory variables, posing a threat to the SPF's reliability. Consequently, at some point in time, SPF calibration does not provide a sufficient level of reliability, and SPF development (i.e., using the most recent data) should be done. However, at present, there are no methods by which road authorities can determine when SPF development should be undertaken. The rule-of-thumb in practice is to develop SPFs every 5 to 7 years, but there is no evidence to demonstrate that this frequency is optimal or that it is optimal for all jurisdictions.

- The generalised linear models, which are the most popular choice for SPF development, come with a number of limitations. Most of the time, there is no prior knowledge of which functional forms/distribution will best fit the data. In addition, they have strict hypotheses regarding the error terms and underlying relationship between dependent and independent variables [14] which constrains the ability of the model to fit the data. This raises the question if non-parametric models can be incorporated within the EB process for NS and if such models would perform better than conventional NB-EB models.
- Lastly, it is well recognized that crashes are rare events. This gives rise to the problem of small sample size when developing the SPFs. The desirable large-sample properties of various parameter-estimation approaches (for example, maximum likelihood estimation) are often not achieved with small sample sizes. Various methods have been proposed for overcoming this issue, and one such approach is to augment the observed data with synthesized observations. Recent advances in AI methods may provide an opportunity to address the small sample issue for NS using NB-EB models.

1.3 Research Objectives and Methodologies

The motivations described in the previous section lead to four research objectives. These objectives are identified and briefly described below. The thesis is structured around these objectives and consequently, the chapter(s) in which each objective is addressed is also identified.

Much of the content of this thesis has already been peer reviewed and published in journals or is currently under-review. These publications are identified in the descriptions below.

1: Develop an objective method to determine if SPFs should be redeveloped

SPFs are developed on a yearly basis for three data sets from three regions. A method is defined to capture the actual monetary benefit associated with using the most recently developed SPFs for network screening instead of using recalibrated existing SPFs. The actual benefit is calculated on a yearly basis for each data set. Model(s) are developed to estimate these benefits using available data (e.g. change in AADT, crash counts) to be compared with the redevelopment cost. Results are compared with several benchmarks. The objective is fulfilled in Chapter 2 of the thesis, and has been published as the following journal and conference articles:

Zarei, M., Hellinga, B., & Izadpanah, P. (2022). A Benefit-Cost Based Method to Determine When Safety Performance Functions (SPFs) Should be Redeveloped for Use in Intersection Network Screening. *Transportation Research Record*, 2676(11), 239–249.

Zarei, M., & Hellinga, B. (2022). A quantitative method to determine when safety performance functions used for network screening should be redeveloped. *In Proceedings of the Transportation Research Board (TRB) annual meeting.*

2: Develop a method for estimating the monetary benefit of improving AADT accuracy

A simulation-based method is proposed that can quantify the monetary benefit of improving AADT accuracy. Using the proposed method, a sensitivity analysis is conducted to investigate the sensitivity of network screening outcomes for a range of different conditions. The objective is fulfilled in Chapter 3 of the thesis, and has been published as the following journal article:

Zarei, M., & Hellinga, B. (2022). Method for Estimating the Monetary Benefit of Improving Annual Average Daily Traffic Accuracy in the Context of Road Safety Network Screening. *Transportation Research Record*.

3: Develop a non-parametric EB estimation method for network screening

CGAN are used as crash predictive models and combined with EB method to produce EB estimates. The results are compared with conventional EB estimation method (using negative binomial regression models) in terms of hotspot identification using both real and simulated crash data sets. The objective is fulfilled in Chapters 4 and 5 of the thesis, and has been published as the following journal articles:

Zarei, M., Hellinga, B., & Izadpanah, P. (2022). CGAN-EB: A Non-parametric Empirical Bayes Method for Crash Frequency Modeling Using Conditional Generative Adversarial Networks as Safety Performance Functions. *International Journal of Transportation Science and Technology*, ISSN 2046-0430.

Zarei, M., Hellinga, B., & Izadpanah, P. (2023). Application of Conditional Deep Generative Networks (CGAN) in Empirical Bayes Estimation of Road Crash Risk and Identifying Crash Hotspots. *International Journal of Transportation Science and Technology*, ISSN 2046-0430.

4: Develop a crash data augmentation method for SPF development

A data augmentation method for crash frequency data based on conditional generative adversarial networks (CGAN) is proposed. Method is evaluated in terms of hotspot identification, estimation accuracy, and model goodness-of-fit for real and simulated crash data sets. This objective is fulfilled in Chapter 6, which is currently under review for publication by:

Zarei, M., Hellinga, B. (2023). Crash Data Augmentation Using Conditional Generative Adversarial Networks (CGAN) for Improving Safety Performance Functions. *Transportmetrica*.

Chapter 2

Method for Estimating the Monetary Benefit of Improving Annual Average Daily Traffic Accuracy in the Context of Road Safety Network Screening

This chapter is based on the following journal article:

Zarei, M., & Hellinga, B. (2022). Method for Estimating the Monetary Benefit of Improving Annual Average Daily Traffic Accuracy in the Context of Road Safety Network Screening. *Transportation Research Record*.

In this journal paper I was the first author and was responsible for the writing of the article. The paper was edited by Dr. Hellinga and Dr. Izadpanah. I also developed the crash predictive models and cost-benefit framework and the Python code.

2.1 Introduction

In this chapter, it is attempted to address the first challenge presented in Figure 1.1, namely the presence of error in input data. To this end, a method is proposed that can be used to quantitatively estimate the monetary benefit of improving the accuracy of Annual average daily traffic (AADT) in terms of network screening (NS) outcomes. The use of the proposed method is illustrated through the application to a real-world crash data set. In addition the sensitivity of the network screening results to AADT error in different conditions (i.e. dispersion, sample mean, sample size) is investigated to determine which conditions will benefit the most from AADT accuracy improvement.

Safety performance functions (SPFs) are crash predictive models that estimate the frequency of crashes for different crash types as a function of roadway geometry and operational characteristics. SPFs that can be developed using both statistical models and deep learning models [15, 16] are an integral part of the Road Safety Management Process (RSMP) proposed by the Highway Safety Manual (HSM) [3], a systematic approach by which transportation authorities can identify and prioritize traffic sites (e.g., intersections, roadway segments, and/or ramps) for more detailed safety evaluations. This process is commonly called Network Screening (NS) and the identification and prioritization of roadway locations is also referred to as hotspot identification.

AADT is a commonly used input to safety performance functions (SPFs) which are crash predictive models that estimate the frequency of crashes for different crash types as a function of roadway geometry and operational characteristics, and is frequently estimated by applying temporal factors to expand short-term counts. As a result, there is always some level of error in AADT estimates that can have direct impact on network screening outcomes. Such errors may result in an inefficient use of limited resources for safety improvements.

AADT error can be reduced by either improving estimating procedures [17, 18], collecting counts at more sites (e.g. placing more permanent/short-term count stations), or increasing the duration of short-term counts. The latter two approaches provide more accurate AADT data, but at a higher cost. The challenge for all jurisdictions is to decide how much should be spent to improve AADT accuracy. Assuming jurisdictions make this decision on a benefit/cost basis, then the key question is to determine the benefit of improved AADT accuracy in NS outcomes.

This chapter is structured as follows: The next section provides the relevant background on network screening using EB estimates, AADT estimation methods, and methods that have been suggested to incorporate AADT uncertainty in road safety performance

measures. The Methodology section describes the proposed methodology. The results of implementing the framework are discussed in a sensitivity analysis in the Methodology section and in an example application in Example Application section. Finally, the last section provides conclusions and recommendations.

2.2 Background

2.2.1 Network screening using empirical Bayes (EB) estimates

Network screening is a process of ranking crash hotspots (also known as prone sites or sites with promise) and the first step in the highway safety management process. For the past three decades, an empirical Bayesian (EB) approach has been the most popular method for this purpose [4] in which the long-term expected crash frequency is estimated by combining prior and current information in order to rank sites. The prior information in the EB method comes from calculating a sample mean and variance from a reference group of sites similar to those under evaluation, or from a calibrated safety performance function (SPF) that relates the crash frequency of the reference sites to their characteristics. The point estimates of the expected mean and the variance are then combined with the observed site-specific crash count to obtain an improved estimate of a site's long-term expected crash frequency [19]. In mathematical form, if y is the observed number of crashes at the site, which is Poisson-distributed, and λ is the site's expected long-term average number of crashes; the EB estimator of λ is as follows:

$$EB = w \times E(\lambda) + (1 - w) \times y \tag{2.1}$$

where EB denotes the empirical Bayes estimate of the site's expected long-term crash counts, $E(\lambda)$ is estimated by the crash prediction model, and the weight w is a function of the mean and variance of λ and is always a number between 0 and 1.

Many parametric and non-parametric crash prediction models have been used in the literature. For parametric models, the NB [20], the Poisson-lognormal [21], the Conway–Maxwell–Poisson [22], the Poisson-Tweedie [23], the Sichel [24], and for non-parametric models deep neural networks [16, 25], and support vector machines [26] are some examples.

Among all the above-mentioned models, the NB model (Eq. 2.2) has been the most frequently used in practice for predicting crashes and the corresponding EB estimate can be calculated as follows:

$$f(y|\mu,\alpha) = \frac{\Gamma(y+1/\alpha)}{\Gamma(1/\alpha)\Gamma(y+1)}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)\left(\frac{1}{1+\alpha\mu}\right) \tag{2.2}$$

$$EB^{NB} = w \times \mu + (1-w) \times y \tag{2.3}$$

$$w = \frac{1}{1+\alpha\mu} \tag{2.4}$$

where y is the observed number of crashes per year, α is a dispersion parameter, and μ is the number of crashes predicted by the NB model given site inputs such as AADT, geometric characteristics, etc.

2.2.2 AADT estimation methods

One of the most challenging inputs for the network screening is AADT data which often suffers from a considerable amount of error and uncertainty [13]. Installing permanent count stations or traffic detection cameras on every link in the network is impractical and cost-prohibitive and consequently several methods are described in the literature and used in practice to estimate AADT based on site specific short-term counts and data from nearby permanent count stations and/or site characteristics. The traditional factoring approach [27], which is most commonly used in practice, divides the permanent count station sites into a set of TPG such that all sites within a TPG have similar temporal traffic patterns. Expansion factors (hourly/daily/monthly) are computed for each TPG on the basis of the counts from the permanent count stations associated with each TPG. Then for each site in the network for which AADT is required, the associated TPG is identified and the AADT is estimated by expanding the short term count from the site with the TPG expansion factors. The AADT estimates from this method are subjected to error due to incorrect grouping of permanent counts stations into traffic pattern groups, incorrect association of a TPG for the short term count site, and the random day-to-day variability in the traffic counts [18].

A number of alternate AADT estimation approaches have been proposed to improve AADT estimation accuracy. Tsapakis and Schneider proposed a support vector machine method to determine from which factor group expansion factors should be obtained for a given short term count [28]. Other approaches based on geostatistical or machine learning methods have been proposed that make use of site and land use characteristics (both

with and without short term count data) to estimate AADT. For example, segment-based modelling has been demonstrated to significantly improve the accuracy of heavy vehicle volume predictions by integrating spatial characteristics and homogeneity of road segments [29, 30]. Lam et al proposed a neural network model to expand short-term counts in Hong-Kong and compared it to a regression model [31]. Similarly Sharma et al proposed a neural network model to estimate AADT expansion factors for low volume roads [18]. The input for the network was input factors which are the hourly volumes divided by the total volume for one or more short-period traffic counts in the sample divided by equivalent number of sample days (this was named sample average daily traffic - SADT). The target output is the actual expansion factor which is the true AADT divided by total volume for one or more short-period traffic counts in the sample. More recently, Tawfeek and El-Basyouny applied a deep neural network (DNN) to estimate AADT of minor roads at rural, stop-controlled intersections and compared with the estimation from a linear regression model [32]. Model input included AADT of major road and site characteristics such as number of lanes, presence of median/right-turn/and left-turn lanes, and distance to nearest urban area. The results show about 35% improvement in estimation of minor road AADT in comparing with the linear regression method. In addition they showed that the SPFs developed from the neural network estimations of the minor road AADT showed better fit to the data.

It has also been proposed to use various other complementary data to improve the accuracy of AADT estimation. Jiang et al showed that AADT estimation accuracy could be improved through the use of aerial imagery [17]. More recently, Zhang and Chen showed that using vehicle probe data can significantly improve the accuracy of AADT estimates [33]). Other recent studies have examined the use of GPS data [34] and cell phone data [35] for improving AADT estimates.

Despite the wide spread use of traditional AADT estimation methods and the ubiquitous need for AADT data, there are few studies that quantify the magnitude of the error (uncertainty) in these AADT estimates. A study by the FHWA (2015) is likely the most comprehensive examination of AADT estimation accuracy and the factors that influence accuracy (e.g. short term count duration, methods for forming TPGs, and types of expansion factors) [36]. This study used 13 years of data (24 hour counts from 365 days of the year) from 320 permanent count stations capturing a wide range of AADT for 9 different road functional classes across 32 states in the USA. A Monte Carlo process was used to repeatedly draw (1000 trials) a randomly selected short term count from the year of count data available for each site and then to estimate the AADT ($AADT_{Est}$). Relative estimation error is reported as $\frac{AADT_{Est}-AADT_{Obs}}{AADT_{Obs}}$, where $AADT_{Obs}$ is the AADT computed from all 365 days of counts from that site. The study reports the 95% confidence limits

of the relative error (uncertainty range) aggregated across all road classes and states to be -28% to 36% when short term counts are 24h in duration and taken during weekdays. However, the report also shows that this uncertainty range varies substantially depending on the road functional class and state (the half width of the uncertainty range varied from less than 10% to 70%). The study does not characterize the shape of the distribution.

Milligan et al characterized the accuracy of AADT estimated from short-term counts using data from 69 permanent count stations in Manitoba, Canada [37]. They performed their study in a similar manner to the FHWA study but instead of computing expansion factors from a factor group containing multiple permanent count stations, they computed the expansion factors from the single closest permanent count station belonging to the same traffic pattern group as the short term count site (they term this the IPC method). Using the IPC method and their set of data, they found the relative estimation error to follow a Normal distribution with an uncertainty range of -50% to 50% when using a 24h duration short-term count.

If it is assumed relative errors are normally distributed with a mean of zero, then from the above two studies, we can conclude that the standard deviation of the relative estimation error ranges from 5% to 35%.

Note that Milligan et al also provide a summary of the AADT estimation errors from a number of older studies from the literature. These studies either do not report the uncertainty range or do so for conditions that are not commonly encountered in practice (i.e. short term count duration of a month or greater).

2.2.3 Incorporating AADT uncertainty in network screening

The importance of AADT uncertainty to road safety analysis, and specifically network screening, has also been studied. Several studies have attempted to quantify AADT estimation errors and then formulate network safety regression modelling frameworks that directly include this error. Maher and Summersgill investigated AADT uncertainty in modeling crash data with Poisson distribution using two approaches [38]. In the first, a simulation approach is used to change AADT according to a lognormal distribution with a variance to mean ratio of 10 %. The safety performance function (SPF) calibrated on the simulated data showed about 20% smaller values for coefficient estimations than SPFs developed based on original AADT data. In other words, randomizing traffic volume variables lead to a bias (underestimate) in their corresponding SPF coefficient. In the second approach, they proposed a formal functional model to explicitly account for AADT uncertainty in the log-likelihood function. In the proposed functional form, crash and AADT

values have different Poisson models which can be integrated into one log-likelihood function for parameter estimations. Then using an iterative algorithm, the maximum likelihood of solution for parameters can be found. The simulation results indicate the mentioned biases in parameter estimation are very much reduced when the modified functional form is applied.

El-Basyouny et al proposed the Measurement Error Negative Binomial (MENB) model, to consider uncertainty in the AADT estimates and to provide a better estimate of the over-dispersion parameter [39]. The MENB approach, which is inspired from models in epidemiological research, has three elements of response (i.e. dependent variable), exposure (i.e. traffic flow) and measurement error (i.e. uncertainty in traffic flow). The MEBN method outperforms (i.e. better goodness-of-fit results) the conventional negative binomial approach in the case of large AADT measurement errors. In a more recent study, Musunuru and Porter simulated AADT estimates considering a Normally distributed measurement error and compared the corresponding SPF models to the SPF model developed on the true AADT values [40]. They found that when measurement error is not included in road safety regression modeling, regression coefficient estimates are biased toward zero.

All of the mentioned studies have proposed different methods to either improve the AADT accuracy or include the uncertainty of the AADT data in order to obtain more reliable SPFs. However, to the knowledge of the authors, there is no framework in the literature that can quantify the magnitude of the potential benefit associated with either improving AADT estimates or incorporating AADT errors into modeling process, in terms of network screening results. The contributions of this chapter are as follows:

1. A simulation-based method for cost-benefit analysis of improving AADT estimates is proposed.
2. Using the proposed method, a sensitivity analysis is conducted to investigate the sensitivity of network screening outcomes for a range of different conditions.
3. The application of the proposed method is presented in a real-world example.

.

2.3 Methodology

In this section, a method that can quantify the impact of AADT error in terms of network screening outcomes is proposed. Several different metrics have been used to quantify the

accuracy of network screening in this regard. Our goal in this chapter is to quantify this impact in terms of monetary loss so that jurisdictions can make decisions on a benefit-cost basis. The proposed method consists of four modules as presented in Fig. 2.1. Each module is described in the following subsections.

2.3.1 M1: Crash Data Simulation Module

The method begins with simulating true long-term expected crash frequencies (k) and the crash counts (y) based on the given SPF, dispersion parameter (α), AADT data and other features of the sites (X) using the process which has been proposed in [41] and been used in previous simulation studies [42, 43]:

$$y \sim Poisson(k);$$

$$k = \mu \times e^{\epsilon};$$

$$\mu = \text{SPF}(AADT, X)$$

$$\exp(\epsilon) \sim gamma(1, \alpha);$$

$$PSI = \max(k - \mu, 0).$$

Note that this module also finds the hotspot true ranking (R in Figure 2.1) based on true potential for safety improvement (PSI) values, which is the difference between k and what is normal for similar sites [44]. As a result, this ranking is the optimal ranking of hotspots based on the simulated crash count with the highest potential for crash reduction given appropriate counter measures.

2.3.2 M2: Modified AADT Generation Module

Following the same approach adopted by [40], in this module, the AADT for each site is modified to include the measurement error (ϵ) given the measurement error variance σ^2 as follows:

$$AADT_{modified} \sim \text{Normal } (AADT, \sigma^2) \tag{2.5}$$

$$\sigma = AADT \times \nu \tag{2.6}$$

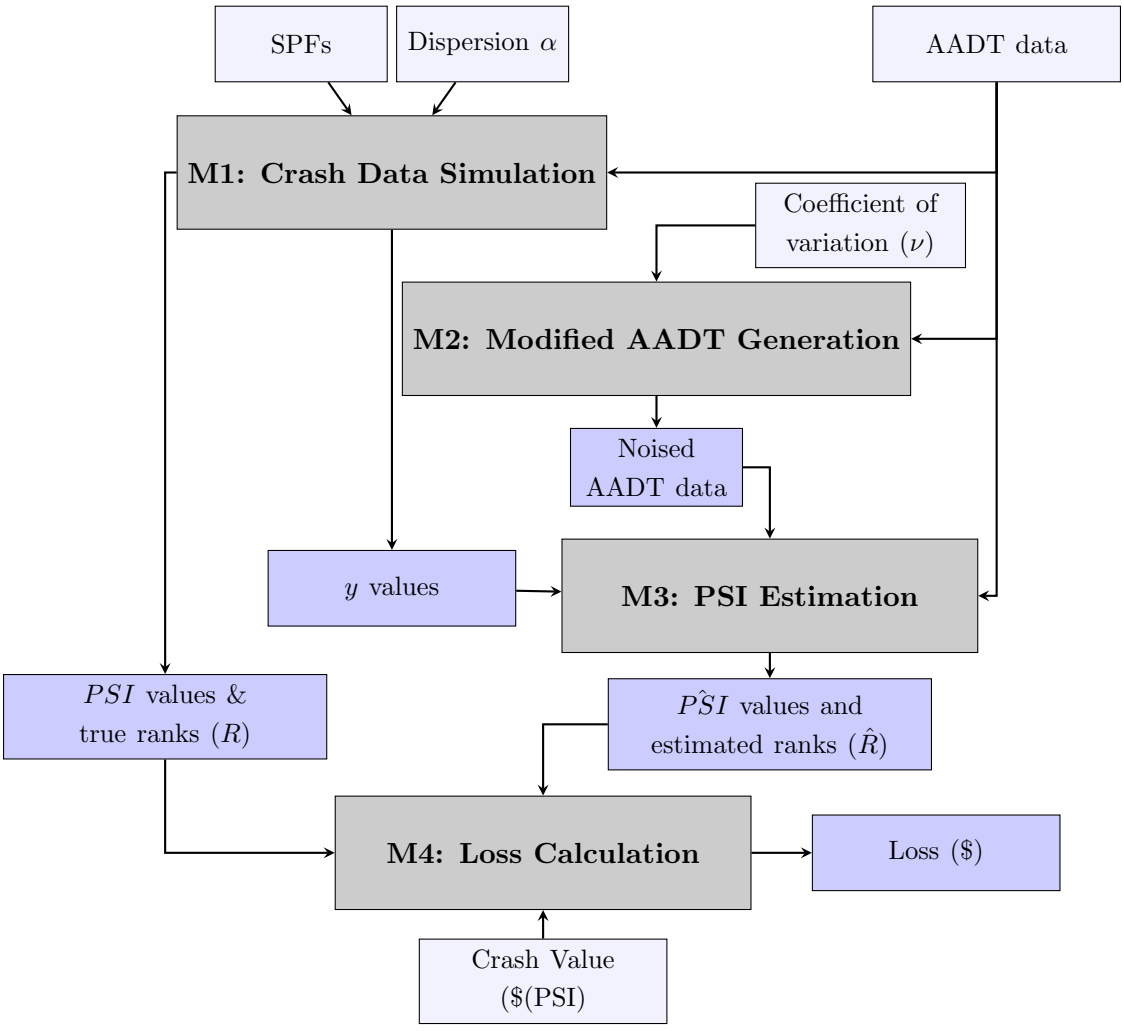


Figure 2.1: Input/output flow of four modules in the proposed framework

where $AADT$ is the true AADT value, σ^2 is the error variance, and ν is COV. COV can be estimated using the estimated AADT values for a sample of sites with known AADT. With the modified AADT values we are trying to replicate the real world condition in which the AADT estimates include some amount of measurement error.

2.3.3 M3: PSI Estimation Module

Module M3 computes the EB estimates given the modified AADT data and simulated crash counts (K) by fitting a NB model. Then $\hat{PSI}$ is computed [44]:

$$\hat{PSI} = \max(EB^{NB} - \hat{\mu}, 0) \tag{2.7}$$

where EB^{NB} is based on Eq. 2.3 and $\hat{\mu}$ is the predicted number of crashes by the fitted NB model. Using the $\hat{PSI}$, the ranking based on modified data is then calculated (these rankings are denoted $\hat{R}$).

2.3.4 M4: Loss Calculation Module

When erroneous AADT data leads to incorrect ranking of sites, then sites with a lower PSI are included in the list of top n hotspots in place of sites that actually have a higher PSI. Given that the objective of identifying the top n hotspots on the basis of PSI is to be able to concentrate resources and implement safety improvement measures at those sites where the potential for safety improvements is the greatest, incorrectly identifying sites results in the application of resources to lower priority sites and represents an opportunity cost [45] which is called *potential loss*. This module estimates a monetary value of the potential loss ($Loss_\nu$) due to using modified AADT based on the value of COV (i.e. ν). To this end, the PSI Difference (ΔPSI) is computed for top n hotspots given the estimated rankings from the modified data ($\hat{R}$), true PSI values, and the true ranking (R):

$$\Delta PSI_{\nu,n} = \sum_{R,n} PSI - \sum_{\hat{R},n} PSI \tag{2.8}$$

where $\sum_{R,n} PSI$ is the sum of PSI values for the true top n hotspots, and $\sum_{\hat{R},n} PSI$ is the sum of PSI values for the top n hotspots identified using the modified AADT data. ΔPSI reflects the difference in the PSI values across the true top n hotspots and the PSI values across the n hotspots when using the AADT data containing error. The social cost of crashes can be established by crash type (e.g. property damage only (PDO) crashes,

injury crashes, and fatal crashes) [46, 3, 47]. Consequently, the loss can be expressed as a monetary value as follows:

$$Loss_{\nu} = \Delta PSI_{\nu,n} \times \$(\text{PSI}) \tag{2.9}$$

$$\$(\text{PSI}) = \frac{N_{PDO} \times \$_{PDO} + N_{Injury} \times \$_{Injury} + N_{Fatal} \times \$_{Fatal}}{N_{PDO} + N_{Injury} + N_{Fatal}} \tag{2.10}$$

where N and \$ represent the number and societal cost of crashes of a given severity type.

2.4 Sensitivity Analysis

In this section, the proposed method is used to investigate the sensitivity of NS outcomes for a range of different conditions. In order to simulate conditions that are convincingly similar to the range of conditions typically observed in empirical crash data, the parameters of these experiments (i.e. dispersion, sample mean and sample size) are based on the reported parameters related to crash datasets in eight published chapters summarized in [42], as follows:

- Four dispersion parameter (α) values: 0.25, 0.5, 1.5, 3.0
- Three sample sizes (i.e. number of sites): 250, 500, 1000
- Four sample means (crashes/year): 0.7, 1.9, 5.1, 11.3

For the purposes of the sensitivity analysis, it is assumed that sites are road segments and that segments are all of the same length. Consequently, the following SPF is used:

$$\hat{\mu} = exp(\beta \times ln(AADT) + \beta_0) \tag{2.11}$$

where β is set be to 1.18 and β_0 is set to be -13.15, -12.15, -11.15 and -10.35 for sample mean of 0.7, 1.9, 5.1 and 11.3 respectively. These values are chosen to be close to SPF parameters fitted to the available data. A real AADT data set is used for sampling the AADT values for each scenario as a means of ensuring a realistic distribution of values. The data is from the Highway Safety Information System (HSIS) collected for divided 4-lane segments from urban freeways in Washington State for the year 2017.

Using the above-mentioned setup and the framework presented in Figure 2.1 the below steps are followed for each scenario:

1. For each scenario and AADT sample, M1 simulates the long-term expected crash frequency (k) and the observed crash counts (K), and computes the true ranking (R) based on PSI.
2. M2 modifies the values of the AADT sample based on a given coefficient of variation (ν in Equation 2.6).
3. M3 uses the simulated observed crash counts (K) and the noised AADT data to compute the EB estimates, PSI, and hotspot ranking ($\hat{R}$).
4. M4 compares the suggested ranking from M3 ($\hat{R}$) with the true ranking (R) for the top 10% hotspots and calculates the ΔPSI values.

The ΔPSI results for 48 scenarios are presented in Figure 2.2 as box-plots. Each box-plot represents the ΔPSI results over 200 replications for each ν value (ranging from 0% to 100% in 20% intervals). The mean values of the replications are shown by a green triangle. The percentage change of ΔPSI mean values as a function of COV are shown in Figure 2.3.

From these results it can be observed that the sensitivity of network screening varies across different ν values but there is a general increase in ΔPSI as ν increases. In some conditions, errors in AADT values seem to have very little effect in terms of network screening outcomes while in others the network screening outcomes appear to be quite sensitive to the accuracy of AADT.

In Equation 2.8, larger sample sizes provides larger ΔPSI (Figure 2.2). This is expected as ΔPSI is calculated over top 10% hotspots which include more sites for larger sample sizes. To avoid this scale effect, the relative change in ΔPSI is computed as:

$$\%\Delta PSI(\nu, n) = \frac{\Delta PSI_{\nu,n} - \Delta PSI_{\nu=0,n}}{\Delta PSI_{\nu=0,n}} \times 100 \tag{2.12}$$

Based on these relative changes as depicted in Figure 2.3, it can be seen that the impact of sample size on the increasing trend of ΔPSI versus ν is negligible.

The results show that the impact of AADT errors on network screening outcomes is mostly impacted by the average crash frequency (i.e. sample mean) and the magnitude of

Figure 2.2: ΔPSI results

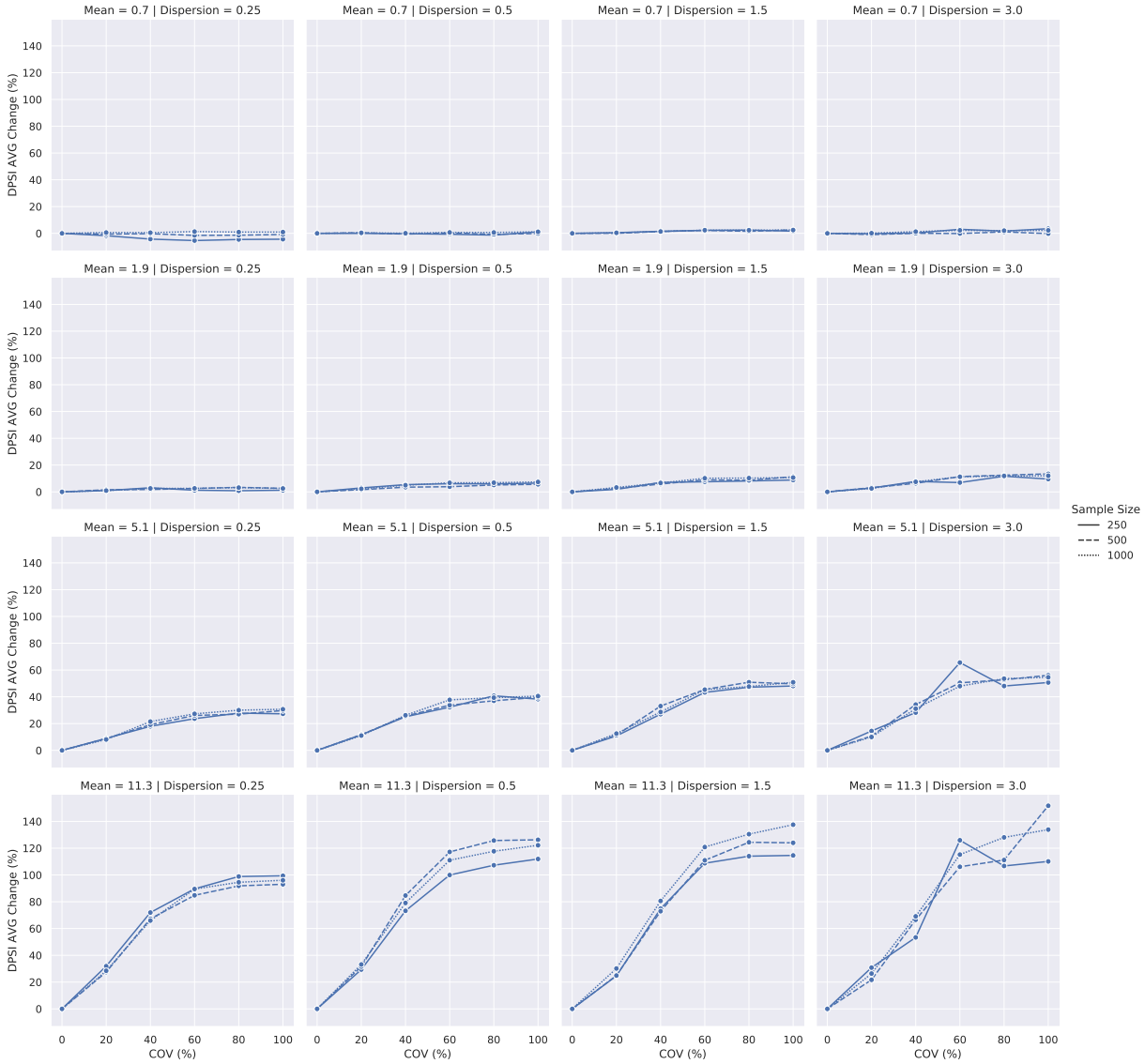


Figure 2.3: % Change of ΔPSI (shown as DPSI in the graphs) mean values

the AADT error. The sample size and dispersion parameter value impacts are negligible. This is evident from the plots in Figure 2.3. When mean = 0.7 crashes/year/site, for all dispersion values, the relative change for the average of ΔPSI values are mostly less than about 5%. In contrast, the relative change reaches a maximum of approximately 60% and 120% when sample mean is 5.1 and 11.3 respectively. These observations can be explained by rewriting PSI equation based on model prediction (μ), observed crash number (y) and dispersion parameter (α) as follows:

$$\begin{aligned} PSI &= \max(EB - \mu, 0) \\ &= \max((1-w) \times (y-\mu), 0) \\ &= \max\left(\frac{\alpha\mu \times (y-\mu)}{1+\alpha\mu}, 0\right) \end{aligned} \tag{2.13}$$

Based on Equation 2.13, PSI has a non-zero value when $y > \mu$. When the sample mean is very small, then most sites have zero crash counts (i.e. y) with $PSI = 0$ and the top hotspots will always include those few sites with non-zero crash counts. Consequently, when the sample mean is very small then PSI-based ranking mostly depends on crash counts y which is not influenced by AADT accuracy. Another way to explain this is to take the derivative of PSI with respect to μ that is the main term in Equation 2.13 directly affected by AADT error:

$$\frac{d(PSI)}{d(\mu)} = \frac{\alpha y + 1}{(1+\alpha\mu)^2} - 1 \tag{2.14}$$

In Equation 2.14, for very small sample means, y and μ tend to be close to zero resulting in $\frac{d(PSI)}{d(\mu)}$ being close to zero. On the other extreme, when the sample mean is very large, large y and μ values will be expected which makes $\frac{d(PSI)}{d(\mu)}$ become close to -1. Similarly, for very small dispersion parameter α, $\frac{d(PSI)}{d(\mu)}$ becomes zero. In other words, PSI is more sensitive to μ as the only term that can be affected by AADT error in Equation 2.13 when sample mean and dispersion parameter are larger. And this is exactly what is observed in the results in Figures 2.2 and 2.3.

The results in this section suggest that the potential loss related to using inaccurate AADT values can have a wide range of impacts on the network screening outcomes depending on conditions such as the magnitude of the inaccuracy, sample size, SPF parameters, and sample mean. This means that the impact of AADT inaccuracy on network screening

outcomes is not constant but must be quantified for each jurisdiction (maybe even each location type) separately. Thus there is an increased need for utilizing the proposed method to estimate how much investment in improving AADT accuracy is warranted. In the next section, a real-world data set is used to show the application of the proposed method for a specific jurisdiction.

2.5 Example Application

For this section, the crash data set from 2014-2017 for 865 intersections located in London, ON, Canada are used. There are four types of intersections as follows:

- G1: Signalized 4-legged intersections
- G2: Signalized 3-legged intersections
- G3: Unsignalized 4-legged intersections
- G4: Unsignalized 3-legged intersections

The SPFs are developed for each group separately and the results are presented in Table 2.1. Note that the crash mean and dispersion parameter for each intersection group type are within the range of the sample mean and dispersion parameter values evaluated with the simulation step, providing additional confirmation that the simulation sensitivity analysis considered a realistic range of crash dataset characteristics.

Table 2.1: Developed SPFs for each intersection group

Group	**SPF**	**Dispersion**	**Crash Mean**	**# Sites**	**AADT (mean)**
G1	$e^{(1.68\times AADT-15.3)}$	0.2	8.3	315	28900
G2	$e^{(0.704\times AADT-5.99)}$	0.7	3.1	87	24900
G3	$e^{(0.732\times AADT-6.48)}$	0.4	1.0	200	8650
G4	$e^{(0.686\times AADT-6.74)}$	0.8	0.6	263	1120
All	-	0.5	4	865	15400

Note: All SPF coefficients are found to be statistically significant (p-value < 0.05).

CURE plots were developed to understand how well the SPFs from Table 1 fit the data (Figure 2.4). The residuals show the differences between historical (observed) and

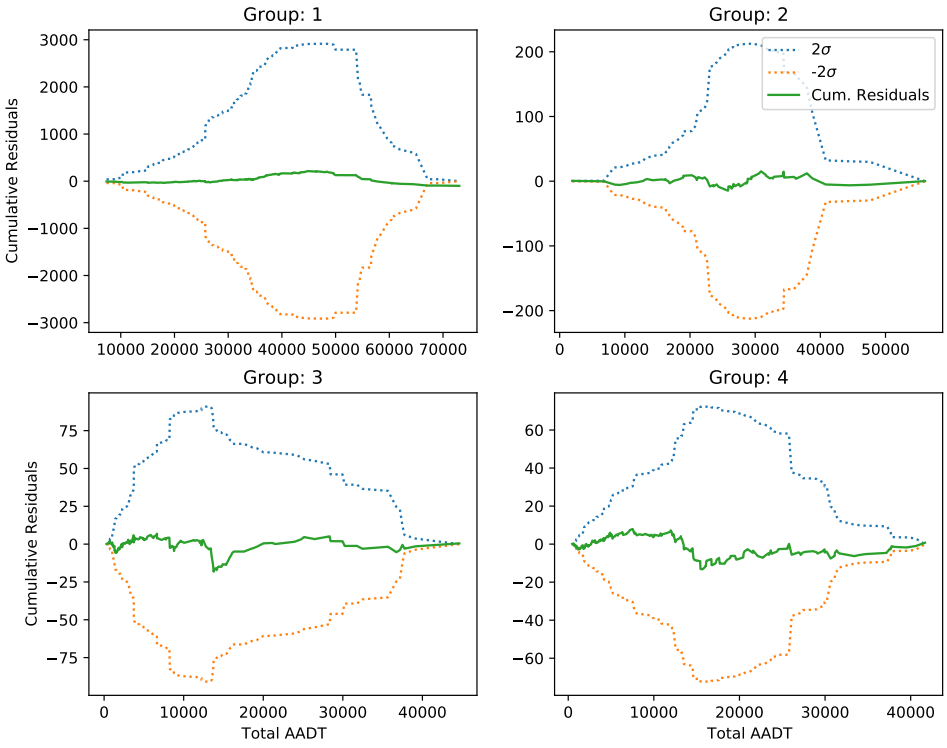
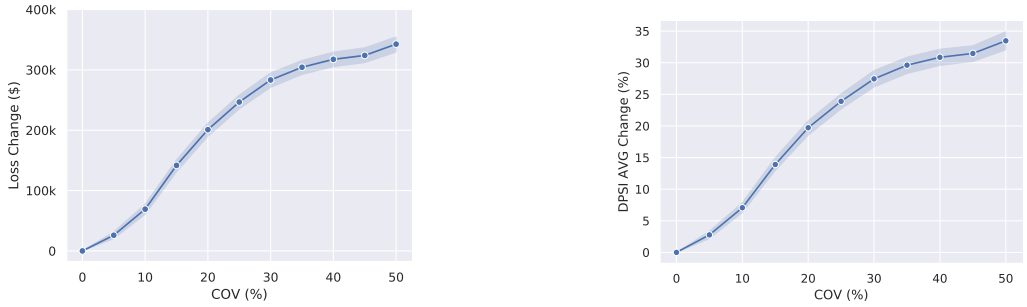


Figure 2.4: Cumulative Residual (CURE) plots for developed SPFs



(a) Loss change ($) versus COV of AADT error for London data set

(b) % Change of ΔPSI versus COV of AADT error for London data set

Figure 2.5: Loss change ($) and % Change of ΔPSI plots for London, ON data set

predicted crash frequencies. An SPF is deemed to be unbiased when the CURE plot falls within two standard deviations [48]. Based on the CURE plots in Figure 2.4, the residuals for all SPFs are within the two standard deviations (σ) indicating that SPFs are well-fit:

$$\sigma = \left[\sigma_i^2 \times \left(1 - \frac{\sigma_i^2}{\sigma_T^2}\right)\right]^{0.5} \tag{2.15}$$

where σ_i^2 is the cumulative sum of squared residuals until the element i, and σ_T^2 is the total cumulative sum of squared residuals.

In order to estimate the monetary loss due to inaccurate AADT data the steps described in Figure 2.1 are followed. It is assumed that the crash value (i.e. $\$(PSI)$) is $15,000 and network screening is performed to find the top 10% hotspots. For different values of ν in Figure 2.1, ΔPSI for top 10% hotspots is calculated using 250 different simulated data sets based on the SPFs and dispersion values (Table 2.1), and then converted into a monetary value using $\$(PSI)$ (Equation 2.9). In addition, the mean absolute percentage error (MAPE) for SPF coefficients (i.e. b_0 and b_1) and dispersion parameter given each ν value are computed to evaluate the impact of AADT error on model parameters.

The simulation results are presented in Figure 2.5a, 2.5b and 2.6. In Figure 2.5b, $\%\Delta PSI(\nu, n)$ from Equation 2.12 versus different values of ν values are presented. These results can be compated to those from the simulation sensitivity analysis by considering the mean crash rate of 4 and dispersion parameter of 0.5 from Table 2.1. The $\%\Delta PSI(\nu, n)$ values are close to the case with mean = 5.1 and dispersion = 0.5 in Figure 2.3.

Figure 2.5a shows the average monetary loss change ($Loss_\nu - Loss_{0\%}$) and its 95% confidence interval associated with each COV value in comparing with the base line (i.e.

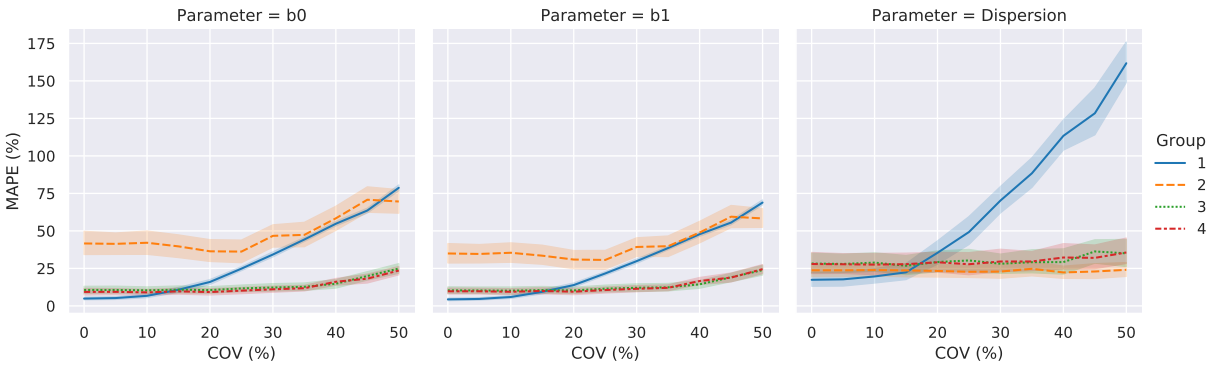


Figure 2.6: Mean Absolute Percentage Error (MAPE) for SPF parameter estimations

$\nu = 0$) which can be used for decision making regarding improving the accuracy of AADT values. Note that the expected benefit is non-linear, and consequently the expected benefit is a function of both the starting AADT accuracy (i.e. COV) and the improvement in accuracy (i.e. change in COV). For example, improving AADT accuracy by decreasing the COV from 30% to 10% is estimated to provide a benefit of approximately $200,000 where as improving AADT accuracy by decreasing COV of 50% to 30% provides about $60,000. In practice, the jurisdiction could compare the estimated expected benefit with the cost of achieving a given improvement in AADT accuracy and make an informed decision.

Finally, Figure 2.6 shows the mean absolute percentage error of model parameter estimations for each group and its 95% confidence interval associated with each COV value. As expected, estimation error tends to increase when AADT error is increased. Also the rate of increase is higher for larger sample means (e.g. G1 and G2) - a result that is consistent with the the results in Figure 2.3.

2.6 Conclusions, Limitations, and Recommendations

AADT data is a key input for developing road safety performance functions used in road safety network screening. AADT is typically estimated by applying temporal factors to expand short-term counts. The error in estimated AADT data have direct impact on safety performance measures. In this chapter, a simulation-based method is proposed that can quantify the monetary benefit of improving AADT accuracy. The proposed method is used over different simulated crash data conditions to study the sensitivity of network screening outcomes to AADT error. According to the simulation results, the sensitivity

of NS outcomes is highly dependent on crash data characteristics such as dispersion and sample mean, with crash data sets with larger sample mean and dispersion having higher sensitivity to AADT error. Additionally, the proposed method was used in a real example to quantify the cost of inaccuracies in AADT on the NS outcomes that can assist jurisdictions in understanding the benefits of investing in methods to improve AADT accuracy. The proposed method can be used with any safety performance function structure.

One of the limitations of the proposed method is the lack of knowledge regarding the nature of AADT errors. In this study, the COV of AADT errors is assumed to be constant for all sites and that the errors are Normally distributed. However, there is a need for future research to establish the characteristics of the AADT estimation errors to confirm that these assumptions are valid. It's important to note that if future research demonstrates that the COV of AADT estimation errors varies as a function of the AADT and/or follows something other than the Normal distribution, then the proposed simulation method can still be used by modifying Equation 2.6.

Furthermore, in this chapter, the proposed method is used for road segments. For application to intersections, in which the AADT of major and minor roads are both inputs for model development, there may be error correlation that should be considered.

Lastly, the proposed method monetizes the cost of inaccurate AADT values on the basis of the average cost of crashes (by type) and the change in PSI associated with more accurate AADT values. PSI quantifies the potential for safety improvements at a site in terms of reduction in crashes per year by comparing the EB crash frequency estimate for the site to the expected crash frequency predicted for sites with the same characteristics. However, achieving this reduction in crashes requires the implementation of safety improvement measures which vary in cost and effectiveness depending on the site characteristics, the nature of the crashes, countermeasure, etc. The proposed method for quantifying the value of improved AADT does not account for these variations in countermeasure costs and effectiveness.

Chapter 3

A Benefit-Cost Based Method to Determine When Safety Performance Functions (SPFs) Should be Redeveloped for Use in Intersection Network Screening

This chapter is based on the following journal article:

Zarei, M., Hellinga, B., & Izadpanah, P. (2022). A Benefit-Cost Based Method to Determine When Safety Performance Functions (SPFs) Should be Redeveloped for Use in Intersection Network Screening. *Transportation Research Record*, 2676(11), 239–249.

In this journal paper I was the first author and was responsible for the writing of the article. The paper was edited by Dr. Hellinga and Dr. Izadpanah. I also developed the crash predictive models and cost-benefit framework and the Python code.

3.1 Introduction and Background

In the previous chapter, a method is proposed to determine the value of improving AADT data accuracy which is a critical data for developing SPFs. In this chapter, we focus on one of the challenges in developing SPFs: deciding between redevelopment and re-calibration. A cost-benefit analysis method has been proposed that can be used to objectively determine whether a set of regionally developed SPFs used for road safety network screening should be redeveloped or recalibrated. SPFs that can be developed using both statistical models and deep learning models [15, 16] are an integral part of the Road Safety Management Process (RSMP) proposed by the Highway Safety Manual (HSM) [3], a systematic approach by which transportation authorities can identify and prioritize traffic sites (e.g., intersections, roadway segments, and/or ramps) for more detailed safety evaluations. As mentioned in the previous chapter, this process is commonly called Network Screening (NS) and the identification and prioritization of roadway locations is also referred to as hotspot identification.

Though different NS methods have been proposed in the literature [4, 6, 3] a common approach is to determine the excess expected average crash frequency using SPFs with empirical Bayesian adjustments [7, 4]. Regardless of the specific NS method used, in order for SPFs to be applied within the NS process, the functional form, the set of independent variables, and the coefficients of the SPFs must all be determined through one of the following two methods [9]:

1. Development of SPFs specific for the subject jurisdiction, or
2. Calibration of existing SPFs.

In the first method, *SPF Development*, the full statistical model development process is undertaken. A database for the subject jurisdiction is compiled consisting of crash frequency data, traffic volume and geometry data for each site in the network. Statistical model development is undertaken to determine the most appropriate structure of the SPF model, the statistically significant independent variables, and the model coefficients. The literature suggests that developing jurisdiction-specific SPFs for the purpose of network screening takes 40 to 100 staff-hours [9] but this will vary depending on the number of SPFs to be developed. Furthermore, this model development process requires specialized expertise in statistical modeling. These cost and expertise requirements often deter transportation authorities from undertaking SPF development on a regular basis.

In the second method, *SPF Calibration*, a pre-existing SPF is updated through a simple calibration process based on the database of the subject jurisdiction. There are several methods for calibrating crash models (i.e. SPFs) proposed in the literature, some of which are presented in Table 3.1. The HSM method is the most straight forward method and is more common in practice. The pre-existing SPF can be either an SPF that had been developed for the subject jurisdiction some time ago, or it can be a borrowed SPF that was originally developed for some other jurisdiction. SPF Calibration requires less time and resources [49] but conceptually could provide less reliable results than the SPF Development method [50].

Table 3.1: Calibration methods

Reference	**Method**	**Note**
[3]	$N_{pred}^{cal} = N_{Pred} \times \frac{N_{obs}^{tot}}{N_{pred}^{tot}}$	HSM method
[51]	$N_{pred}^{cal} = N_{Pred} \times \frac{N_{obs}^{tot}}{N_{pred}^{tot}}$	HSM method but estimate dispersion parameter based on local samples
[52]	$N_{pred}^{cal} = A \times N_{Pred}^{B}$	Estimate regression parameters, A and B, based on local sample
[53]	$N_{pred}^{cal} = \frac{\sum_{i=1}^{K} N_{Obs}^{i}}{K}$	Prediction for each site is the average of its K nearest sites based on their predictions values

[1] N_{Pred} = predicted crash frequency, N_{pred}^{cal} = calibrated prediction, N_{obs}^{tot} = total observed crash count

Jurisdictions typically conduct NS on an annual basis for the purposes of planning and implementing safety interventions; however, there exists very little prior work in the literature examining methods by which jurisdictions can determine when SPF development is justified. Shirazi et al. [54] proposed a procedure to determine when SPFs should be re-calibrated. They introduced a C-proxy parameter that is the ratio of the total number of observed crashes to a rough estimate of the total number of predicted crashes. If the relative difference between the C-proxy of the current year and the reference-year (i.e. the latest year that the SPFs were recalibrated) is greater than a certain threshold (e.g. 10%), it is recommended to recalibrate the SPFs.

While the above noted approach is simple to use, it poses three challenges:

1. It is not clear what the threshold should be or how to establish the threshold,

2. Even if a threshold is established, it is not clear that the threshold is transferable to other jurisdictions or even if the same threshold should be used for different site types (ie. intersections vs road segments) within the same jurisdiction, and

3. The method attempts to inform jurisdictions when they should recalibrate SPFs versus continuing to use their existing SPFs. The method does not provide any insight into when to redevelop SPFs.

Considering these limitations, this chapter proposes a method that enables practitioners to decide whether or not there is a need to undertake full SPF development.

This chapter is structured as follows. The NS method used in the study is described in the next section. The empirical data used in this study are described in Section 3.3, followed by the research methodology in Section 3.4. The results are described in Section 3.5. Finally, conclusions and recommendations are presented in Section 3.7.

3.2 Network screening based on Potential for Safety Improvement (PSI)

There are several methods in the literature for network screening such as crash frequency, crash rate, empirical Bayes (EB) estimate, and potential for safety improvement (PSI) [6]. It has been shown that the methods which apply the EB adjustment (i.e. PSI and EB) perform better than methods that do not use the EB adjustment [6, 7, 44]. Performance has been evaluated on the basis of a variety of metrics including false identification test, site consistency test, method consistency test, total rank differences test, and Poisson mean differences test [5, 4]. The PSI method, which is a refinement of the EB approach, is used in this study because it estimates how much we can improve the safety of a hotspot and this can be interpreted as a monetary benefit of network screening. PSI is the difference between the EB estimate and the predicted crash frequency μ, as follows:

$$PSI = EB - \mu = w \times \mu + (1 - w) \times Obs - \mu \tag{3.1}$$

where Obs is the observed crash frequency and predicted crash frequency, μ, is derived from a SPF which usually has the following form:

$$\mu = e^{\beta_0 + \beta\ \mathbf{X}} \tag{3.2}$$

where X is a vector of features (e.g. exposure, lane width, shoulder), β is a vector of model coefficients for the corresponding input features, and β_0 is intercept. Though many different statistical models have been proposed, the model coefficients are commonly estimated by fitting a negative binomial model to the data, which can account for over-dispersion of crash data (i.e. the variance exceeds the mean of the crash data) [14, 55]. Consequently, w in Eq. 3.1 can be calculated as a function of mu and dispersion parameter (k) of the negative binomial model:

$$w = \frac{1}{1 + k\mu} \tag{3.3}$$

3.3 Study data

The data used in this study consists of crash records, traffic volumes in the form of average annual daily traffic (AADT) and geometric characteristics for intersections within the Regional Municipalities of Niagara (2010-2018), Peel (2008-2017) and Halton (2007-2016), and within the City of London (2008-2018), all of which are located in south western Ontario, Canada. There are four types of intersections in each data set including 4-legged signalized, 3-legged signalized, 4-legged unsignalized, and 3-legged unsignalized. The summary statistics of the data sets are presented in Table 3.2.

3.4 Methodology

The proposed methodology is based on a benefit-cost analysis and consists of the following steps:

1. Develop SPFs and PSI values for all periods for each data set.
2. Define and quantify the benefit associated with using the most recent SPFs.
3. Develop a model(s) to estimate the benefit using available data to be compared with the redevelopment cost.

Note that it is assumed that the most recent SPFs that are developed based on more recent data provides more reliable network screening results since it represent the most recent behavior of system. In the following each of the above three steps is explained in details and the corresponding results are presented.

Table 3.2: Descriptive statistics of the data

Measures	Statistics	Peel	Niagara	Halton	London
Sites	Number	746	593	427	865
Total Crashes	Max	90	35	57	79
	Mean	6.2	2.3	3.7	3.8
	STD	10.6	3.4	6.2	7.0
Major Street AADT	Max	110,707	63,102	76,830	61,487
	Mean	25,805	11,712	20,242	14,400
	Min	291	374	297	156
	STD	16,851	6,964	11954	11,222
Minor Street AADT	Max	55,742	27,518	45,113	42,481
	Mean	6,136	3,979	4,925	3,561
	Min	2	60	7	7
	STD	9,178	3,827	6,211	4,937

3.4.1 Develop SPFs and Estimate PSI values

Two time periods are defined. The *Base period* is the first P years and the *Development period* is the most recent P years of data associated with each of the data sets. The *Base year* is the first year of the base period and the *Current year* is the last year of the development period. Note that the recommended range for P in literature is 3 to 5 years [56, 3]. Based on these periods, we also define three sets of SPFs as follows:

- **Base SPFs (B-SPFs):** the SPFs that are developed on the basis of the first P years of data from each data set (i.e. for $P = 4$, the Base SPFs are developed using data from 2008-2011 for Peel, 2010-2013 for Niagara, 2007-2010 for Halton, and 2008-2011 for London)

- **Calibrated SPFs (C-SPFs):** the adjusted B-SPFs using a calibration factor (C) [3]:

$$N_{pred}^{cal} = N_{Pred} \times C \tag{3.4}$$

  where C is the ratio of the total observed number of crashes (N_{obs}^{tot}) by the total number of crashes predicted by applying the Base SPFs (N_{pred}^{tot}) to the most recent P year period:

$$C = \frac{N_{obs}^{tot}}{N_{pred}^{tot}} \tag{3.5}$$

- **Redeveloped SPFs (R-SPFs):** the SPFs that are developed on the basis of the most recent P years of data.

The NS process is performed twice on each data set for each year after the base period. The first uses the C-SPFs and the second uses the R-SPFs. For example, consider the application of the NS process to the Peel region data set for 2014 and $P = 4$ years. In the first approach (C-SPFs) SPF coefficients have been computed on the basis of the base period (2008-2011) data. N_{Pred}, the output from these SPFs, is then adjusted by the calibration factor C which is computed on the basis of 2010-2014 data. In the second approach (R-SPFs), SPF coefficients have been computed on the basis of the most recent 4-years of data (i.e. 2010-2014). The NS process uses EB to combine the predictions from the SPFs with observed data. For both SPF approaches, the EB adjustment is made using the most recent P years of data (2011-2014 in this example) and then the PSI measure is computed for each intersection. Intersections are then ranked in descending order of PSI.

3.4.2 Define and quantify the benefit of R-SPF

Assuming that the re-developed SPFs (R-SPFs) provides more reliable results, the benefit of using R-SPFs versus calibrated SPFs (C-SPFs) can be quantified in terms of the net impact of changes to the sites identified as part of the top $n\%$ hotspots in terms of the total PSI:

$$PPB_{n\%} = \frac{\sum_R PSI_R - \sum_C PSI_R}{\sum_R PSI_R} \times 100 \tag{3.6}$$

where

$PPB_{n\%}$: Percentage PSI benefit (%) considering top $n\%$ hotspots

$\sum_R PSI_R$: Sum of the PSI in EPDO unit from all top $n\%$ hotspots resulting from NS using R-SPFs

$\sum_C PSI_R$: Sum of PSI in EPDO unit from all top $n\%$ hotspots resulting from NS using C-SPFs

It is important to note that the EPDO is used to convert fatal and injury crashes into PDO crashes based on the equivalent societal costs of crash severity outcomes [3].

$PPB_{n\%}$ shows that how much more PSI we could achieve by using R-SPFs instead of C-SPFs. Since PSI is in units of crashes, $PPB_{n\%}$ can be converted into a monetary value using the societal cost of one PSI unit ($(PSI)) [46, 3, 47] as follows:

$$\text{Benefit}(\$) = PPB_{n\%} \times \left(\sum_{R} PSI_R \right) \times \$(\text{PSI}) \tag{3.7}$$

where $(PSI) can be estimated based on the counts and societal costs [46, 3, 47] of property damage only ($\$_{PDO}$), injury crashes ($\$_{Injury}$), and fatal crashes ($\$_{Fatal}$):

$$\$(\text{PSI}) = \frac{N_{PDO} \times \$_{PDO} + N_{Injury} \times \$_{Injury} + N_{Fatal} \times \$_{Fatal}}{N_{PDO} + N_{Injury} + N_{Fatal}} \tag{3.8}$$

The monetary benefit of using redeveloped SPFs (i.e. Benefit($)) can be compared to the cost of redeveloping the SPFs to determine whether or not redeveloping SPFs is justified. Note that in this approach the benefits and costs are specific for the local jurisdiction and reflect local values of social costs of crashes as well as costs to redevelop SPFs.

3.4.3 Develop a model(s) to estimate the R-SPF benefit

The previous section presented PPB, which can be used to quantify the benefit of adopting R-SPFs but calculating PPB requires that the jurisdiction carry out redevelopment of SPFs and thus, in its current form, this measure is not useful for decision making. Consequently, in this section, we attempt to find a method for estimating PPB from data that is readily available without having to redevelop SPFs.

When the underlying characteristics that influence the relationships captured by the SPFs are the same between the base period and the development period, then we expect the NS outcomes using both C-SPFs and R-SPFs will be the same and PPB will be close to zero. Conversely, if the underlying characteristics between two periods are quite different, then the C-SPFs will be unable to accurately depict the development period, resulting in a higher PPB. As a result, if we can figure out the "difference" between the two periods, we can estimate PPB without needing to use R-SPFs.

There are three types of information (number of crashes; traffic volumes; and time since SPFs were last redeveloped) that are readily available and can be used to describe the difference between two periods, as well as have a potential correlation with PPB. For each type of information, we define a trend measure that captures the absolute relative change between the base period and the current period

$$\%Crash = \frac{|Crash_B - Crash_R|}{Crash_R} \tag{3.9}$$

$$\%AADT = \frac{|AADT_B - AADT_R|}{AADT_R} \tag{3.10}$$

$$\%Time = \frac{N_{year}}{P} \tag{3.11}$$

where subscript B represents the base period, R represents the redevelopment (current) period, $Crash$ is the total number of crashes in the network, $AADT$ is average of total AADT (or other exposure measure), N_{year} is the number of years between the current year and the base year, and P is the period length in year.

These measures are used to develop simple linear models that can be used for estimating PPB:

- $PPB_{Crash} \approx \beta_{Crash} \times$ %Crash
- $PPB_{AADT} \approx \beta_{AADT} \times$ %AADT
- $PPB_{Time} \approx \beta_{Time} \times$ %Time

β_{Crash}, β_{AADT}, and β_{Time} are the slope values for each PPB estimation model which are estimated in the following sections for the mentioned data sets.

3.5 Results

The methodology as described in the previous section was carried out using the four data sets. The following steps were performed for each data set, for three different period lengths (i.e. $P = 3, 4, 5$), and for each year after the base period:

1. R-SPFs and C-SPFs were developed. B-SPFs were developed for base period.
2. Using R-SPF, the PSI_R values were estimated and used for ranking hotspots
3. Using C-SPF, the PSI_C values were estimated and used for ranking hotspots
4. $PPB_{n\%}$ was computed for $n = 5\%, 10\%, 15\%, 20\%$.
5. $\%Crash$, $\%AADT$ and $\%Time$ were calculated.

Figure 3.1 shows the PPB values plotted against $\%Crash$, $\%AADT$, and $\%Time$ for each data data set. A linear least squares regression model (without intercept) for estimating PPB was fit for each trend measure and each data set.

All slope values are statistically significant at the 95% confidence level. The results indicate that the linear regressions based on $\%Crash$ are most accurate and those based on $\%AADT$ are least accurate. These relationships can be used for estimating PPB and decision making.

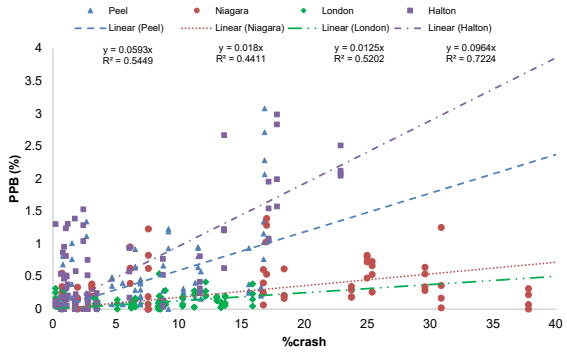
Each year following the base period, a decision should be made on whether or not to redevelop SPFs for each region. The true correct decision is made based on the true PPB value, which is calculated by comparing the NS results from R-SPFs and C-SPFs, and a PPB threshold:

$$PPB_{threshold} = \frac{\$(\text{R-SPF})}{\$(\text{PSI}) \times \sum_{n\%} \text{PSI}} \times 100 \tag{3.12}$$

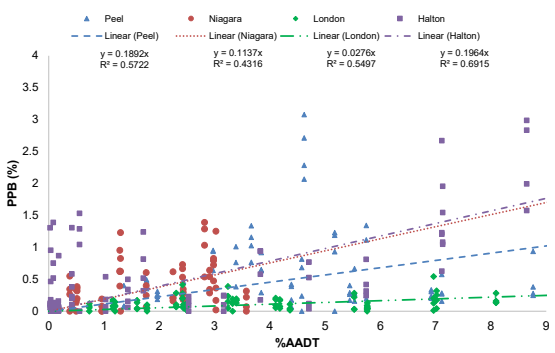
For instance, if the R-SPF cost ($(R-SPF)) is \$30,000, $(PSI) is \$15000, and sum of PSI values in top $n\%$ hotspots is 1000 EPDO units, then the PPB threshold will be 0.2%. The model decision is based on comparing the PPB threshold and the PPB estimated using one of the following models:

- Crash model: $PPB_{Crash} \approx \%Crash \times \beta_{Crash}$
- AADT model: $PPB_{AADT} \approx \%AADT \times \beta_{AADT}$
- Time model: $PPB_{Time} \approx \%Time \times \beta_{Time}$
- CAT model: $PPB_{CAT} \approx AVG(PPB_{Crash} + PPB_{AADT} + PPB_{Time})$

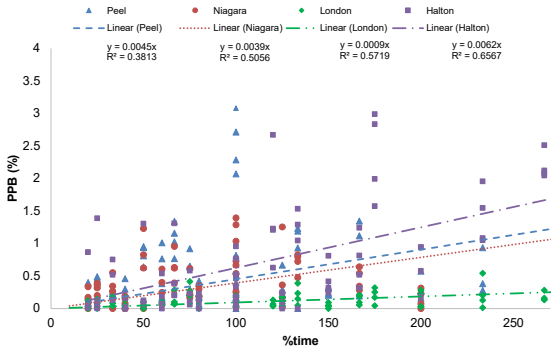
The CAT model is proposed to consider all three trend measures together and compare its performance with other individual models. In order to evaluate the performance of these models, three benchmark models are used as follows:



(a) %Crash



(b) %AADT



(c) %Time

Figure 3.1: PPB versus %Crash, %AADT, %Time

- R model: Redevelop SPFs every year
- P model: Redevelop after P years (i.e. period length)
- C model: Never redevelop SPFs (always use C-SPFs)

Finally, the performance of each model is quantified using the following accuracy metric over each data set over a range of PPB thresholds:

$$Accuracy = \frac{Correct\ Decisions}{Total\ Number\ of\ Decisions} \times 100 \tag{3.13}$$

where *Total Number of Decisions* = the number of years in the data set in which we would like to make a decision about SPF redevelopment, and *Correct Decisions* = number of years in which the model suggestion was correct according to the true PPB value and the specified PPB threshold.

The PPB threshold is a function of the redevelopment cost, the societal cost of a crash, and the total PSI, all of which vary by jurisdiction. Table 3.3 indicates the PPB threshold value for various combinations of typical values of redevelopment cost (expressed as a multiple of the societal cost of a crash) and the total PSI expected from the network screening process (which is a function of the size of the network and the percent of sites considered for the hotspots of interest). The range extends from 0.02% to 1.00%. The threshold value is very small when the cost of redevelopment is small relative to the cost of a crash and when the total PSI is large. Similarly, the threshold value is large when the cost of redevelopment relative to the cost of a crash is large and when the total PSI is small. When the threshold is small, we want to redevelop the SPFs and when the threshold is large, it is sufficient to use the calibrated SPFs.

Table 3.4 shows the average, standard deviation, and coefficient of variation (COV) of total PSI for each data set over the range of available years. As indicated for all four data sets used in this study, the total PSI in top hotspots for each year exhibited substantial temporal stability suggesting that jurisdictions are able to make a sufficiently accurate estimate of total PSI from the last time they carried out NS using redeveloped SPFs.

The accuracy of the models were assessed over the range of threshold values from 0.02% to 1.0%. The accuracy of each model for each data set is presented in Figure 3.2. These results are based on using the PPB models shown in Figure 3.1. The relationships between model accuracy and PPB threshold appear to follow a similar pattern in all four data sets and the "R" and "C" benchmark models behave as expected. For example, the "R" model

in all data sets starts at roughly 100 percent accuracy and decreases as the PPB threshold increases. Conversely, the "C" model starts at approximately 0% accuracy and increases in accuracy as the PPB threshold increases.

3.6 Discussion

Based on the results presented in the previous section, all four proposed models appear to be more accurate than the bench mark methods over the majority of the PPB threshold range and all exhibit a "U"shape as expected. Their accuracy is similar to the "R" model, which is the best model for low PPB threshold values, and then is similar to the "C" model, which is the best model for high PPB threshold values.

Although the graph for London follows the same trend, the intersection of the "C" model and "R" model (where both models have the same accuracy of 50%) is closer to zero than for the other three data sets, indicating that for this data set, there is a much greater range for which using calibrated SPFs is the best choice. This result indicates that for this data set there were no significant changes to the underlying relationships between crashes and site characteristics between the base period and the following years and consequently the SPFs are more temporally stable than for the other three jurisdictions. This is additional evidence that the determination of when to redevelop SPFs must be specific for the jurisdiction rather than employing the same fixed update interval for all locations. It can be seen that all four proposed models are able to capture the unique characteristics of each jurisdiction.

Table 3.3: PPB threshold (%) based on Total PSI and cost ratio ($\frac{\$(\text{R-SPF})}{\$(\text{PSI})}$)

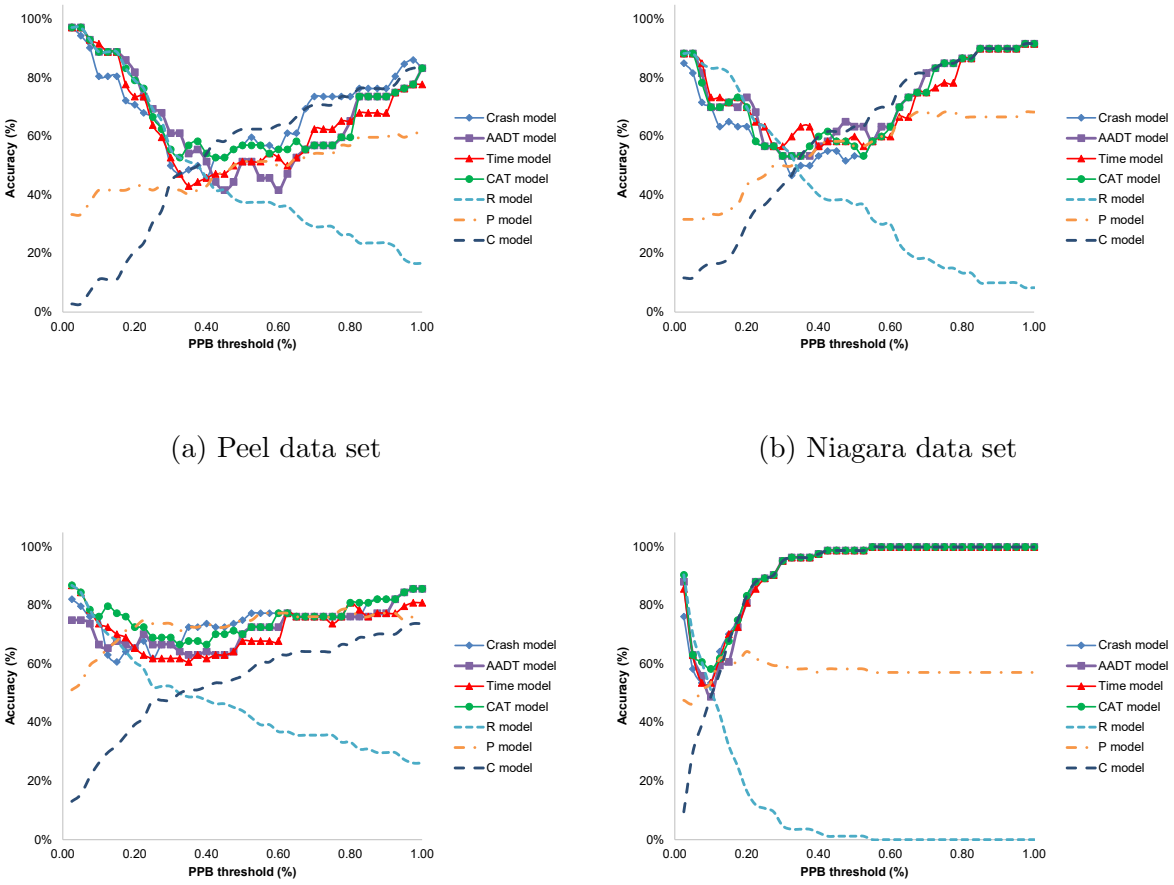
		Total PSI in top hotspots				
		500	**1000**	**1500**	**2000**	**5000**
$\frac{\$(\text{R-SPF})}{\$(\text{PSI})}$	**1**	0.20%	0.10%	0.07%	0.05%	0.02%
	2	0.40%	0.20%	0.13%	0.10%	0.04%
	3	0.60%	0.30%	0.20%	0.15%	0.06%
	4	0.80%	0.40%	0.27%	0.20%	0.08%
	5	1.00%	0.50%	0.33%	0.25%	0.10%

To compare the performance of the proposed models with the benchmarks, the accuracy of each model is averaged across the PPB threshold range of 0.02-1.0% and the results

Table 3.4: Temporal variation of total PSI for top 10% hotspots

	Total PSI in top 10% hotspots			
	Peel	Niagara	Halton	London
Mean	630	195	268	423
SD	48	26	9	31
COV	7.6%	13.3%	3.4%	7.3%

are shown in Table 3.5. The results reveal that, on average, all four proposed models outperform all three benchmarks. Based on average accuracy, the best model is the CAT model with 77.1 % average accuracy, which improves on the best benchmark, the "C" model with 63.9% average accuracy, by about 20.6%. Among the proposed models, the Time model has the lowest average accuracy of 75.3% and the Crash model has the second best average accuracy of 76.4%.



(a) Peel data set

(b) Niagara data set

(c) Halton data set

(d) London data set

Figure 3.2: Accuracy of models over four data sets

Table 3.5: Average accuracy of proposed and benchmark models with optimized slopes for each data set

	Proposed Models				Benchmark Models		
Data set	**Crash model**	**AADT model**	**Time model**	**CAT model**	**R model**	**P model**	**C model**
Peel	**68.8**%	66.0%	65.0%	67.7%	47.5%	49.1%	52.5%
Niagara	70.0%	**73.0**%	72.5%	72.0%	39.6%	55.6%	60.4%
Halton	74.8%	72.2%	71.7%	**76.0**%	46.3%	73.0%	53.8%
London	92.0%	91.8%	92.1%	**92.6**%	11.1%	57.4%	88.9%
AVG	76.4%	75.8%	75.3%	**77.1**%	36.1%	58.8%	63.9%

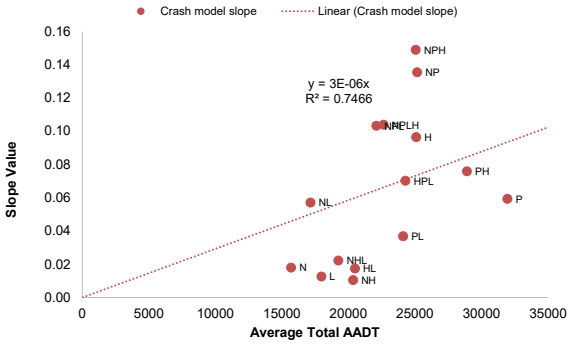
The previous evaluation of model accuracy assumes that the jurisdiction specific optimal slope value for each model is known (i.e. calculated once using the same process described in previous section). Though jurisdictions can calibrate these values, this imposes additional effort, so there is motivation to provide an alternative approach. Our investigation showed that there is a relationship between the model slope values and the average of total AADT (AADT of major road + AADT of minor road) per site in each data set. In order to confirm this, all combinations of the four data sets were created for the overlapped years to prepare a wider range of data sets:

- NP data set: based on Niagara and Peel data sets for 2010-2017
- NL data set: based on Niagara and London data sets for 2010-2018
- PL data set: based on Peel and London data sets for 2008-2017
- HL data set: based on Halton and London data sets for 2008-2016
- PH data set: based on Peel and Halton data sets for 2008-2016
- NH data set: based on Peel and Halton data sets for 2010-2016
- NPL data set: based on Niagara, Peel and London data sets for 2010-2017
- NPH data set: based on Niagara, Peel and Halton data sets for 2010-2016
- NHL data set: based on Niagara, Halton and London data sets for 2010-2016
- HPL data set: based on Halton, Peel and London data sets for 2008-2016
- NPLH data set: based on Niagara, Peel, London, and Halton data sets for 2010-2016

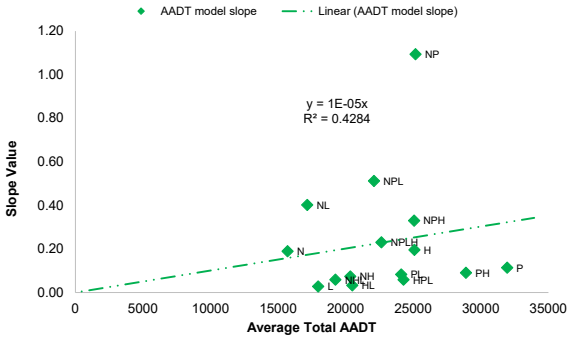
For each of the above data sets the process of fitting the linear models to PPB values as a function of %Crash, %AADT, and %Time as described in the previous section was followed. The resulting slope values for each model are plotted against average total AADT per site for each data set and presented in Figure 3.3. Based on these results, the slope values for the proposed Crash, AADT, and Time models can be approximated using average total AADT per site (slopes are significant at the 95% confidence level). This means that for a similar change in the measures (i.e. %Crash, %Time, %AADT) in jurisdictions with higher average AADT values, more benefit is expected than in jurisdictions with lower average AADT values. The results show that this estimation is more accurate for the

Crash and Time models with uncentered-R^2 score of 0.75 and 0.67 respectively than for the AADT model with uncentered-R^2 score of 0.43.

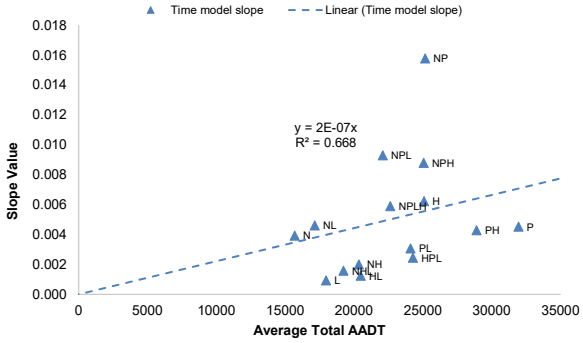
The slope estimation models from Figure 3.3 were used to estimate slopes for each of the four proposed PPB estimation models and then the average accuracy was determined over the PPB threshold range of 0.02-1.0% for Peel, Niagara, Halton, and London. The results are presented in Table 3.6. As expected, the models are less accurate when using estimated slopes rather than optimized slopes and this is more severe for London in which using optimized slopes can improve the accuracy of the proposed decision making models by approximately 50%. For the other three data sets, the improvement from using optimized slopes versus estimated slopes is less than 10%.



(a) Crash model



(b) AADT model



(c) Time model

Figure 3.3: Model slope values versus average total AADT

Table 3.6: Average accuracy of proposed and benchmark models with estimated slopes for each data set

	Proposed Models				Benchmark Models		
Data set	**Crash model**	**AADT model**	**Time model**	**CAT model**	**R model**	**P model**	**C model**
Peel	60.4%	56.8%	**63.9%**	62.5%	47.5%	49.1%	52.5%
Niagara	57.3%	**73.5%**	72.7%	68.5%	39.6%	55.6%	60.4%
Halton	**75.9%**	70.2%	73.4%	75.8%	46.3%	73.0%	53.8%
London	66.8%	50.1%	68.3%	61.3%	11.1%	57.4%	**88.9%**
AVG	65.1%	62.6%	**69.6%**	67.0%	36.1%	58.8%	63.9%

One of the limitations of the analysis described in this chapter is that it is based on data spanning at most 11 years and consequently the range of variation of both average AADT and total number of crashes is limited. Thus, the linear relationships calibrated between PPB and trend measures (i.e. $\%Time$, $\%AADT$ and $\%Crash$) may not be linear over a larger range of values of trend measures. Future studies can examine this by using data sets with a longer time period. Also, the proposed method is only tested and verified on intersections as the traffic sites. There are other types of sites in a traffic network (e.g. road segments) that need to be considered as well. Investigating the applicability of the proposed method for other types of sites is another task for future studies.

3.7 Summary of Methodology

In this chapter, a cost-benefit analysis method has been proposed that can be used to determine whether or not a set of regionally developed SPFs used for road safety network screening should be redeveloped. The key assumption is that the most accurate network screening results are obtained by using SPFs that have been developed using the more recently available crash and network data. Based on this assumption, a PSI percentage benefit (PPB) has been defined to capture the benefit of using redeveloped SPFs (R-SPFs) instead of using calibrated SPFs (C-SPFs) and found a set of models to estimate PPB from readily available data. The proposed methodology requires the following information:

1. $\frac{\$(\text{R-SPFs})}{\$(\text{PSI})}$: cost ratio where \$(R-SPF) is the cost of redeveloping SPFs and \$(PSI) is the monetary value of a PSI unit. The latter can be estimated using Equation 3.8 which requires the number of PDO, injury and fatal crashes as well as their societal costs. A reasonable estimate is typically between \$10,000 to \$30,000.

2. $\sum_{n\%} PSI$: total PSI for top $n\%$ hotspots which can be obtained from the previous NS results. n% is the percentage of top hotspots of interest. Total PSI will vary with the size of the network (i.e. number of sites) and n%.

3. $\overline{AADT}_{total}$: Average total entering AADT per intersection (if the optimized slopes for the PPB estimation models are not available)

4. %Crash: the absolute percentage difference of total crash counts between the base period and development period

5. %AADT: the absolute percentage difference of average total AADT per site between the base period and development period

6. %Time: the difference between current year and base year divided by period length (i.e. P) multiplied by 100

Based on the above data and the following steps, a decision can be made:

1. Calculate the PPB threshold (%) using Eq. 3.12.
2. Estimate the slope values for each PPB estimation model if the optimized values are not available:
   - $\beta_{Crash} \approx 2.92 \times 10^{-6} \times \overline{AADT}_{total}$
   - $\beta_{AADT} \approx 1.01 \times 10^{-5} \times \overline{AADT}_{total}$
   - $\beta_{Time} \approx 2.21 \times 10^{-7} \times \overline{AADT}_{total}$
3. Estimate the expected PPB from one of the following models:
   - $PPB_{Crash} \approx \beta_{Crash} \times$ %Crash
   - $PPB_{AADT} \approx \beta_{AADT} \times$ %AADT
   - $PPB_{Time} \approx \beta_{Time} \times$ %Time
   - $PPB_{CAT} \approx AVG(PPB_{Crash}, PPB_{AADT}, PPB_{Time})$
4. If the estimated PPB is greater than the threshold then redevelopment is recommended (i.e. the benefits of redevelopment is greater than its costs).

3.8 Conclusions and Recommendations

This study presents a methodology for jurisdictions to determine when to redevelop SPFs economically, rather than relying on recalibrated older versions. The approach uses three trend measures and a benefit metric based on societal costs of crashes to estimate benefits before redevelopment. The method, tested on real datasets, outperforms existing benchmarks by 13% in decision-making accuracy. Despite the fact that the suggested method performs well over the data sets used in this study, it is recommended to further validate the models for a larger set of jurisdictions and longer time periods. Furthermore, the approach was developed and tested exclusively on intersections. Future research can test and refine the method for other types of sites such as road segments.

Chapter 4

CGAN-EB: A Non-parametric Empirical Bayes Method for Crash Frequency Modeling Using Conditional Generative Adversarial Networks as Safety Performance Functions: Preliminary Performance Analysis

This chapter is based on the following journal article:

Zarei, M., Hellinga, B., & Izadpanah, P. (2022). CGAN-EB: A Non-parametric Empirical Bayes Method for Crash Frequency Modeling Using Conditional Generative Adversarial Networks as Safety Performance Functions. *International Journal of Transportation Science and Technology*, ISSN 2046-0430.

In this journal paper I was the first author and was responsible for the writing of the article. The paper was edited by Dr. Hellinga and Dr. Izadpanah. I also developed the CGAN-EB and GLM-based models models using Python.

4.1 Background

In earlier chapters, benefit-cost analysis methods have been developed to estimate the value of improving AADT accuracy as well as the value of redeveloping SPFs in terms of NS results (Figure 1.1. In this chapter, we focus on addressing functional form limitation of traditional safety performance functions through using more flexible modeling techniques.

Developing a reliable crash predictive model (or SPF), is a complex process owing to some inherent characteristics of crash data such as excess zero values (i.e. low sample mean), temporal/spatial correlation, multicollinearity (i.e. the high degree of correlation between two or more independent variables), and over-dispersion (i.e. variance is greater than mean) to name a few [14]. Several statistical modeling approaches have been used to deal with these issues that are mainly based on Poisson and Negative Binomial (NB) models or their variants in the form of mixture models or zero-inflated models [57, 58, 59, 60]. Practically, it is recommended that the Poisson regression model is estimated as an initial model, and if over-dispersion is found, then both negative binomial and zero-inflated count models could be considered [61].

There are several challenges to developing SPFs using parametric modeling. Selecting the best parametric model and establishing a functional form that best fits the given crash data can be technically challenging and time/effort intensive and the choice can have a significant influence on the outcome of network screening [62, 63]. One alternative which avoids these challenges is to use non-parametric or semi-non-parametric modeling approaches. For instance in [43], it was shown that a semi-non-parametric Poisson (SNP) model performs better than NB model when calibrated to simulated data sets with various distribution of error terms. The NB model was found to substantially overestimate the effect of lane width on crash frequency reduction relative to the SNP model due to the SNP model's more robust estimation of unobserved heterogeneity.

Another option to deal with the aforementioned challenges of parametric models is using fully non-parametric or data-driven models such as deep neural networks (DNN). Several studies have shown that such models have better fitting and predictive performance than parametric ones in the context of crash data modeling [64, 65, 66, 67, 68]. These computational models are able to extract inherent features in the data minimizing the efforts required for feature selecting and feature engineering that is essential for parametric models. They have also demonstrated superior performance in dealing with multi-colinearity, which is the non-independence of predictor variables [69]. Moreover, significant progress has been made to address the criticism that DNN-based models are black-boxes [70, 71]. In one of the recent works, a novel method has been proposed that can be used to describe

the inner working of deep learning models [72]. The method includes visualization and feature importance criteria over the input and hidden layers of the network that can be used for data/feature importance analysis.

Non-parametric approaches, such as DNN models, have been applied to different traffic safety problems to overcome these limitations. Some of the recent applications of deep learning models in crash data analysis include a crash count model with an embedded multivariate negative binomial model [64], developing a global safety performance function [67], real time crash predictions [73, 74], pedestrian near-accident detection [75], crash severity prediction [76, 77], and crash data augmentation [78, 79]. However, using a DNN based model for the purpose of hotspot identification and integrating it with the EB method is quite rare. Hence, there is a need for an EB estimation approach that takes advantage of DNN models and does not have the limitations of parametric approaches. In this chapter, using a powerful deep learning technique called Conditional Generative Adversarial Network (CGAN), a DNN-based model is combined with the EB method and compared with the traditional NB-EB approach in terms of model fitting, predictive performance and network screening results. In the next section the concept of CGAN is briefly described.

4.2 Conditional Generative Adversarial Network (CGAN)

GANs are a relatively recently developed type of deep generative models from the unsupervised machine learning field that can implicitly model any kind of data distribution and generate new samples and has achieved tremendous success in many fields (e.g., image/video synthesis/manipulation, natural language processing, classification) in recent years [80, 81, 82]. GANs consist of simultaneously training two deep neural networks, a generator which produces synthetic samples that mimic the characteristics (i.e. distribution) of the real (observed) data, and a discriminator which tries to distinguish between the synthetic samples coming from the generator and the real samples from the original data set. This competition between the generator and discriminator has been formalized as a min-max optimization problem and shown that at Nash equilibrium of the contest between the generator and the discriminator, the generator can capture the distribution of the observed data (for more information refer to [80]).

In a conditional-GAN [83], referred to as CGAN, both the generator and the discriminator are conditioned on some data which could be a class label or a feature vector if we wish to use it for regression purposes. For the regression, the training steps are presented in Figure 4.1.

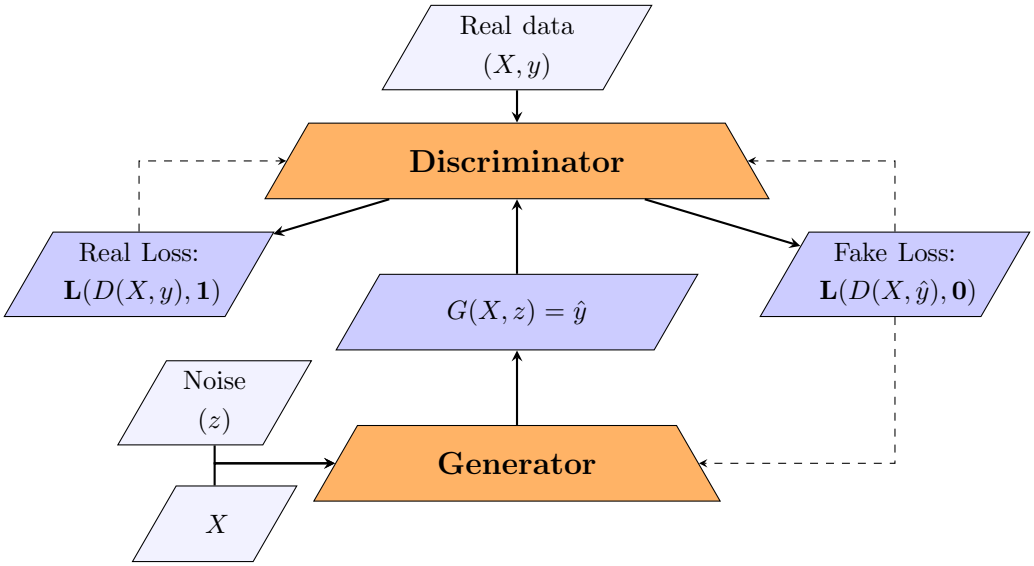


Figure 4.1: CGAN training structure (X is feature vector, y is the dependent variable, z is a noise value from a normal distribution $N(0, 1)$)

The CGAN training begins by assigning two sets of random weights to both the discriminator and generator neural networks. Next, real loss value which shows the ability of the discriminator to recognize the real instances (i.e. y) is calculated based on $D(X, y)$, unit vector (i.e. $\mathbf{1}$) and a loss function (e.g. binary cross entropy). Note that $D(X, y) \in [0, 1]$ is the output of the discriminator using (X, y) as input and $G(X, z) = \hat{y}$ is the output of generator using (X, z) as input. Fake loss value which shows the ability of the discriminator to recognize fake instances (i.e. $\hat{y}$) is calculated based on $D(X, \hat{y})$, zero vector (i.e. $\mathbf{0}$) and a loss function (e.g. binary cross entropy). Here $D(X, \hat{y})$ is the output of the discriminator using $(X, \hat{y})$ as input. Then, the weights of the discriminator are updated based on the objective to minimize total loss (i.e. real loss + fake loss) and the weights of the generator will be updated based on its objective to maximize fake loss. This cycle continues until the stop condition (e.g. maximum number of epochs) is met. In an ideal training condition, both fake loss and real loss converges to 0.5 which indicates that it is impossible to distinguish between input real data and synthetic data because they are samples of the same distribution [83]. In the context of crash prediction models, X, y and $\hat{y}$ represent site characteristics, crash count, and the generated crash count by generator respectively.

The generator mimics the underlying distribution of real data so it can generate m samples $(y_{ij}; j = 1, m)$ for site i and this set of samples is conditioned on the feature vector X_i. The mean of the samples can then be interpreted as the prediction of the model [84].

This process can be formalized as a min-max function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p(x)}\left[\log(D(x))\right] + \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D(G(z))\right)\right] \tag{4.1}$$

where $p(x)$ is the training data distribution, $p(z)$ is the prior distribution of the generative network, and z is a noise vector sampled from the model distribution $p(z)$ such as the Gaussian or uniform distribution.

GAN and its variants have been rarely used in crash data analysis. The review of the literature indicates very little application of GAN and its variants to crash data analysis problems. It has been observed that GAN is used to generate traffic data related to crashes in order to overcome data imbalance in real-time crash prediction [79]. In other transportation areas, GAN have been recently used for network traffic prediction [85] and traffic flow data imputations [86]. In this study, we propose an EB framework using CGAN, named CGAN-EB, as an alternative to NB models for SPFs in the network screening process.

4.3 CGAN-EB framework

As proposed by Hauer [19], given y as the observed number of crashes and k as the expected crash count, the EB estimator of k can be calculated as follows:

$$E(k|y) = w \times E(k) + (1-w) \times y \approx w \times E(y) + (1-w) \times y \tag{4.2}$$

where the weight w is a function of the mean and variance of k and is always a number between 0 and 1:

$$w = \frac{E(k)}{E(k) + VAR(k)} \tag{4.3}$$

In the above equations, if k is gamma distributed, then the resulting y follows an NB distribution (Eq. 4.4). As a result, the NB-EB estimate can be derived as follows:

$$f(y|\mu,\alpha) = \frac{\Gamma(y+1/\alpha)}{\Gamma(1/\alpha)\Gamma(y+1)}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)\left(\frac{1}{1+\alpha\mu}\right) \tag{4.4}$$

$$EB^{NB} = w \times \mu + (1-w) \times y \tag{4.5}$$

$$w = \frac{1}{1 + \alpha\mu} \tag{4.6}$$

where y is the observed number of crashes per year, α is the dispersion parameter, and μ is the number of crashes predicted by the NB model.

Note that the mean and the variance of y are $E[y] = \mu$ and $var(y) = \mu + \alpha\mu^2$ respectively. If $\alpha \to 0$, the crash variance equals the crash mean and the NB distribution converges to the Poisson distribution.

In order to derive the EB estimates using a CGAN model (EB^{CGAN}), we need $E(k)$ and $VAR(k)$ (see Eq. 4.3) which can be approximated using the samples (e.g. $m = 500$ samples to obtain stable estimates while being fast in calculations) taken from a trained CGAN model given the feature vector (X) of any given site:

$$E^{CGAN}(k) \approx \frac{\sum_{j=1}^{m} CGAN_j(X)}{m} \tag{4.7}$$

$$Var^{CGAN}(k) \approx \frac{\sum_{j=1}^{m} (E^{CGAN}(k) - CGAN_j(X))^2}{m - 1} \tag{4.8}$$

where $CGAN_j(X)$ is the j-th sample from total m samples obtained from the CGAN model when provided with X, the feature vector of a given site, as input. Also, the weight factor w in Eq. 4.2 when using a CGAN model can be defined as follows:

$$w^{CGAN} \approx \frac{E^{CGAN}(k)}{E^{CGAN}(k) + Var^{CGAN}(k)} \tag{4.9}$$

Based on the above equations, EB^{CGAN} can be formulated as:

$$EB^{CGAN} = w^{CGAN} \times E^{CGAN}(k) + (1 - w^{CGAN}) \times y \tag{4.10}$$

4.4 Evaluation Methods

In order to compare the performance of CGAN and NB, two approaches have been employed in this study: 1) using a real-world crash data, and 2) using simulation experiments.

4.4.1 Performance Evaluation using Real-world Crash data set

The first approach uses a real data set from the Highway Safety Information System (HSIS) to evaluate the proposed CGAN-EB method. Evaluation considers (i) model fit to the available crash frequency data; (ii) model performance for predicting crash frequency for another time period; and (iii) network screening performance (i.e. identifying hotspots). The fit performance of the models have been evaluated using three common criteria including MAE, MAPE and coefficient of determination (R^2 score). The predictive performance of the models are assessed using a test set (e.g. crash data of another period). The network screening performance is evaluated using four tests including the SCT, the MCT, and RDT that are presented in [87, 88], and the PDT which has been proposed in this study. All these tests are based on the assumption that, in the absence of significant changes, detected hotspots in time period i should remain hazardous in the subsequent time period $i+1$ [89].

The Site Consistency test (SCT) was designed to measure the ability of a network screening method to consistently identify a hazardous site (i.e. hotspot) in subsequent observational periods. The higher the SCT score, the better the network screening method is. The SCT score for roadway segments is calculated using the following equation [90, 87]:

$$SCT_v = \frac{\sum_{r=1}^{R} C_{r,v,i+1}}{\sum_{r=1}^{R} L_{r,v}} \tag{4.11}$$

where $C_{r,v,i+1}$ is the number of crashes at a site in the time period $i+1$ that is ranked r as identified by method v using the data from time period i, $L_{r,v}$ is the corresponding length of r^{th} ranked site (in miles or km), and R is the rank threshold that is used as cut-off in the hotspot identification.

The method consistency test (MCT) measures the number of hotspots detected in time period i that are then also detected in the subsequent time period $i+1$. The greater the MCT score, the more reliable and consistent the method is. The MCT score can be expressed as:

$$MCT_v = |\{x_{r=1}, x_{r=2}, ..., x_{r=R}\}_i \cap \{x_{r=1}, x_{r=2}, ..., x_{r=R}\}_{i+1}| \tag{4.12}$$

where $x_{r=1}, x_{r=2}, ..., x_{r=R}$ are the hotpots that are ranked $1; 2; ...; R$ respectively by the network screening method v.

In Rank Difference Test (RDT), the ranking of hotspots in two consecutive periods are compared. The method with a smaller RDT score is considered a superior method. This score can be calculated as follows:

$$RDT_v = \sum_{r=1}^{R} |r - R(x_{r,v,i+1})| \tag{4.13}$$

where $R(x_{r,v,i+1})$ is the rank of hotspot x in period $i+1$ which has been ranked r^{th} in period i by method v.

Finally, the Prediction Difference Test (PDT) which is a revised version of total performance test [91], quantifies the difference between the total number of crashes (EB estimate) for the top ranked hotspots in period i and the total number of crashes estimated for the top ranked hotposts in period $i+1$. The PDT score is computed as follows:

$$PDT_v = \sum_{r=1}^{R} |EB_{x_r,i,v} - EB_{x_r,i+1,v}| \tag{4.14}$$

where $EB_{x_r,i,k}$ is the EB estimate of the long-term mean of crashes for site x_r with rank of r in period i by method v, and $EB_{x_r,i+1,v}$ is the corresponding EB estimate for period $i+1$. The method with a smaller PDT score is considered a superior method.

4.4.2 Performance Evaluation using Simulation Experiments

This section describes a simulation experiment aiming to compare the performance of CGAN-EB versus NB-EB. In a simulation environment, it is possible to establish a priori sites that are hazardous and assess whether the competitive methods can correctly identify them [88]. For this experiment, five crash data sets are simulated which have the same independent variables (i.e X data) as the real-world crash data set but have different crash frequencies (i.e. y data) that are randomly sampled from a Poisson distribution with mean and variance of k which is computed based on a given SPF as follows [41, 42, 43]:

$$y_i \sim Poisson(k_i) \tag{4.15}$$

$$k_i = \mu_i \times \exp(\epsilon_i) \tag{4.16}$$

$$\mu_i = \text{SPF}(X_i) \tag{4.17}$$

$$\exp(\epsilon_i) \sim gamma(1, \alpha) \tag{4.18}$$

where α is the dispersion parameter. Note that ϵ represents the unobserved heterogeneity following the log-gamma distribution as assumed in the NB model. It is worth mentioning that these simulation settings are completely consistent with the NB model assumptions regarding the error term distribution and log-linear relationship between dependent and independent variables.

After simulating all data sets for each experiment, we have trained both NB-EB and CGAN-EB models using each training data set. In order to compare CGAN-EB versus NB-EB in terms of their performance as crash hotspot identification methods, two error measures, FI test and PMD test proposed in [87], and the mean absolute percentage error (MAPE) of EB estimates have been employed. These tests have been specifically used for comparing hotspot identification methods in a simulation environment where the truth is known. The FI test calculates the number of sites that are erroneously categorised as hotspots, and the PMD test is the mean absolute difference of the true Poisson means for true hotspots and the suggested hotspots by a method. In this study, a normalised version of these tests are proposed and used to compare across different number of top hotspots. The normalized FI, PMD and MAPE are computed as follows:

$$FI_m = \frac{|\{h_{r=1}, h_{r=2}, ..., h_{r=R}\} - \{x_{r=1}, x_{r=2}, ..., x_{r=R}\}_m|}{|\{h_{r=1}, h_{r=2}, ..., h_{r=R}\}|} \tag{4.19}$$

$$PMD_m = \frac{\sum k_h - \sum k_x}{\sum k_h} \tag{4.20}$$

$$MAPE = \sum_{i=1}^{N} \frac{|k_i - EB_i|}{k_i} \tag{4.21}$$

In Equation 4.19, $x_{r=1}, x_{r=2}, \ldots, x_{r=R}$ are the sites suggested as hotspots by the method m that are ranked $1; 2; \ldots; R$ respectively, and $h_{r=1}, h_{r=2}, \ldots, h_{r=R}$ are the true hotspots that are ranked based on the true Poisson means (i.e. k values). R is the rank threshold that is used as a cut-off in the hotspot identification. In Equation 4.20, $\sum k_h$ is the sum of true Poisson means for true hotspots and $\sum k_x$ is the sum of true Poisson means for the sites suggested as hotspots by method m. Both FI and PMD have a minimum value of zero (when the method discovers all true hotspots) and a maximum value of 1 (when none of the true hotspots are detected.

Table 4.1: Summary of Characteristics for Individual Road Segments in the WA Data (Numerical Variables)

Numerical Variable	**Minimum**	**Maximum**	**Mean (SD)**
Number of Crashes per year	0	60	0.9 (2.5)
Left shoulder width 1 [LSW1 (ft)]	0	22	2.1 (2.5)
Left shoulder width 2 [LSW2 (ft)]	0	20	2.2 (2.6)
Median width [MW (ft)]	2	750	40.1(41.6)
Right shoulder width 1 [RSW1 (ft)]	0	24	8.6 (3.6)
Right shoulder width 2 [RSW2 (ft)]	0	22	8.5(3.5)
Segment length [L (mi)]	0.01	2.02	0.1 (0.1)
AADT (F)	4328	178,149	46618 (27725)

Note: 1 and 2 indices after LSW and RSW are related to two opposite directions of the road. Mean AADT computed over all sites and over 6 years

4.5 Empirical crash data set description

The crash data set used in this study is from the Highway Safety Information System (HSIS) collected for divided 4-lane segments from urban freeways from 2012 to 2017 in Washington State, USA. The data set covers a total of 3085 individual road segments with length ranging from 0.01 to 2.02 miles (0.016 to 3.25 km). 2047 segments are located in rolling terrain type and the rest (1038) are located in level terrain. Table 4.1 provides a summary of statistics for seven numerical variables of the road segments in the Washington State (WA) data and the correlation of the 7 aggregated variables is displayed in Figure 4.2. The blue color represents positive correlations, whereas the red tint represents negative correlations. The darker the color, the greater the correlation. The results show that, as expected, LSW1, LSW2, RSW1 and RSW2 are highly correlated so only one of them appeared in the final model.

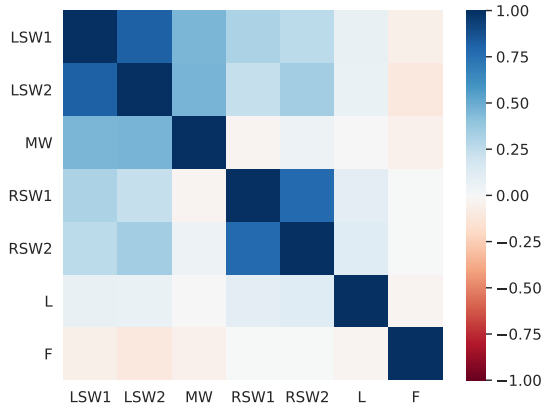


Figure 4.2: Variable correlation

4.6 Developing models

The data described in the previous sections were divided into two 3-year periods P1 (2012-2014) and P2 (2015-2017) in order to develop models (NB and CGAN) and compare their performances based on the described evaluation methods. The details of the model development process for the NB-EB method and the CGAN-EB method are explained in the following sections.

4.6.1 NB models

The NB modeling results for the WA data are provided in this section. A NB model requires the relation of crash frequency and the set of explanatory variables to be specified. We have used the following GLM for this purpose:

$$\mu = exp(\beta_0 + \beta_L ln(L) + \beta_F ln(F) + \beta \mathbf{X}) \tag{4.22}$$

where

μ : estimated crash frequency

L : segment length (miles)

F : traffic volume (e.g. AADT)

$\mathbf{X}$: vector of other features (LSW1, RSW1, etc.)

β_L, β_F, β : NB model coefficients

β_0 : intercept

Note that, for two features F (AADT) and L (segment length), their natural log forms are used so that their transformations lead to the case of zero crashes for zero values. This is a common functional form in crash data modeling of road segments [67, 92].

The coefficients of the models have been estimated using StatsModels [93], a Python-based module providing different classes and functions for statistical modeling and analysis. The dispersion parameter, α is determined using auxiliary Ordinary Least Square (OLS) regression without constant [94]. The NB model coefficients and goodness-of-fit statistics have been estimated using maximum likelihood method and are presented in Table 4.2. Note that other variables including LSW1, LSW2, RSW1, and terrain type were not statistically significant. The dispersion parameter, α, is significantly different from zero which confirms the appropriateness of the NB model relative to the Poisson model. Consistent with expectation, the features MW (median width), F (AADT), and L (segment length) are all positively correlated with crash frequency, and increasing the width of right shoulders decreases the crash frequency. Note that the SPF coefficients in Table 4.2 are used in Equation 4.17 to generate simulated crash frequencies.

4.6.2 CGAN models

The CGAN models for this study have been developed using Keras [95], an open-source deep neural network library developed in Python. The architectures of the generator and discriminator are presented in Figure 4.3. These architectures are designed based on suggested architectures in [84] for using CGAN as a regression model. *DenseLayer(n)* in Figure 4.3 is a regular deeply connected neural network layer with n nodes and *ConcatLayer(n)* concatenates a list of inputs. The model configuration parameters are set as follows:

- Activation functions: Exponential Linear Unit (ELU), Rectified Linear Unit (ReLU), and Sigmoid [96]
- Optimizer: Adam [97]

Table 4.2: The NB Model Coefficients for the WA Data

Variables	**Coefficient (β)**	**SE**	**p-value**
Intercept	-11.8	0.36	0.00
ln(L)	0.902	0.018	0.00
ln(F)	1.33	0.033	0.00
RSW2	-0.0627	0.0051	0.00
MW	0.00200	0.00040	0.00
α	0.836	0.052	0.00
Deviance	6.73E+03	na	na
χ^2	1.03E+04	na	na
AIC	1.71E+04	na	na
BIC	-7.77E+04	na	na

SE = Standard Error, na = Not Available, α: Dispersion Parameter, χ^2 = Pearson's Chi-Square Statistic
AIC = Akaike Information Criteria, BIC = Bayesian Information Criterion

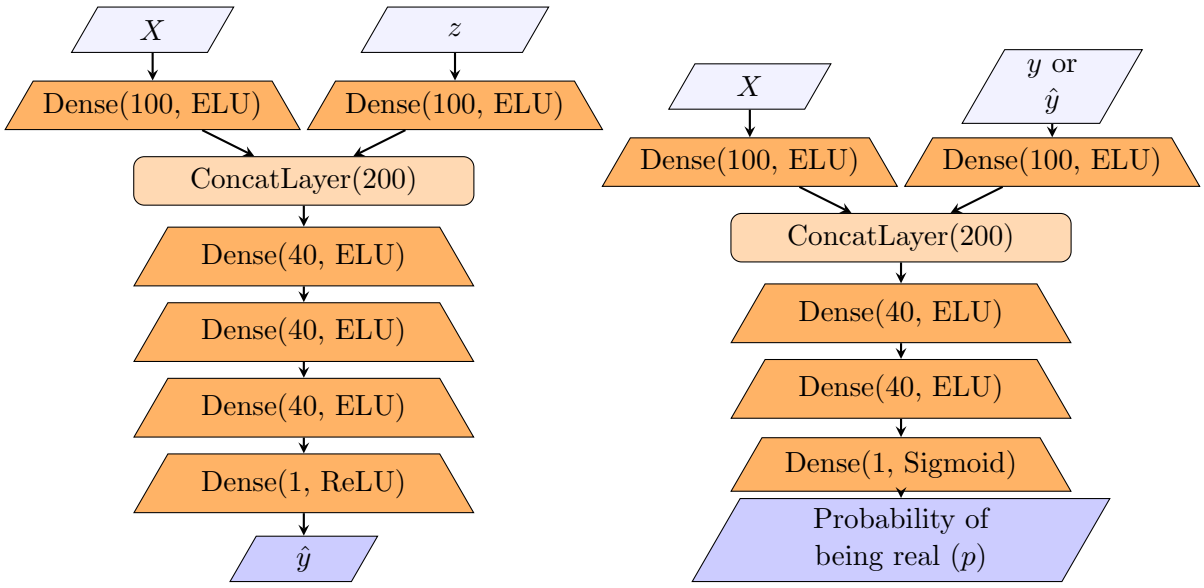
- Number of epochs: 1000
- Batch size: 100
- Learning rate (both generator and discriminator): 0.001
- Learning rate decay (both generator and discriminator): 0.0001

These values are optimized through monitoring the loss function for discriminator and generator during training process. Note that for the hidden layers for both generator and discriminator we have used ELU function, and for the output layer ReLU and Sigmoid functions are used for generator and discriminator respectively.

4.7 Results and discussion

4.7.1 Performance Results: Real-world Crash Data

In Table 4.3, the NB and CGAN models are compared in terms of regression fit of each model developed on the crash data in period P1. The criteria selected for this purpose



(a) Generator architecture

(b) Discriminator architecture

Figure 4.3: CGAN architectures. The input layer of generator includes normalized feature vector with size of 8 and a noise value ($z \sim N(0, 1)$), and input layer of discriminator includes same feature vector and crash count (i.e. y)

include mean absolute error (MAE), mean absolute percentage error (MAPE) and coefficient of determination (R^2 score). According to all three measures, the CGAN models fit the empirical data (i.e. train data set) better than the NB models. The main reason for this is that NB models are constrained to a specified functional form in Equation 4.22 but there is no such constraints in CGAN models.

Table 4.3: Regression evaluation results for CGAN and NB models for WA data

Criteria	**NB**	**CGAN**	**CGAN Improvement**
MAPE	0.56	0.51	8.9%
MAE	0.74	0.69	6.4%
R^2 score	0.38	0.46	21%

MAE = mean absolute error, MAPE = mean absolute percentage error

The results in Table 4.3 evaluate the models in terms of fit to the training data (i.e. period P1). To validate the models and avoid the problems of over-fitting [72] the NB and CGAN models are evaluated on their predictive performance by comparing the accuracy of their predictions for crash counts of period P2 data set as a test set. The results are presented in Table 4.4.

Table 4.4: Predictive performance results for CGAN and NB models for WA data over test data set (P2)

Criteria	**NB**	**CGAN**	**CGAN Improvement**
MAPE	0.48	0.46	5.2%
MAE	0.91	0.88	3.6%
R^2 score	0.34	0.39	15%

MAE = mean absolute error, MAPE = mean absolute percentage error

The results in Table 4.4 indicate that the CGAN shows better performance over the test data set (i.e. P2) which confirms the model has not over-fitted the data. The CGAN model outperformed the NB model in terms of predictive performance because it employs deep neural networks rather than the NB model's limited linear relationship. In the following

sections, multiple more tests and experiments are carried out to evaluate the performance of the CGAN model in terms of network screening.

In order to compare the performance of NB-EB and CGAN-EB in terms of network screening results, the EB estimate for two time periods (P1 and P2) was computed for each road segment. The ranking of hotspots was based on the crash rate (i.e. the EB estimate of the crash frequency divided by the segment length). Tables 4.5 and 4.6 present the observed crash counts over period P1 and the corresponding EB estimates by different methods for the top 10 hotspots identified by the CGAN-EB and NB-EB. Within the top 10 ranked sites, the two methods identified 9 of the same sites and these sites have a similar rank order. As expected, the model predictions are very different from the observed crash count in some cases but the corresponding EB estimates are much closer. The reason is that both methods (CGAN-EB and NB-EB) are using observed crash counts in the EB estimation.

Table 4.5: EB estimates and crash predictions for top 10 hotspots identified by CGAN-EB for Period P1

CGAN-EB rank	Crash Count	L	CGAN Pred.	CGAN-EB	NB Pred.	NB-EB	NB-EB rank
1	29	0.12	18.5	24.4	7.8	26.2	1
2	15	0.06	4.3	10.5	3.0	11.6	2
3	81	0.42	30.7	73.2	18.3	77.2	3
4	23	0.09	2.5	14.0	2.2	15.7	4
5	3	0.03	7.3	4.5	2.6	2.9	22
6	12	0.06	3.9	8.3	2.7	9.1	5
7	97	0.66	36.4	89.5	39.1	95.3	6
8	72	0.51	28.1	65.9	20.5	69.2	8
9	89	0.63	29.3	80.8	23.6	85.8	7
10	39	0.3	25.5	36.7	13.2	36.9	9

Table 4.6: EB estimates and crash predictions for top 10 hotspots identified by NB-EB for Period P1

NB-EB rank	Crash Count	L	CGAN Pred.	CGAN-EB	NB Pred.	NB-EB	CGAN-EB rank
1	29	0.12	18.5	24.4	7.8	26.2	1
2	15	0.06	4.3	10.5	3.0	11.6	2
3	81	0.42	30.7	73.2	18.3	77.2	3
4	23	0.09	2.5	14.0	2.2	15.7	4
5	12	0.06	3.9	8.3	2.7	9.1	6
6	97	0.66	36.4	89.5	39.1	95.3	7
7	89	0.63	29.3	80.8	23.6	85.8	9
8	72	0.51	28.1	65.9	20.5	69.2	8
9	39	0.3	25.5	36.7	13.2	36.9	10
10	21	0.12	2.1	10.3	2.2	14.4	23

The performance of the two methods in identifying hotspots was investigated by conducting the four tests described in Section 4.4. The results of these tests are presented in Table 4.7. The values of each test for the top 2.5%, top 5%, top 7.5%, and top 10% hotspots are provided. In the last column of the table, the relative improvement of CGAN-EB over NB-EB is also presented (positive values means CGAN-EB showed better performance).

Table 4.7: Test scores for the NB-EB and proposed CGAN-EB models

Test	**Top % Hotspots**	**NB-EB**	**CGAN-EB**	**Improvement**
Site Consistency Test (SCT) score	2.5%	494	551	12%
	5.0%	390	415	6.4%
	7.5%	335	385	15%
	10.0%	308	332	7.9%
			AVG	10%
Method Consistency Test (MCT) score	2.5%	53	56	5.7%
	5.0%	96	97	1.0%
	7.5%	144	149	3.5%
	10.0%	194	201	3.6%
			AVG	3.4%
Rank Difference Test (RDT) score	2.5%	72	70	2.8%
	5.0%	138	135	2.2%
	7.5%	176	170	3.4%
	10.0%	211	194	8.1%
			AVG	4.1%
Prediction Difference Test (PDT) score	2.5%	6.2	3.6	41%
	5.0%	4.3	2.9	33%
	7.5%	3.5	2.1	39%
	10.0%	3.1	1.9	38%
			AVG	38%

The SCT scores in Table 4.7 indicates that CGAN-EB outperformed NB-EB in identifying the top 2.5%, 5%, 7.5%, and 10% of hotspots with the highest crash frequency in P2 and improved the scores about 10% on average. The MCT scores show a similar trend to the SCT scores and indicate an average improvement over all four cases of 3.4%. This means that the proposed CGAN-EB approach provides greater consistency than NB-EB in terms of the hotspots identified in the two periods (P1 and P2). In terms of the RDT test, CGAN-EB had better performance for all hotspot levels with an average improvement of 4.1% . Finally, on the basis of PDT scores, the estimates by CGAN-EB are much more consistent over the two periods than by NB-EB suggesting that the total EB estimates of crashes across the hotspots is much more similar in periods P1 and P2 for the CGAN-EB model than the NB-EB model. All of these four tests examine the consistency of the network screening outcomes across two consecutive time periods (P1 and P2). The superior performance of the proposed CGAN-EB method over the NB-EB suggests that it has better temporal transferability than NB-EB, but conclusive statements about temporal and spatial transferability needs further investigation in future studies.

4.7.2 Performance Results: Simulated Crash Data

Using the equations described in Section 4.4.2, five different data sets have been simulated and used to compare the performance of CGAN-EB and NB-EB methods. To this end, the dispersion parameter and sample mean are selected to be similar to the values from the real crash data set (mean = 0.9, dispersion = 0.84) and the following SPF based on the results in Table 4.2 is used:

$$\mu = \exp\left(0.002\text{MW} - 0.0627\text{RSW2} + 1.33\ln(F) - 14\right) \tag{4.23}$$

Here it is assumed that all segments have the same length and the hotspot ranking is performed using EB estimates from each method. For each simulated crash data set, FI, PMD for 2.5%, 5%, 7.5% and 10% top hotspots and the MAPE of EB estimates over all samples have been computed. The average of five values for each metric is presented in Table 4.8.

Table 4.8: Test scores for the NB-EB and proposed CGAN-EB models using simulated data sets

Test	Top % Hotspots	NB-EB	CGAN-EB	CGAN-EB improvement
FI score	2.5%	0.309	0.301	3%
	5%	0.310	0.308	0%
	7.5%	0.324	0.320	1%
	10%	0.334	0.334	0%
			AVG	1%
PMD score	2.5%	0.124	0.121	2%
	5%	0.118	0.118	0%
	7.5%	0.119	0.119	0%
	10%	0.126	0.123	3%
			AVG	1%
MAPE (EB)	-	2.733	2.015	26%

According to the results in Table 4.8, CGAN-EB slightly improved network screening performance in terms of FI and PMD tests (around 1%), while significantly reducing the MAPE of EB estimations (about 26%). This suggests that, while CGAN-EB did not improve network screening performance or hotspot ranking, it did improve overall accuracy in EB estimations.

However, when interpreting these results, it is important to recall that for these simulation experiments, all simulation settings were selected to be completely consistent with the assumptions of the NB-EB method (i.e. conditions are set to optimize NB performance, creating the most difficult conditions for the CGAN-EB to perform better than the NB-EB). Despite this, the performance of CGAN-EB was at least as good as NB-EB in terms of FI and PMD and much better in terms of the accuracy of the EB estimates. As a result, it is expected that CGAN-EB performance relative to the NB-EB performance improves when the simulation settings are changed to become less consistent with the NB assumptions (and arguably a condition that would more accurately reflect conditions encountered with real world data applications). Furthermore, there are several other factors that are known to impact the performance of the NB-EB method such as sample size, sample mean, and dispersion magnitude. Additional investigations are required to examine the performance of the proposed CGAN-EB model across these factors .

4.8 Conclusions

In this chapter, CGAN-EB, a novel EB estimation approach based on CGAN models - a powerful deep generative model - is proposed and its performance is compared to the frequently used NB-EB model. CGAN-EB and NB-EB are used to model both real-world and simulated crash data sets, and are compared in terms of model fit, predictive performance and network screening results. The results from real-world crash data show that the CGAN model has better ability to fit the crash data, and provide more accurate predictions over the same test data set. Several tests have been conducted using both real-world and simulated crash data sets to evaluate the network screening performance of CGAN-EB, and the average score across the four different thresholds of hotspots for all four tests indicate that the proposed CGAN-EB model performs better than NB-EB particularly in terms of consistency of suggested hotspots and the accuracy of EB estimations. All of this evidence indicates that the proposed CGAN-EB method is a powerful crash modelling approach crash frequency modeling and for network screening with performance that is equal to or better than the conventional NB-EB approach.

Despite these very promising results, there remain a number of additional questions regarding the CGAN-EB model proposed in this chapter that need to be investigated. Although we carried out a simulation based evaluation in which we know the true safety state of each site, more work is required to examine the sensitivity of the CGAN-EB performance benefits (vs conventional methods) to key attributes such as the number of observations, the nature of the crash data (e.g. mean and dispersion), etc. Also, spatial and temporal transferability is an important aspect in SPF development which is not been examined in this chapter. In the next chapter, these two questions are more thoroughly investigated.

Chapter 5

CGAN-EB: A Non-parametric Empirical Bayes Method for Crash Frequency Modeling Using Conditional Generative Adversarial Networks as Safety Performance Functions - Sensitivity and Transferability Analysis

This chapter is based on the following journal article:

Zarei, M., Hellinga, B., & Izadpanah, P. (2023). Application of Conditional Deep Generative Networks (CGAN) in Empirical Bayes Estimation of Road Crash Risk and Identifying Crash Hotspots *International Journal of Transportation Science and Technology*, ISSN 2046-0430.

In this journal paper I was the first author and was responsible for the writing of the article. The paper was edited by Dr. Hellinga and Dr. Izadpanah. I also developed the CGAN-EB and GLM-based models models using Python libraries.

5.1 Introduction and Background

In previous chapter, we proposed the CGAN-EB approach and compared its performance to the conventional NB-EB approach used real-world crash data to address the functional form limitation of traditional SPFs (Figure 1.1). Those results showed that CGAN is better able to fit the crash data, produce more accurate crash frequency predictions over the same test data set, and more consistent crash hotspots over different time periods. We compared the performance of CGAN and NB using a real-world data set in which the true values for expected crash counts are unknown and consequently it was not possible to assess the accuracy of the models with respect to truth. Also, individual real-world data sets can represent only a small range of conditions (e.g. dispersion, mean crash rate, sample size). In addition, spatial and temporal transferability is an important asset in SPF development and the previous work did not examine these aspects of the CGAN-EB approach. As a result, the purpose of this study is to address the following two research questions:

- Under what conditions (i.e. sample size, sample mean crash rate, dispersion parameter) does CGAN-EB provide better performance than the conventional NB-EB approach? To this end, we carry out a simulation study for which truth is known and consequently the accuracy of the models can be quantified. Also, we use simulated data to explicitly represent a wide range of conditions, encompassing the range of conditions likely to be encountered in real-world applications. Assessing model performance over the full range of conditions enables us to identify under which conditions the CGAN-EB model provides improvements over the NB-EB approach.
- Does the CGAN-EB approach provide similar or better temporal and spatial transferability compared to the NB-EB approach? In this chapter, the temporal and spatial transferability of CGAN are compared with NB using real-world crash data sets from two different jurisdictions.

5.2 Simulation Experiments

5.2.1 Crash Data Simulation Process

A simulation environment has been frequently used for traffic safety studies [98, 99, 43] specifically because it provides the following two benefits over using empirical data; (a) a

range of specified conditions (e.g. modifying dispersion, mean, data size) can be evaluated, and (b) the true crash risk values (i.e. expected crash counts) at each location are known [42]. The simulation experiments used to evaluate the performance of CGAN-EB versus NB-EB are described in this section. Twelve experiments are designed and presented in Table 5.1 to investigate the impact of dispersion parameter, sample mean crash rate, and sample size on the performance of CGAN-EB versus NB-EB. In order to simulate conditions that are convincingly similar to conditions in empirical crash data sets, the parameters of these experiments (i.e. dispersion, sample mean and sample size) are based on the reported parameters related to crash data sets in eight published chapters summarized in [42]. It should be noted that the sample sizes for each experiment are chosen to be close to or larger than the recommended minimum sample size given each sample mean (according to [8]) to provide the best conditions for NB model development and to minimize the unreliably estimated dispersion parameter. The minimum recommended sample sizes for sample means of 1.5 crashes/year and 12 crashes/year are 700 and 200 respectively.

Table 5.1: Experiments

	Low Dispersion ($\alpha = 0.5$)		High Dispersion ($\alpha = 1.5$)	
Sample size	Low Mean (1.5)	High Mean (12)	Low Mean (1.5)	High Mean (12)
2000	E1	E2	E3	E4
1000	E5	E6	E7	E8
500	E9	E10	E11	E12

For each of these experiments, five training data sets are randomly generated based on the experiment parameters and using the following steps which have been proposed in [41] and been used in previous simulation studies [42, 43]:

1. Generate a random feature vector with the size of 4 (X_1, X_2, X_3, X_4) from a uniform distribution on $[0, 1]$.

2. Generate the corresponding count Y_i given that the mean for observation i is gamma distributed with the dispersion parameter α and mean equal to 1:

$$Y_i \sim Poisson(\lambda_i);$$
$$\lambda_i = \exp(\beta_0 + 0.05X_1 - 0.05X_2 + X_3 - X_4 + \epsilon_i);$$
$$\exp(\epsilon_i) \sim gamma(1, \alpha).$$

3. Steps (1) and (2) are repeated until the sample size associated with each experiment is reached.

Note that ϵ represents the unobserved heterogeneity following the log-gamma distribution as assumed in the NB model. The sample mean is controlled by β_0 which is equal to 0.5 for "low mean" experiments and 2.5 for "high mean" experiments. It is worth mentioning that these simulation settings are completely consistent with the NB model assumptions regarding the error term distribution, log-linear relationship between dependent and independent variables, constant dispersion parameter, and independence of features.

Figure 5.1 illustrates the distribution of simulated crash counts (i.e. Y_is) for one data set of each experiment.

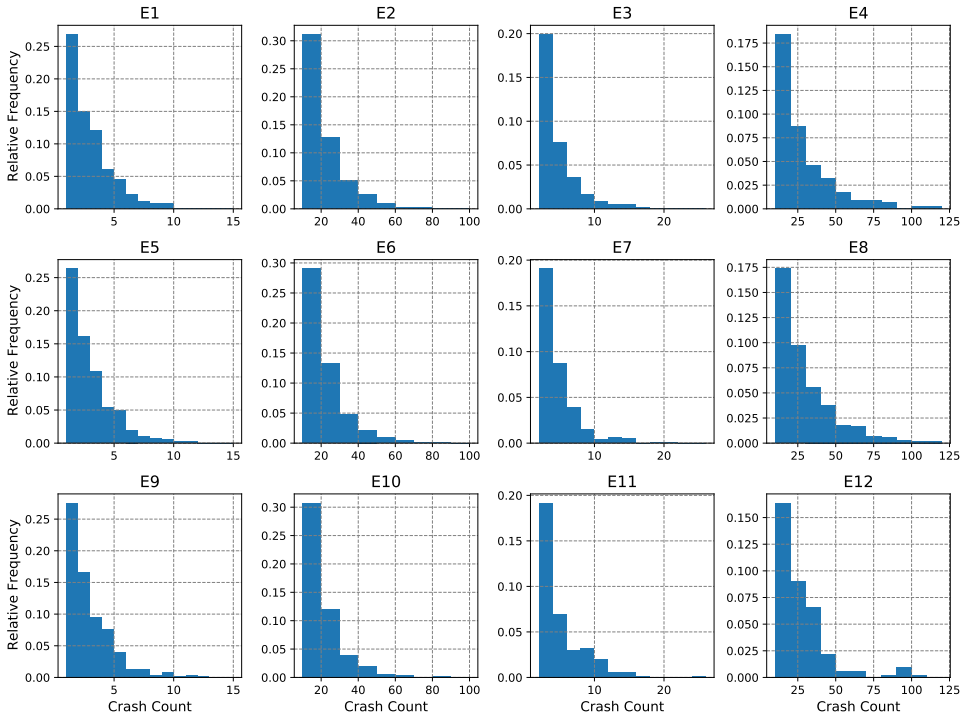


Figure 5.1: Distribution of simulated crash counts for one data set of each experiment

5.2.2 Model Training Process

After simulating all data sets for each experiment, we have trained both NB-EB and CGAN-EB using each training data set. The NB models have been developed using StatsModels

[93], a Python-based module providing different classes and functions for statistical modeling and analysis. The dispersion parameter (i.e. α) is determined using auxiliary Ordinary Least Square (OLS) regression without constant [94]. The CGAN models for this study have been developed using Keras [95], an open-source deep neural network library written in Python. The architectures of the generator and discriminator are presented in Figure 5.2. These architectures are designed based on suggested architectures in [84] for using CGAN as a regression model. *DenseLayer(n)* in Figure 5.2 is a regular deeply connected neural network layer with n nodes and *ConcatLayer(n)* concatenates a list of inputs. The model configuration parameters are set as follows:

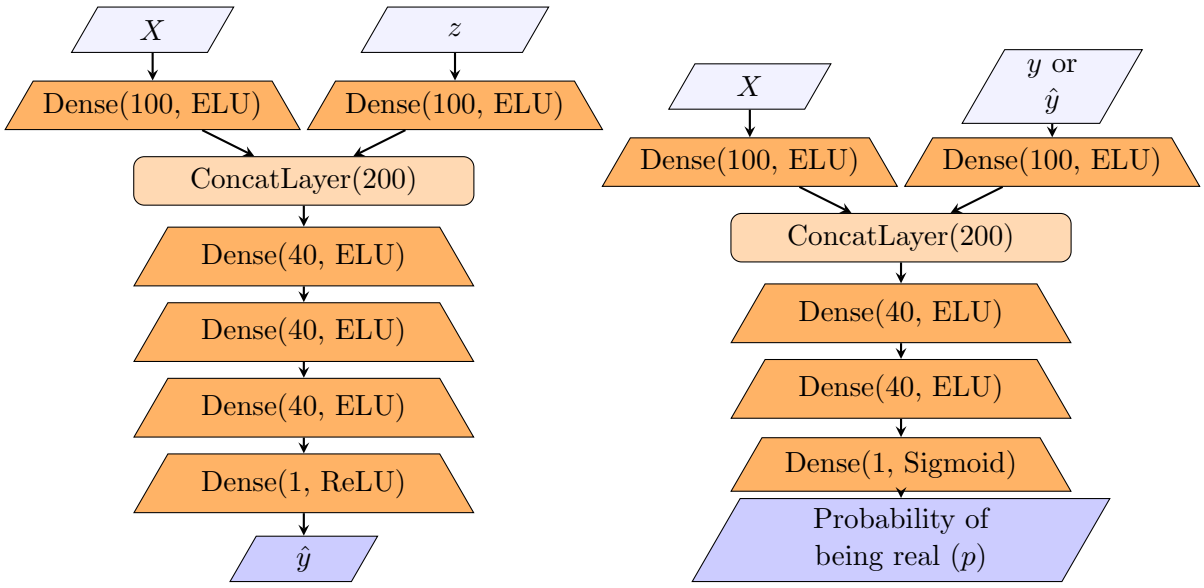
- Activation functions: Exponential Linear Unit (ELU), Rectified Linear Unit (ReLU), and Sigmoid [96]
- Optimizer: Adam [97]
- Number of epochs: 500
- Batch size: 100
- Learning rate (both generator and discriminator): 0.001
- Learning rate decay (generator): 0.001
- Learning rate decay (discriminator): 0.0

5.2.3 Evaluation Methods

Three error measures have been used to compare CGAN-EB versus NB-EB in terms of their performance as crash hotspot identification methods:

1. False Identification (FI) test
2. Poison Mean Difference (PMD) test
3. Mean Absolute Percentage Error (MAPE) of EB estimates

FI and PMD tests are proposed in [87] and have been specifically used for comparing hotspot identification methods in a simulation environment where the truth is known. The FI test calculates the number of sites that are erroneously categorised as hotspots, and the PMD test is the mean absolute difference of the true Poisson means for true hotspots and the hotspots identified by a method. Because we are comparing a variety of conditions and data sets in this simulation study, a normalised version of these tests is proposed and used. MAPE is used to compare the performance of CGAN-EB and NB-EB in terms of the estimation accuracy for Poison mean. In numerical form, the normalized FI and PMD tests and MAPE are as follows:



(a) Generator architecture

(b) Discriminator architecture

Figure 5.2: CGAN architectures. The input layer of generator includes normalized feature vector with size of 8 and a noise value ($z \sim N(0, 1)$), and input layer of discriminator includes same feature vector and crash count (i.e. y)

$$FI_m = \frac{|\{h_{r=1}, h_{r=2}, ..., h_{r=R}\} - \{x_{r=1}, x_{r=2}, ..., x_{r=R}\}_m|}{|\{h_{r=1}, h_{r=2}, ..., h_{r=R}\}|} \tag{5.1}$$

Where $x_{r=1}, x_{r=2}, \ldots, x_{r=R}$ are the sites suggested as hotspots by the method m that are ranked $1, 2, \ldots, R$ respectively, and $h_{r=1}, h_{r=2}, \ldots, h_{r=R}$ are the true hotspots that are ranked based on the true Poisson means. R is the rank threshold that is used as a cut-off in the hotspot identification.

$$PMD_m = \frac{\sum \lambda_h - \sum \lambda_x}{\sum \lambda_h} \tag{5.2}$$

Where $\sum \lambda_h$ is the sum of the true Poisson mean crash frequencies for true top R hotspots and $\sum \lambda_x$ is the sum of true Poisson mean crash frequencies for the sites suggested as the top R hotspots by method m.

$$MAPE_m = \left(\sum_i^R \frac{|EB_i^m - \lambda_i|}{\lambda_i} \right) \times \frac{1}{R} \tag{5.3}$$

Where λ_i and EB_i^m represent the true Poisson mean crash frequency and the EB estimate of crash frequency using method m for the site with rank i, respectively.

Both FI and PMD tests have a minimum value of zero (when the method discovers all true hotspots) and a maximum value of 1 (when none of the true hotspots are detected.

Note that for each experiment, five pairs of NB and CGAN models are trained using five different training data sets and each model was evaluated using five separate simulated test data sets in order to avoid potential over-fitting of models. This produced 25 test replications for each experiment.

5.2.4 Results and Discussion

The goal of the twelve experiments presented in Table 5.1 is to compare CGAN-EB with NB-EB in terms of hotspot identification performance (using FI and PMD) and predictive performance (using MAPE) under different sample sizes, sample mean and dispersion parameters. This comparison for each experiment is performed using FI, PMD and MAPE tests for 2.5%, 5%, 7.5%, and 10% top hotspots which are common thresholds of identified hotspots reported in literature [87, 100]. Since there are 25 replications for each experiment, there are 25 test results for each set of hotspots. The average of these 100 simulation

results (i.e. 25 replications times 4 hotspot thresholds) and the percentage change in FI, PMD and MAPE tests for CGAN-EB in comparing to NB-EB are presented in Table 5.2.

Table 5.2: Performance results of CGAN-EB versus NB-EB in terms of Average FI, PMD and MAPE tests

Mean	α	Size	Ex	FI test			PMD test			MAPE test		
				NB	CGAN	Dif(%)	NB	CGAN	Dif(%)	NB	CGAN	Dif(%)
Low	Low	2000	E1	43%	43%	0%	14%	14%	0%	30%	28%	**7%**
		1000	E5	41%	40%	2%	13%	13%	0%	29%	28%	**3%**
		500	E9	45%	45%	0%	15%	15%	0%	31%	28%	**10%**
	High	2000	E3	33%	32%	3%	9%	9%	0%	27%	27%	0%
		1000	E7	33%	33%	0%	10%	10%	0%	29%	29%	0%
		500	E11	33%	32%	3%	9%	9%	0%	26%	26%	0%
High	Low	2000	E2	19%	19%	0%	3%	3%	0%	12%	12%	0%
		1000	E6	19%	19%	0%	3%	3%	0%	12%	12%	0%
		500	E10	21%	20%	5%	3%	3%	0%	13%	13%	0%
	High	2000	E4	12%	12%	0%	1%	1%	0%	10%	10%	0%
		1000	E8	12%	12%	0%	1%	1%	0%	9%	9%	0%
		500	E12	14%	14%	0%	2%	2%	0%	11%	11%	0%

"Dif(%)" = relative difference computed as (NB - CGAN)/NB x 100%, "Ex" = experiment number,
Bold values show statistically significant improvements by CGAN-EB (p-value of paired sample t-test < 0.05).

Based on the results in Table 5.2, FI values have a range of 12% to 45%, PMD values have a range of 1% to 15%, and MAPE values have a range of 9% to 31%. As expected there is a general correlation between all test values meaning that the models with lower MAPE (better predictive performance) showed better performance in FI and PMD (better hotspot ranking). Also, the test values are lower (i.e. better model performance) in the experiments with higher sample mean (i.e. E2, E6, E10, E4, E8, E12) as compared to the experiments with lower sample mean (i.e. E1, E5, E9, E3, E7, E11). For experiments with the same sample mean, those with higher dispersion produced lower errors. This observation was expected as it has been shown that crash data characterized by a low sample mean can seriously affect the estimation of the dispersion parameter [8]. Also, in the EB method, larger weights are assigned to observed crash counts when sample mean and/or dispersion parameter are large (see Equation 4.10 and 4.5). As a result, EB estimates will mostly depend on the observed counts (especially for the top hotspots which usually experience large observed counts when sample mean and/or dispersion is large) rather than the SPF model predictions.

Regarding the impact of sample size, the results suggest that there is not a significant impact on the error magnitudes. This was expected as we set the sample size close to or greater than the minimum sample size recommended by [8] for reliable estimation of SPF model parameters. Consequently, if we had evaluated sample sizes much smaller than the minimum recommended sample size, we would expect to observed increases in test scores (i.e. decreased model performance).

Regarding CGAN-EB versus NB-EB comparison, we performed a paired sample t-test (DOF = 100, p-value = 0.05) for each experiment for FI, PMD and MAPE results. The outcomes indicate that the difference between FI and PMD results of the two methods are not statistically significant in all 12 experiments. Consequently, we can conclude that for these cases, the CGAN-EB method performance is no different than the NB-EB method. However, the improvement in MAPE for E1(7%), E5(3%) and E9(10%) were found to be statistically significant and for these three experiments, CGAN-EB performed better than the NB-EB.

Recall that for these 12 simulation experiments, all simulation settings were selected to be completely consistent with the assumptions of the NB-EB method, and despite this, the performance of CGAN-EB was at least as good as NB-EB. As a result, it is expected that CGAN-EB performance relative to the NB-EB performance improves when the simulation settings are changed to become less consistent with the NB assumptions. To that purpose, we ran four more experiments (F5, F6, F7, F8 referred to as F-experiments) using the similar simulation settings as E5, E6, E7, E8 (referred to as E-experiments) for comparison purposes, but with a log-nonlinear functional form instead of a log-linear functional form

at step 2 of the simulation in Section 5.2:

$$\lambda_i = \exp(\beta_0 + 0.05X_1^{0.5} - 0.05X_2^{0.5} + X_3^2 - X_1X_4 + \epsilon_i) \tag{5.4}$$

Using the same FI, PMD, and MAPE tests, the performances of CGAN-EB and NB-EB are compared within each experiment. The results, which are presented in Table 5.3, indicate that more significant improvements are achieved by CGAN-EB over NB-EB in these F-experiments and in the E-experiments.

The paired sample t-test results for the FI, PMD and MAPE tests show that the CGAN-EB model performs better (statistically significant decrease in mean test score) for the low sample mean experiments (F5 and F7). For the high sample mean cases (F6 and F8), the differences of the two methods were not statistically significant for 5 of the six cases.

Comparing the test results in Table 5.3 to those in Table 5.2 it can be observed that the performance improvement of the CGAN-EB model relative to the NB-EB model is larger for the the F-experiments than for the E-experiments and that these performance improvements are statistically significant for the low crash mean experiments (i.e. F5 and F7).

These results are consistent with expectation. When sample mean crash frequency is large, the EB network screening results are more heavily influenced by the observed site crash frequencies and consequently, improved representation of the underlying crash data distribution through the use of CGAN provides small improvements in network screen outcomes. However, when mean crash frequency is small, then the accuracy of the SPF plays a more prominent role in the network screening outcomes, and the enhanced capabilities of the CGAN to represent the distribution of the crash data, particularly when the data do not conform exactly to the assumptions of the NB model, provide significant improvements in the network screening outcomes.

Table 5.3: Performance results of CGAN-EB versus NB-EB in terms of FI, PMD and MAPE tests

Mean	α	Size	Ex	FI test			PMD test			MAPE test		
				NB	CGAN	Dif(%)	NB	CGAN	Dif(%)	NB	CGAN	Dif(%)
Low	Low	1000	F5	40%	39%	**3%**	13%	13%	**2%**	29%	28%	**6%**
	High	1000	F7	32%	31%	**3%**	9%	9%	**3%**	28%	27%	**3%**
High	Low	1000	F6	19%	19%	0%	3%	3%	0%	12%	11%	**1%**
	High	1000	F8	12%	12%	0%	1%	1%	0%	9%	9%	0%

Dif(%): relative difference, Ex: experiment,
Bold values show statistically significant improvements by CGAN-EB (p-value of paired sample t-test < 0.05).

In summary, the results in this section suggest that when all conditions and assumptions are set in favor of NB-EB, then CGAN-EB performs at least as good as NB-EB. However, when we change some of the simulation conditions such that they are not consistent with the assumptions of the NB model (such as the functional form) the difference between the two methods becomes more obvious and the CGAN-EB shows better performance especially when the sample mean is low. In the next section, both models are compared in terms of temporal and spatial transferability.

5.3 Evaluation of Model Transferability

The temporal and spatial transferablity of the CGAN and NB models were evaluated using empirical data sets from two different jurisdictions (Peel and Niagara). We used real-world data sets because there are numerous factors (traffic volumes, rules of the road, network structure, enforcement, land use, driving behaviour, vehicle types and technologies, demographics of travellers, pedestrian and cyclist volumes, etc.) that cause temporal and spatial differences in crash data that are nearly impossible to account for in simulation. Consequently, we selected these two jurisdictions which are quite different in terms of traffic volume and crash counts. The following experiments were conducted to evaluate the spatial (experiment T1 and T2) and temporal (experiments T3 and T4) transferability of the models:

- T1: CGAN and NB models are trained using Peel data (2010-2013) and tested on Niagara data (2010-2013)
- T2: CGAN and NB models are trained using Niagara data (2010-2013) and tested on Peel data (2010-2013)
- T3: CGAN and NB models are trained using Peel data (2010-2013) and tested on Peel data (2014-2017)
- T4: CGAN and NB models are trained using Niagara data (2010-2013) and tested on Niagara Data (2014-2017)

Note that using a four-year period of data for model development is based on Highway Safety Manual (HSM) recommendation and previous study results [56, 3].

5.3.1 Crash Data Sets

The data consists of fatal crash records and injury crash records, traffic volumes in the form of average annual daily traffic (AADT) for minor and major approaches of urban four-legged signalized intersections within the Regional Municipalities of Niagara and Peel from 2010 to 2017. Both of these jurisdictions are located in south western Ontario, Canada. The summary statistics of the data sets are presented in Table 5.4.

Table 5.4: Descriptive statistics of the data

	Measures	AADT (major)	AADT (minor)	FI Crashes per year
Peel	sites	329	329	329
	mean	33746	11694	2
	std	14367	10877	2.5
	min	2505	24	0
	25%	23868	3867	0
	50%	33956	7744	1
	75%	44472	15923	3
	max	79071	55742	25
Niagara	sites	223	223	223
	mean	14059	6145	0.7
	std	6620	4249	1
	min	1762	113	0
	25%	9489	2821	0
	50%	12959	5227	0
	75%	17770	8578	1
	max	63102	27518	8

FI = Fatal and injury crashes per year

5.3.2 Results and Discussion

CGAN and NB models are developed using the same procedure as done for the simulation experiments but in this case, the training data are the observed field data. The resulting NB models have the following functional form and the coefficients are presented in Table 5.5:

$$y = exp(\beta_1 \times \text{AADT}_{major} + \beta_2 \times \text{AADT}_{minor} + \beta_0) \tag{5.5}$$

where $AADT_{major}$ and $AADT_{minor}$ are AADT for major and minor approaches, β_1 and β_2 are model coefficients, and β_0 is intercept.

Table 5.5: Coefficients, dispersion parameter, and fit metrics for NB Models

Train data set	**Variables**	**Coefficient (β)**	**SE**	**p-value**
Peel (2010-2013)	β_0	-11.9	0.61	0.00
	β_1	0.69	0.06	0.00
	β_2	0.6	0.03	0.00
	α	0.28	0.05	0.00
	Deviance	1501	na	na
	χ^2	1610	na	na
	AIC	4709	na	na
	BIC	-7928	na	na
Niagara (2010-2013)	β_0	-9.8	0.61	0.00
	β_1	0.56	0.06	0.00
	β_2	0.48	0.03	0.00
	α	0.21	0.14	0.00
	Deviance	889	na	na
	χ^2	950	na	na
	AIC	1912	na	na
	BIC	-5150	na	na

SE = Standard Error, na = Not Available, α: Dispersion Parameter, χ^2 = Pearson's Chi-Square Statistic
AIC = Akaike Information Criteria, BIC = Bayesian Information Criterion

In order to improve the model performances, the outputs of each Base SPF are adjusted using the Highway Safety Manual calibration method which calculates the ratio of the total observed number of crashes (N_{obs}^{tot}) to the total number of crashes predicted from each model

(N_{pred}^{tot}) for each test data sets [3]:

$$N_{pred}^{cal} = N_{Pred} \times \frac{N_{obs}^{tot}}{N_{pred}^{tot}} \tag{5.6}$$

Mean Absolute Percentage Error (MAPE) computed across all sites in the test data set is used to evaluate the performance of the models in each experiment using each test data set. The results, presented in Table 5.6, show that although the CGAN model seems to outperform NB model on average across all four experiment, the differences in MAPE values are not considerable. These results suggests that CGAN model provides comparable temporal and spatial transferability as NB model.

Table 5.6: Temporal and spatial transferability results for CGAN and NB models

	Experiment	MAPE(NB)	MAPE(CGAN)	% Improvement
Spatial Transferability	T1	0.58	0.59	-1.7%
	T2	0.67	0.66	1.5%
Temporal Transferability	T3	0.59	0.59	0%
	T4	0.66	0.63	4.5%

5.4 Conclusions and Recommendations

In this study, the performance of non-parametric empirical Bayes method based on conditional generative adversarial network (CGAN-EB) is compared with the traditional parametric approach using negative binomial model (NB-EB) in a simulation environment. Several experiments have been conducted, each of which include simulating the data sets with defined parameters, fitting CGAN and NB models, and estimating EB estimates. Then the models are compared based on their ability to detect correct hotspots (using FI and PMD tests) as well as their accuracy of EB estimates (using MAPE). The results show that both models perform better in cases with larger sample mean and dispersion parameters and their performance was not impacted by sample size (but we only examined the influence of sample size for the range of sample size equal to or greater than the minimum sample size recommended by [8]). More importantly, the results showed that the CGAN-EB model provided statistically significant improvements in performance in experiments with low sample means (sample mean of 1.5 in our study) and when the underlying crash

data characteristics do not exactly conform to the assumptions of the NB model. These conditions are commonly experienced in real-world crash data sets [14].

We also examined the spatial and temporal transferability of the models through the use of a time series of crash data from two different jurisdictions. The results showed that the CGAN-EB and NB-EB models provided similar spatial and temporal transferability performance.

These findings are important as they suggest the CGAN-EB model is more robust than the traditional NB-EB model in that it is able to perform as well or better than the NB-EB model over a wide range of conditions and the CGAN-EB model provides statistically significant improvements in performance when the sample mean is low and the dispersion is low or when the sample mean is low and the underlying crash data do not conform exactly to the assumptions of the NB model.

Notwithstanding these highly promising results, a number of questions remain about the proposed CGAN-EB approach that need to be examined in the future.

1. The performance of CGAN is impacted by its configuration (e.g. architecture, size). For all data sets in this chapter, we used a simple network architecture; however, other types of architectures might be better options. Consequently, it is recommended to examine the sensitivity of the CGAN-EB performance as a function of the CGAN configuration and to determine if different configurations are better suited for different types of network screening applications.

2. There are other approaches than NB-EB for network screening such as finite mixture or zero-inflated models, which have been shown to be better alternatives in some conditions [99, 101]. It is recommended to compare the performance of CGAN-EB with such alternatives for crash data sets for which these models are more appropriate than the NB-EB.

3. Finally, studies have shown that using hierarchical full Bayesian method outperforms the standard EB approach in correctly identifying hazardous sites [10, 102]. Future studies can focus on investigating how to combine this method with a non-parametric model such as CGAN model in order to improve the network screening performance.

Chapter 6

Crash Data Augmentation Using Conditional Generative Adversarial Networks (CGAN) for Improving Safety Performance Functions

This chapter is based on the following journal article:

Zarei, M., Hellinga, B.(2023). Crash Data Augmentation Using Conditional Generative Adversarial Networks (CGAN) for Improving Safety Performance Functions *Transportmetrica a: transport science*.

In this journal paper I was the first author and was responsible for the writing of the article. The paper was edited by Dr. Hellinga and Dr. Izadpanah. I also developed the CGAN-EB and GLM-based models models using Python.

6.1 Introduction

In the previous two chapters, we investigated using CGAN models as an alternative to traditional models to create SPFs. In this chapter, CGAN is applied as a crash frequency data augmentation method to address low sample issue in developing SPFs. The proposed method is evaluated by comparing the performance of Base-SPF (SPFs developed using original data) and Augmented-SPF (SPFs developed using original data plus synthesized data) in terms of hotspot identification performance (i.e. False Identification (FI) and Poisson Mean Difference (PMD) tests), accuracy of estimated long-term crash means, crash frequencies, and dispersion parameters. The experiments are conducted using both real-world and simulated crash data sets.

One of the main challenges in developing reliable SPFs is to have a sufficiently large sample size, particularly as crash data sets frequently have small sample means and large dispersion due to preponderance of zeros ([14, 90]). Excess zeros in crash data can create difficulties when modeling and estimating SPFs, as traditional modeling approaches such as Poisson or negative binomial regression may not adequately capture the presence of these excess zeros. This may lead to biased parameter estimates, reduced model fit, and ultimately, less accurate predictions of crash frequencies. Also, the desirable large-sample properties of various parameter-estimation approaches (for example, maximum likelihood estimation) are not achieved with small sample sizes. There are several approaches to address this issue and increase the size of a crash data set. A common method is to use crash data from 3 to 5 consecutive years ([103, 104, 3]). However, this approach introduces several challenges. Annual average daily traffic (AADT) data are rarely available for all locations for all years and the use of imputation introduces errors ([56]). Furthermore, though increasing the number of years of data increases the sample size, it also introduces temporal variations associated with changes in the road network characteristics, driving behaviours, levels of enforcement, vehicle technologies, etc., that confound the underlying SPF relationships. In addition, sometimes the network contains so few sites of a certain type (e.g. roundabouts, ramps) that reliable SPFs cannot be developed even when expanding the number of years of crash data. In this case, the common practice is to combine two or more site types and include categorical variables within the SPF to distinguish site type or to use an SPF developed using data from another jurisdiction.

An alternate approach, and the one that is the subject of this chapter, is to expand the existing crash data set by augmenting it with synthesized data (i.e. data that is artificially produced using statistical models rather than generated by real-world events).

Advances in the field of deep learning models have resulted in the creation of a suite of generative models such as variational auto encoders (VAE) and conditional generative

adversarial networks (CGANs) that can be used to deal with imbalance or small size data sets. VAE and CGAN have been successfully used in the literature for preparing balanced data sets which is critical for developing proper classification models ([105, 79, 106]). However, such oversampling methods have rarely been used for improving SPF development as a regression model [107]. In this chapter, we propose a data augmentation method based on CGAN and evaluate its performance using both simulated and real-world crash data sets.

The remainder of this chapter is organized as follows. The next section presents a background on SPFs, crash data augmentation, and CGAN. Section 6.3 describes the proposed crash data augmentation methodology. The evaluation process and results are discussed in Section 6.4. Finally, conclusions and recommendations for future works are presented in section 6.6.

6.2 Background and Related Works

SPFs are regression models for estimating the average crash frequency of road segments or intersections. Traditionally, SPFs are developed using a statistical model such as negative binomial (NB) model including both traffic and geometric factors. An example of such a parametric SPF based on NB model might be as follows in terms of generic functional form ([48]):

$$\mu = \exp(\beta_0 + \beta \times \ln(\text{AADT}) + \beta_1 X_1 + \beta_2 X2 + ... + \beta_n X_n) \tag{6.1}$$

where μ is the predicted crash frequency, AADT is annual average daily traffic, X_1, X_2, ... , X_n are n roadway geometric variables, and β_0, β, β_1, β_2, β_n are the regression coefficients.

One of the main applications of SPFs is for crash hotspot identification which is also referred to as network screening. Conventional network screening uses an empirical Bayes (EB) approach first proposed by [19] in which the EB estimate for long-term crash mean is calculated using a weighted average between the predicted crash frequency from SPFs and the observed number of crashes. Then sites are ranked based on their EB estimates to identify the top hotspots for further investigation and safety improvement ([3]). Consequently, the accuracy of SPFs can have significant impacts on network screening results which can be converted into monetary values using societal costs of crash types ([108]). The SPF development process can be quite challenging due to methodological issues associated with crash data such as inadequate sample size, over-dispersion, time-varying explanatory variables, and temporal/spatial correlation to name but a few ([14]). In this chapter, we aim

to propose a data augmentation method for improving SPFs by addressing the inadequate sample size issue.

Crash data augmentation has rarely been used for improving the quality of crash frequency prediction models (i.e. for SPF development [107]). However, there are several works that used different methods to augment (or balance) crash data sets for classification purposes (e.g. real time crash prediction/detection, crash severity prediction). In general, there are four main data balancing methodologies used in the literature to handle imbalanced crash datasets (i.e. crash and non-crash events) including random under-sampling of majority class ([109, 110]), random over-sampling of minority class ([111]), synthetic minority oversampling technique (SMOTE) ([112, 113, 114]), and using deep generative models ([105, 79, 72]).

The under-sampling method involves deleting (ignoring) records from the majority class (typically non-crash events) such that the crash/non-crash event ratio is adequate for developing models. However, because this approach involves ignoring a lot of data, this can result in some information loss ([109]). Over-sampling, on the other hand, involves randomly selecting more samples from the minority class (typically crash events) in order to achieve the desired class ratio. However, this approach is susceptible to model over-fitting, and consequently, this method is rarely employed in practice ([111]).

Instead of duplicating samples from the minority class, in the SMOTE method, new samples are synthesized from the minority class. This widely used method works by utilizing a k-nearest neighbour algorithm to create synthetic data ([115]). Thus, it only takes into account the closeness of the samples and the variable correlation is not taken into consideration. As a result, it creates the same number of synthetic samples for all original minority classes, including those near the decision border of the majority and minority classes, increasing the likelihood of class overlapping ([116]). Several variations of the SMOTE method have been proposed to address these issues such as Borderline-SMOTE ([117]), Safe-Level- SMOTE ([118]), Local-Neighborhood-SMOTE ([119]), and Adaptive-SMOTE or ADASYN ([120]).

More recent data augmentation methods include using deep generative models which are neural network based models that are trained to represent an estimation of the underlying distribution of data ([121]) and have been demonstrated to outperform the other three approaches for data augmentation, overcoming their limitations ([105, 79, 72]). In a recent work ([107]), variational auto encoder (VAE) has been employed to fit the non-zero crash counts and generate more synthesized non-zero crash data to balance the ratio of zero-crash to non-zero crash cases. In terms of root mean square error (RMSE) and mean absolute error (MAE) of model estimations, the results show that crash models developed

on balanced data from the trained VAE model outperformed the crash models developed on original imbalanced crash data, and the the crash models developed on balanced data from SMOTE method.

Another type of deep generative models is generative adversarial networks (GANs) that has been widely used in various fields ([122]). In crash data modeling for instance, recent studies ([79, 72]) have shown that real time crash prediction models based on augmented data sets generated by GAN provided the best prediction accuracy as GAN is able to generate data that more closely mimics the characteristics of the real data.

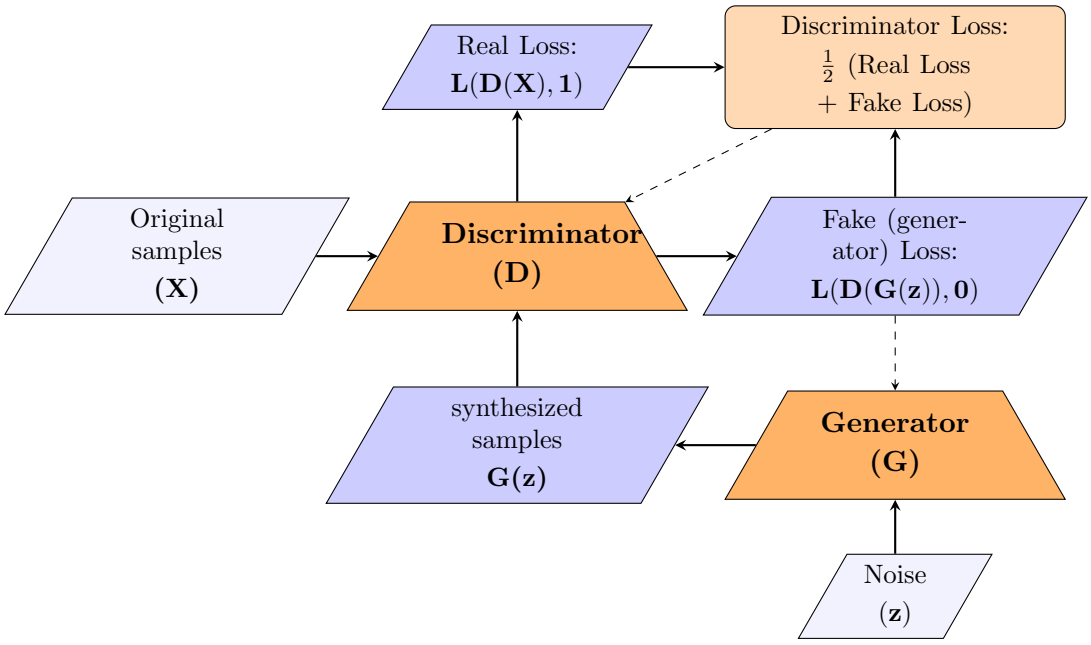


Figure 6.1: GAN training structure

GAN includes competitive training of two neural networks (generator G and discriminator D in Figure 6.1). The generator's goal is to generate samples from the same distribution as the training data, and the discriminator's goal is to determine if the data are real or fake ([123]). Since both networks are trained simultaneously based on each other's feedback, both networks are forced to increase their performances after each cycle. This process can be formalized as a min-max function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p(x)}\left[\log(D(x))\right] + \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D(G(z))\right)\right] \tag{6.2}$$

where $p(x)$ is the training data distribution, $p(z)$ is the prior distribution of the generative network, and z is a noise vector sampled from the model distribution $p(z)$ such as the Gaussian or uniform distribution.

In Figure 6.1, $D(X) \in [0, 1]$ is the output of the discriminator (real/fake classifier) using real data (X) as input, and $G(z)$ is the output of generator using noise (z) as input. Real loss value shows the ability of the discriminator to recognize the real instances (i.e. X), and is calculated based on $D(X)$, unit vector (i.e. $\mathbf{1}$) and a loss function (e.g. binary cross entropy). Fake loss shows the ability of the discriminator to recognize fake instances (i.e. $G(z)$), and is calculated based on $D(X)$, zero vector (i.e. $\mathbf{0}$) and a loss function (e.g. binary cross entropy). In each cycle, the weights of the discriminator network are updated based on the objective to minimize total loss (i.e. real loss + fake loss) and the weights of the generator will be updated based on its objective to maximize fake loss. This cycle continues until the stop condition (e.g. maximum number of epochs) is met. In an ideal training condition, both fake loss and real loss converges to 0.5 which indicates that it is impossible to distinguish between input real data and synthetic data because they are samples of the same distribution ([83]).

GAN and its variants have been widely used in different transportation applications recently including real time crash prediction ([105, 79, 124, 125]), crash frequency modeling [25], traffic flow data prediction ([86, 85]), and autonomous driving ([126, 127]) to name but a few. There are several GAN variants based on different architectures or loss functions proposed in the literature ([128]). One of the popular variants is conditional-GAN (CGAN) ([83]) in which both the generator and the discriminator are conditioned on some data which could be a class label or a feature vector if we wish to use it for regression purposes. The goal of this chapter is to propose a crash data augmentation method based on CGAN that can improve the performance of SPFs. The details of the method are presented in the next section.

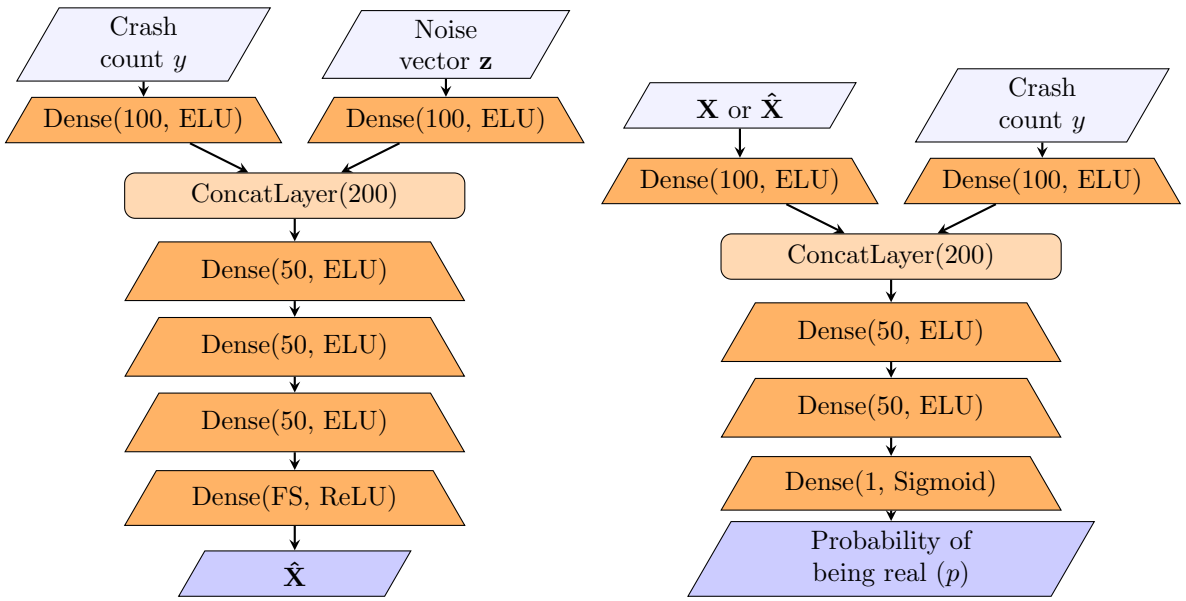
6.3 Methodology

Conditional generative adversarial network, or CGAN for short, is a powerful deep generative model that has seen considerable applicability in many areas in recent years. To the best of the authors knowledge this is the first time CGAN is being used for crash frequency data augmentation for improving safety performance functions as a regression problem. All previous works in this area have used CGAN as a data augmentation method for a classification problem such as crash/non-crash detection and crash severity prediction. The proposed method has two main steps. The first is to design and train CGAN based on the

original crash data set, and then use the trained generator to generate synthesized crash data to develop the SPF.

6.3.1 Design and Train CGAN

For crash frequency data augmentation, we used the generator and discriminator models shown in Figure 6.2. *Dense(n)* in Figure 6.2 is a regular deeply connected neural network layer with n nodes and *ConcatLayer(n)* concatenates a list of inputs. The first layer of each network is an input layer. For the generator, the input layer includes a noise vector (z) with the same size as the feature vector (X) and crash count (y) and the output is a generated feature vector ($\hat{X}$) with size of FS (feature vector size). For the discriminator, crash counts (y), feature vector X as real feature vector and $\hat{X}$ as generated feature vector are included in the input layer. The output of the discriminator is a number between 0 and 1 indicating the probability of being a real sample for the the given input. Also the activation functions used in the networks include Exponential Linear Unit (ELU), Rectified Linear Unit (ReLU), and Sigmoid ([96]).



(a) Generator architecture (b) Discriminator architecture

Figure 6.2: Network architectures. The input layer of generator includes crash count (i.e. y) and a noise vector ($\mathbf{z}$) with the same size of the feature vector (FS), and input layer of discriminator includes normalized feature vector and crash count (i.e. y)

After randomly initializing the weights of each network, in each cycle of CGAN training, the weights of the generator and discriminator networks are updated to minimize the corresponding loss functions ([83]):

$$Loss(D) = -\frac{1}{2}\Big(\log(D(X|y)) + \log[1 - D(\hat{X}|y)]\Big) \tag{6.3}$$

$$Loss(G) = -\log(D(\hat{X}|y)) \tag{6.4}$$

where $D(X|y)$ and $D(\hat{X}|y)$ are the discriminator outputs given feature vector X and generated feature vector $\hat{X}$ conditioned on crash count y. In an ideal equilibrium, both loss values will be equal to $-\log(0.5)$. In this case, the discriminator cannot distinguish between real and fake samples generated by the generator.

6.3.2 Generate Synthesized Crash Data and Develop SPFs

The generator network can be used to generate synthesized crash data once the CGAN has been trained. A random sample from the empirical distribution of crash counts (both crash and non-crash cases) and a random noise vector are input into the generator network for each synthesized data point. This procedure is continued until the desired number of synthesized data is attained. Finally, the safety performance functions are developed using the augmented crash data set (i.e. real plus synthesized data). SPF development can be based on any approach accessible, including machine learning and traditional parametric models. In this study, we have used the NB model as the most common crash frequency modeling approach. The proposed method is implemented and evaluated using both simulated and real-world crash data sets. In the next section the results of these experiments are discussed.

6.4 Experiments and Results

6.4.1 Real-World Crash Data

The performance of the proposed crash data augmentation method is investigated in this section using a real-world crash data set. The data, which is referred to as WA data set throughout this chapter, include fatal and injury (FI) crash counts and traffic volumes

of 3085 individual road segments from the Highway Safety Information System (HSIS) collected for divided 4-lane segments from urban freeways for 2016 and 2017 in Washington State, USA. Table 6.1 provides a summary of statistics for this data set.

Table 6.1: Descriptive statistical features of WA data set

	MW (m)	SW (m)	AADT	Length (m)	Fatal Injury Crash
Sites	3085	3085	3085	3085	3085
Mean	12.2	2.6	49000	135	0.29
STD	12.7	1.1	28412	191	0.91
MIN	0.61	0.0	5082	16.1	0
25% Quantile	4.9	3.0	27385	32.2	0
50% Quantile	12.2	3.0	45875	64.4	0
75% Quantile	14.6	3.0	64507	161	0
MAX	229	7.3	178149	3250	23

Notes: MW = Median Width, SW = Shoulder Width. AADT = Annual Average Daily Traffic volume

Experiment Setup

In the first set of experiments, we investigate the performance of the proposed data augmentation method across different sample sizes. To accomplish this, we draw random samples of different sizes from the WA data set and compare the corresponding "Base-SPFs" (SPFs developed using the drawn sample) and "Augmented-SPFs" (SPFs developed using the augmented sample) with the "True-SPF" which is defined as the SPF developed using the entire WA data set. The sample mean is 0.29 crashes per segment per year. According to [8], the recommended minimum sample size is 4000. The total sample size for WA data set is 6170 observations, which far exceeds the minimum recommended sample size and justifying treating the SPF developed on the entire WA data set as "truth".

The WA data set is divided into a training data set containing 5000 observations (2500 sites) and a testing data set containing 1670 observations (585 sites).

For each experiment we randomly select a sample of observations from the training data set and we vary the sample size by considering five sample size ratios (5%, 15%, 25%, 50%, and 75% of the training data set size (i.e. 5000)). For each experiment we performed the following steps for each sample size ratio:

1. A sub-sample is drawn at random from train data set based on the given sample size ratio.

2. The sub-sample is used to develop Base-SPFs as well as train a CGAN model.

3. The trained CGAN is used to increase the size of the sub-sample to the recommended sample size (i.e. 4000 data points). The set of data containing both the sub-sample of train data set and the data observations synthesized by the CGAN model is called the augmented training sample.

4. The Augmented-SPF is developed using the augmented training sample..

5. For each sample size ratio, these steps are repeated five times.

Model Development

The coefficients of the SPFs have been estimated using a Poisson–gamma regression model with a fixed dispersion parameter in StatsModels [93], a Python-based module providing different classes and functions for statistical modeling and analysis. The dispersion parameter, α is determined using auxiliary Ordinary Least Square (OLS) regression without constant [94]. Also, dummy variables are used for median width (1: median width >= 3m, 0: median width < 3m, where 3m is determined through optimizing goodness-of-fit of the SPF) and shoulder width (1: shoulder width > 0, 0: shoulder width = 0).

The training configuration parameters for training CGAN are set as follows:

- Optimizer: Adam ([97])
- Batch size: 100
- Learning rate (both generator and discriminator): 0.001
- Learning rate decay (both generator and discriminator): 0.001

These values are optimized through monitoring the loss function for discriminator and generator during training process, both of which should converge to -log(0.5).

Detailed Results from A Single Experiment

This section provides a detailed examination of a single experiment before presenting the aggregated results from all experiments for all sample size ratios. A random sub-sample with the size of 2000 (i.e. 50% sample size ratio) was drawn from the train data set and used for developing Base-SPF and training the CGAN. The trained CGAN is then used to augment the sub-sample up to 4000 sample size in order to develop Augmented-SPF.

The distribution of all features for 2000 synthesized data from CGAN and the sub-sample from train data (i.e. with the size of 2000) are presented in Figure 6.3. For all

features, the distribution of the synthesized data closely matches the distribution of the real data visually. To confirm this observation statistically, three tests were carried out:

- The t-test showed that for all cases, there is not enough evidence to conclude that the mean of the feature distributions between synthesized and real data are significantly different (p-values > 0.05).
- The Leven test showed that there is no evidence to conclude that the differences in the variance of the distributions are statistically significant (p-value > 0.05).
- Finally, the Kolmogrove-Smirnov test showed that there is not enough evidence to conclude that the distributions of synthesized data set and real data set are statistically different.

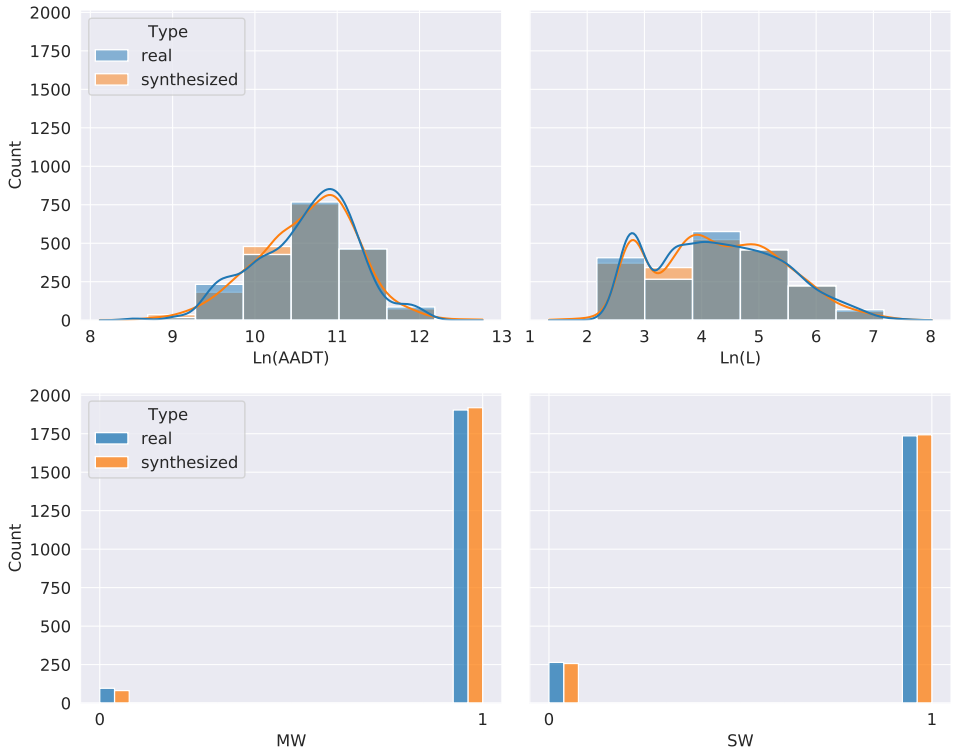


Figure 6.3: Feature distributions for synthesized and real data set

The coefficients and Standard Error (SE) for a Base-SPF and Augmented-SPF for one single experiment with 50% sample size ratio are presented in Table 6.2 and are compared with the coefficients and SE of the True-SPF.

Table 6.2: Coefficients and standard errors for True-, Base- and Augmented-SPFs

	Coefficient Value			Coefficient Standard Error		
	True-SPF	Base-SPF	Aug-SPF	True-SPF	Base-SPF	Aug-SPF
Const.	-19.9	-22.5	-23.3	0.71	1.15	0.67
MW	-0.56	ns	-0.37	0.12	-	0.16
SW	-0.64	-0.78	-0.23	0.095	0.14	0.10
ln(AADT)	1.41	1.55	1.58	0.061	0.10	0.07
ln(L)	0.92	1.03	1.09	0.031	0.05	0.032
Dispersion	1.02	0.60	0.80	0.16	0.22	0.18

ns = not statistically significant (p-value > 0.05), Const. = model constant, Aug-SPF = Augmented-SPF

The goodness of fit for Base-SPFs and Augmented-SPFs have been examined with CURE plots. The CURE plot approach involves graphically observing how well the SPF fits the data set by showing the cumulative residuals for AADT variable. The residuals are computed as the difference between the observed and predicted number of crashes and are ranked from lowest to greatest. In general, a good CURE plot oscillates about zero, and the residuals do not exceed the $\mp 2\sigma$ bounds:

$$\sigma = \left[\sigma_i^2 \times \left(1 - \frac{\sigma_i^2}{\sigma_T^2}\right)\right]^{0.5} \tag{6.5}$$

where σ_i^2 is the cumulative sum of squared residuals until the element i, and σ_T^2 is the total cumulative sum of squared residuals.

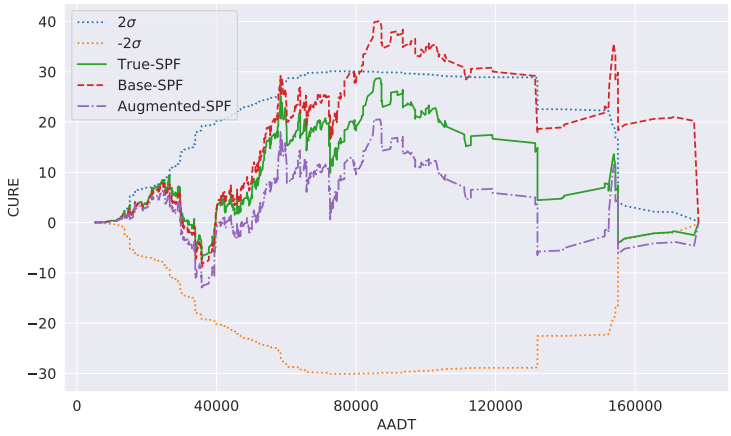


Figure 6.4: Cumulative residual (CURE) plot for Base-, Augmented-, and True-SPF

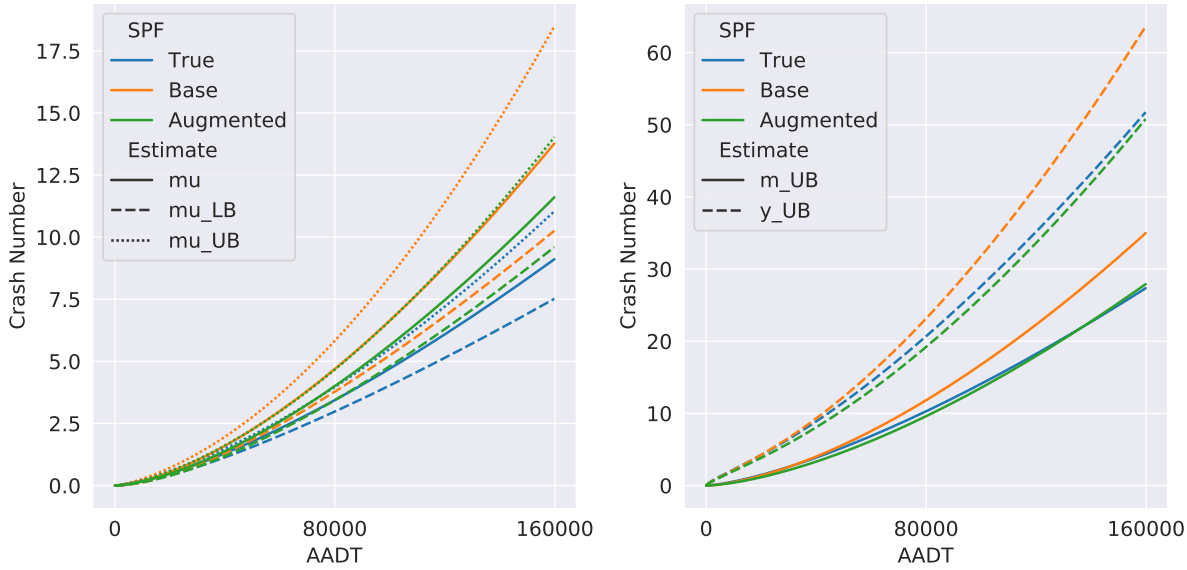


Figure 6.5: 95% confidence intervals for μ (left) and 95% prediction intervals for y and m (right)

In addition, the 95% confidence interval for μ (Equation 6.1), and 95% prediction

interval for y (crash count) and m (Poisson parameter which is referred to as long-term safety ([129, 130])) for a range of AADT are plotted in Figure 6.5. The results in each plot are based on an SPF calculated with the coefficients shown in Table 6.2, in which annual average daily traffic (AADT) is allowed to vary (in increments of 100) between 0 and 160,000 vehicles per day (as this was approximately the range in the WA data set), and segment length is fixed at 1 km. The remaining variables were set to the most common value of each variable in the data set (i.e., MW = 1, SW = 1). For more information about how to calculate these confidence and prediction intervals refer to ([129, 130]).

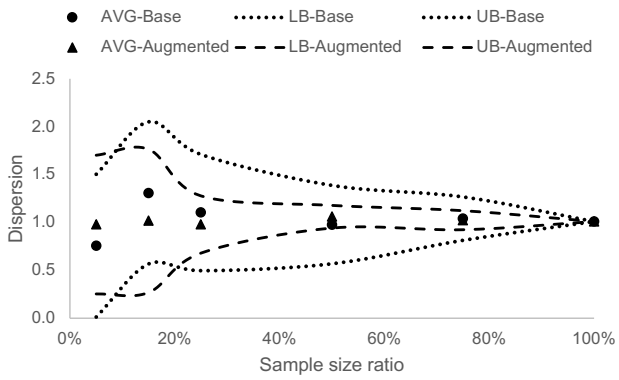
The following observations can be made from the results in Table 6.2, Figure 6.5 and Figure 6.4:

- In the development of the Base-SPF, median width (MW) was not found to be a statistically significant variable but was found to be statistically significant in the Augmented-SPF. Furthermore, Augmented-SPF has lower standard errors than Base-SPF for coefficient estimation.
- Based on the CURE plots for this experiment shown in Figure 6.4, the True-SPF offered the best result, with residuals residing between the $\mp 2\sigma$ boundaries in about 95% of the AADT range. The percentages of in-bound residuals for Augmented-SPF and Base-SPF are about 85% and 60%, respectively, demonstrating that Augmented-SPF produces better outcomes.
- In terms of 95% confidence intervals for mu, as expected, the Base-SPF has the widest range. The range for the Augmented-SPF is narrower and closer to results from True-SPF.
- The lower bound values for the 95% confidence intervals for y and m are not shown because all values for all SPFs, regardless of AADT, were found to be zero. Also, Augmented-SPF was found to produce the more similar upper bounds for y and m to True-SPF than Base-SPF.

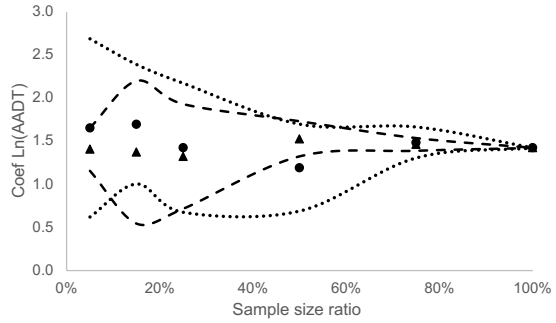
Aggregated Results from All Experiments

The 95% confidence intervals for estimated coefficients and dispersion from Base-SPFs and Augmented-SPFs for each sample size ratio is presented in Figure 6.6. It can be seen that, as expected, when sample size ratio increases the estimated model coefficients and dispersion converge to the corresponding values for True-SPF. In addition, the 95% CI of

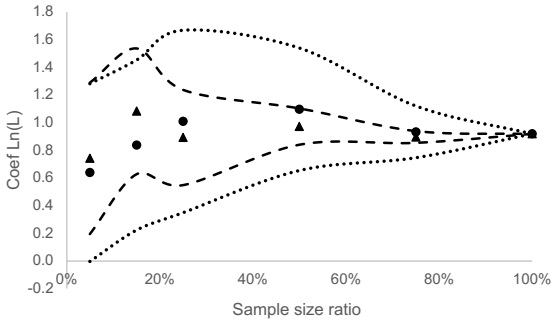
the Augmented-SPFs are usually narrower than the CI of the Base-SPFs which indicates that Augmented-SPF are better models in terms of better fit to the data.



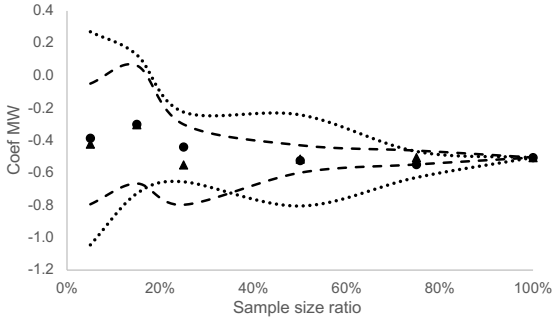
(a) Dispersion



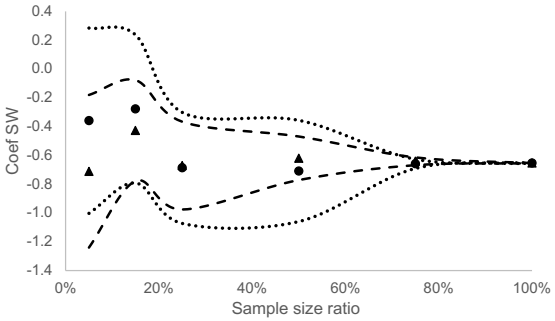
(b) Ln(AADT) coefficient



(c) Ln(segment length) coefficient



(d) Median width coefficient



(e) Shoulder width coefficient

Figure 6.6: 95% CI for estimated model coefficients and dispersion for Base-SPFs vs Augmented SPFs given different sample size ratio

Standard measures of SPF quality, such as Bayesian information criterion (BIC), Akaike information criterion (AIC), and deviance information criterion (DIC) cannot be used to compare the Base and Augmented models because the SPFs are developed on different data sets [131]. Consequently, the SPFs are applied to the test data set to estimate crash counts for each site. The estimates from the developed Base-SPFs and Augmented-SPFs are compared to estimates from the True-SPF in terms of the mean absolute error (MAE) and root mean square error (RMSE) which are calculated as follows:

$$MAE_k = \frac{1}{N}\sum_{i=1}^{N}|\mu_{i,k} - \mu_{i,true}| \tag{6.6}$$

$$RMSE_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mu_{i,k} - \mu_{i,true})^2} \tag{6.7}$$

where k represents either Base- or Augmented-SPF, $\mu_{i,k}$ is the estimated crash count for test site i by SPF k, $\mu_{i,true}$ is the estimated crash count for test site i by True-SPF, and N is the number of test samples.

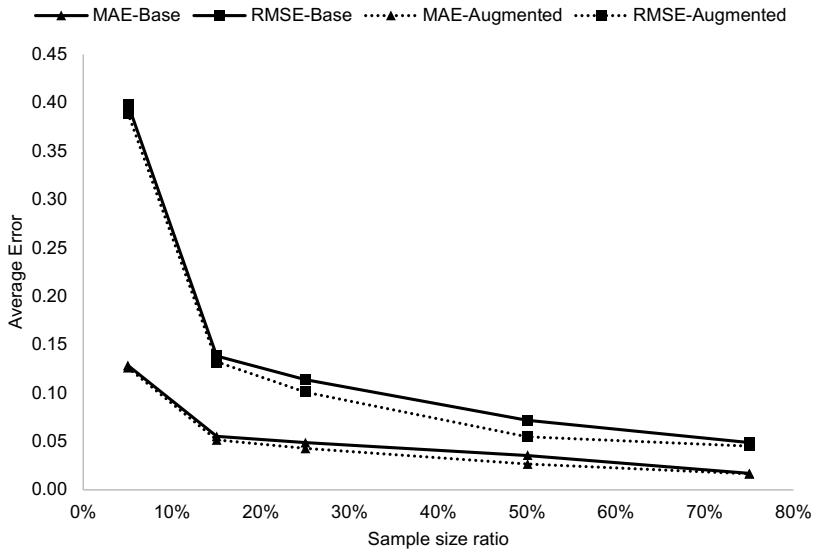


Figure 6.7: Mean absolute error (MAE) and root mean square error (RMSE) for Base-SPFs and Augmented-SPFs as a function of sample size ratio

Table 6.3: Percentage change in MAE and RMSE using data augmentation

	Sample size ratio				
Error Metric	5%	15%	25%	50%	75%
MAE	-1.7%	-6.9%	-12.2%	**-25.1**%	-2.0%
RMSE	-2.3%	-4.5%	-11.2%	**-23.3**%	-7.3%

Note: Bold values represent statistically significant difference (p-value > 0.05).

The MAE and RMSE results are presented in Figure 6.7 and Table 6.3. As expected, the magnitude of the errors reduces as sample size ratio increases as both Base and Augmented-SPFs converge to True-SPF. Of more interest is that the errors for Augmented-SPFs are lower than those for Base-SPFs across all sample size ratios, though the magnitude of improvement varies. The results show that when the sample size ratio is very small or very large, less improvement is obtained. An intuitive explanation is as follows. When the sample size is very small, we cannot train a proper CGAN model to improve by data augmentation or the small sample is not a good representative of the population. Conversely, when the sample size ratio is very large (close to the recommended sample size), reliable SPFs can be developed and therefore there is less opportunity for improvement through data augmentation. Consequently, data augmentation can be considered when there is enough data to train a CGAN but not enough to develop a reliable SPF. This can be decided based on the recommended sample size given the sample mean of the data set.

6.4.2 Simulated Crash Data

In the previous section, we have shown that the proposed data augmentation method can provide statistically significant improvements in the SPF fit and in model predictions. However, that analysis was done for a data set with specific characteristics, notably a fixed dispersion. In this section, we use a simulated environment to evaluate the performance of the proposed crash data augmentation method across a range of sample mean crash rates and dispersion parameter values. Furthermore, simulation enables us to perform the evaluation between Base-SPF and Augmented-SPF under various scenarios while also knowing the truth (i.e. true long-term crash mean). The simulated data sets are created using the following three steps, which is consistent with the approach taken by previous studies ([41, 42, 43, 25]):

1. Generate a random feature vector (X) with the size of 4 (X_1, X_2, X_3, X_4) from a uniform distribution on $[0, 1]$.

2. Generate the corresponding crash count Y from a Poison distribution with the long-term crash mean of λ that is gamma distributed with the dispersion of α:

$$Y \sim Poisson(\lambda);$$
$$\lambda = \exp(\beta_0 + 0.5X_1 - 0.5X_2 + X_3 - X_4 + \epsilon);$$
$$\exp(\epsilon) \sim Gamma(1, \alpha).$$

3. Steps (1) and (2) are repeated until the desired sample size is reached.

It is worth noting that ϵ indicates the unobserved heterogeneity that follows the log-gamma distribution, as specified in the NB model. The sample mean is determined by β_0. As done by other studies ([41, 42, 43, 25]), we set $\beta_0 = 0.5$ to represent a crash data set with a sample mean of 1.6 crashes/year. Following the approach from these same studies, we evaluate four values of the dispersion parameter ($\alpha = [0.5, 1.0, 1.5, 2.0]$) to represent crash data sets with low to high dispersion parameters. The sample size for these experiments is set to 100 which is about 6 times smaller than the minimum sample size recommended by [8] for a sample mean of 1.6. For each dispersion parameter the following steps are used to perform the experiments:

1. A *train data set* with the size of 100 is simulated following the data simulation steps given the dispersion parameter.

2. A CGAN is trained using the simulated *train data set*. The training configuration parameters are set same as previous section.

3. A *Base data set*, *Augmented data set*, and *Evaluation data set* are separately prepared as follows:

   - A *Base data set* with the size of 100 is simulated following the data simulation steps given the dispersion parameter, which is used to develop Base-SPF.
   - An *Augmented data set* with the size of 1000 is prepared by combining *base data set* and 900 data points generated using the CGAN trained in step 2, and the Augmented data set is used to develop Augmented-SPF.
   - An *Evaluation data set* with the size of 100 is simulated following the data simulation steps given the dispersion parameter, and is then used to compare the prediction performance of Base-SPF and Augmented-SPF.

4. The two sets of SPFs (i.e. Base-SPFs vs Augmented-SPFs) are compared in terms of the accuracy of dispersion parameter estimation, crash frequencies (over an unseen prediction test data), EB estimates, and hotspot identification performance.

5. Repeat steps 3 and 4 for 1000 times.

The results are summarized and discussed in the following sections.

Accuracy of Dispersion Estimation

The dispersion parameter for each SPF for each Base data set is determined using auxiliary Ordinary Least Square (OLS) regression without constant ([94]). The 95 % CI of MAE for dispersion estimation for the Base data and the Augmented data given different dispersion parameters are calculated and presented in Figure 6.8.

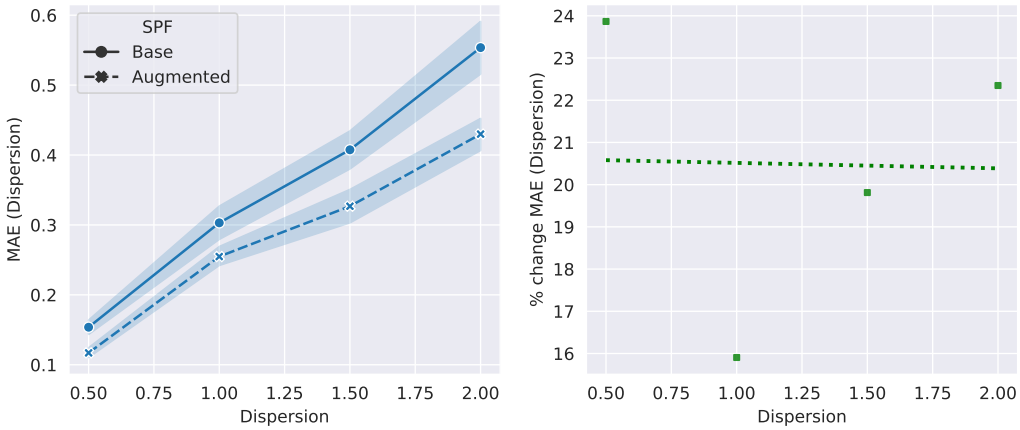


Figure 6.8: MAE (left) and percentage change MAE (right) for dispersion parameter estimates

As expected, the MAE of dispersion increases as the dispersion increases. When the magnitude of MAE is divided by the True dispersion values, then all errors were about 30 %. We can observe from the figure that the dispersion parameter estimation accuracy is better (lower MAE) for the Augmented SPFs than the Base SPFs across all dispersion parameter values. The improved accuracy was statistically significant for all dispersion values and the percentage improvement in dispersion parameter estimation accuracy provided by the Augmented SPF was approximately 20% across all dispersion values.

Accuracy of Crash Frequency Estimations

The prediction performance of Base- and Augmented-SPFs are compared in terms of mean absolute error (MAE) and root mean square error (RMSE) of the SPF predictions for crash frequencies of 1000 prediction test data points. The results for different dispersion parameters are presented in Figure 6.9.

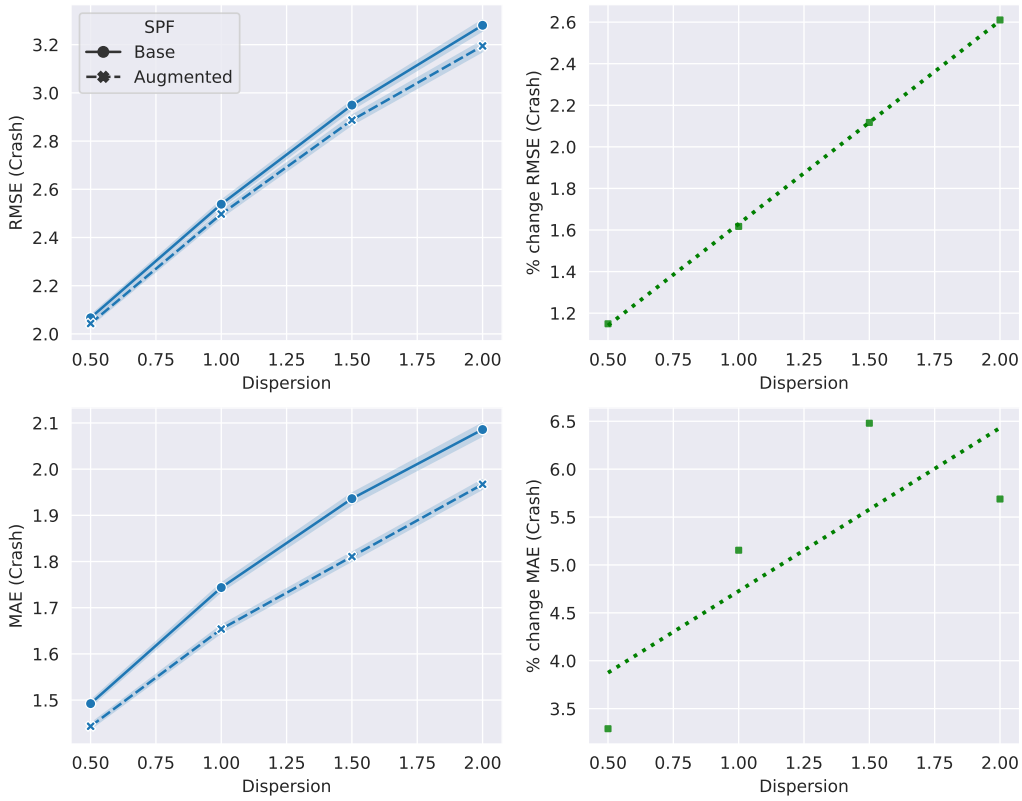


Figure 6.9: RMSE (top left), % change in RMSE (top right), MAE (bottom left) and % change MAE (bottom right) for crash frequency estimates

As expected, crash frequency prediction errors increase as the dispersion parameter value increases. Given that all data sets have the same crash mean, increased dispersion

causes the data sets to contain a small number of sites with a large number of crashes and a large number of sites with zero crash. As a result, the estimation error of the model over data sets with higher dispersion values is increased due to the inclusion of sites with substantially different crash counts than the mean.

But for all dispersion values, the Augmented-SPFs provide more accurate crash frequency predictions than Base-SPFs. The improvement in accuracy as a percent of the Base-SPF error ranges from 1% to 3% for RMSE and 3% to 6% for MAE. The results also indicate that the improvement in crash frequency prediction accuracy provided by data augmentation increases as the data becomes more highly dispersed.

Accuracy of EB Estimates

The prediction performance of Base- and Augmented-SPFs is also compared in terms of mean absolute error (MAE) and root mean square error (RMSE) of the EB estimates. According to the results presented in Figure 6.10, while there is an increasing trend in both MAE and RMSE, the rate of increase for both error metrics decreases as dispersion increases. In addition, the percentage difference between MAE and RMSE of Base- and Augmented-SPFs decreases for larger dispersion values. This can be due to the fact that in the EB method (Equation 6.8), as the dispersion increases, less weight is assigned to SPF predictions and more weight is assigned to observed counts. As a result, regardless of how much improvement in SPF is provided, the weight of SPF prediction in EB method is relatively low for large dispersion values.

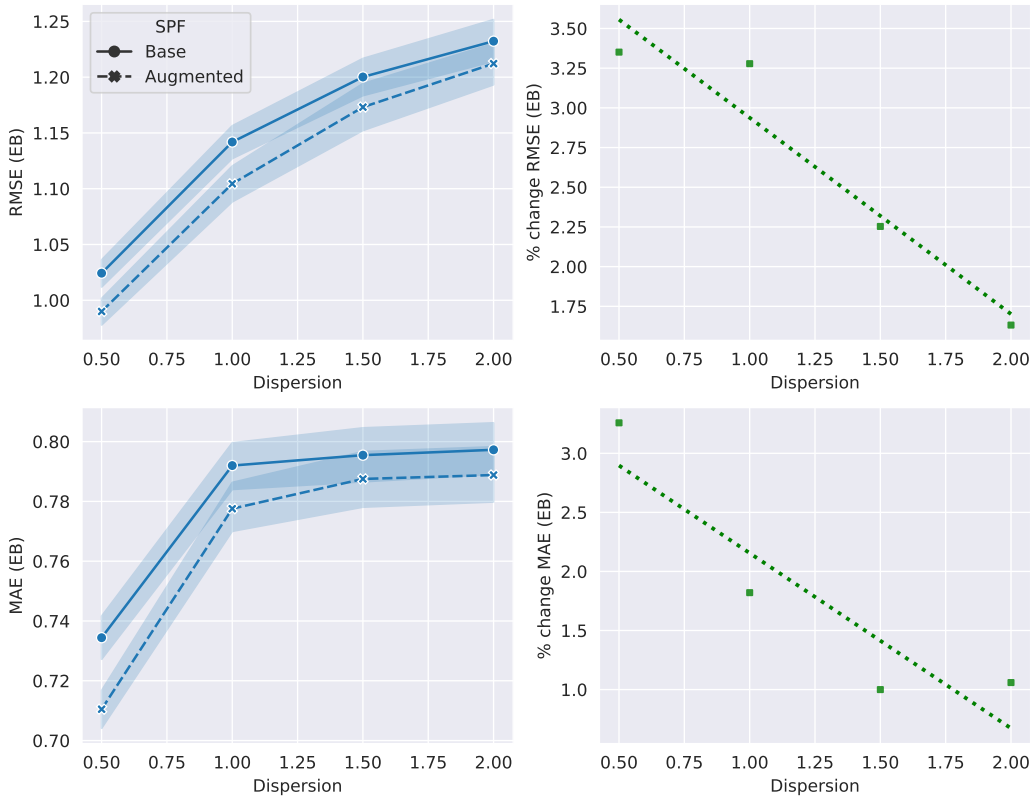


Figure 6.10: RMSE (top left), % change in RMSE (top right), MAE (bottom left) and % change MAE (bottom right) for EB estimates

Hotspot Identification Performance

For each test data set, the hotspot identification process is performed using both base-SPFs and augmented-SPFs. The process includes ranking sites based on the EB estimates (i.e. estimation of long-term crash mean λ) ([19]):

$$EB = \frac{1}{1+\alpha\mu} \times \mu + \frac{\alpha\mu}{1+\alpha\mu} \times y \tag{6.8}$$

where μ is the crash frequency prediction from either base-SPFs or augmented-SPFs. The hotspots suggested based on each SPF can be compared to the true hotspots (ranked based on λ) using False Identification (FI) and Poisson Mean Difference (PMD) tests ([87]). The FI test shows the percentage of sites that are erroneously categorised as hotspots, and the PMD test is the relative difference of the sum of the Poisson means (λ) for true hotspots and the suggested hotspots based on EB estimates. FI and PMD are computed based on a given number of top hotspots. In this study, we calculate FI and PMD for the top 5, 10, 15 and 20 hotspots and then calculate the average of the four values. The FI and PMD results as a function of dispersion parameter value are presented in Figure 6.11.

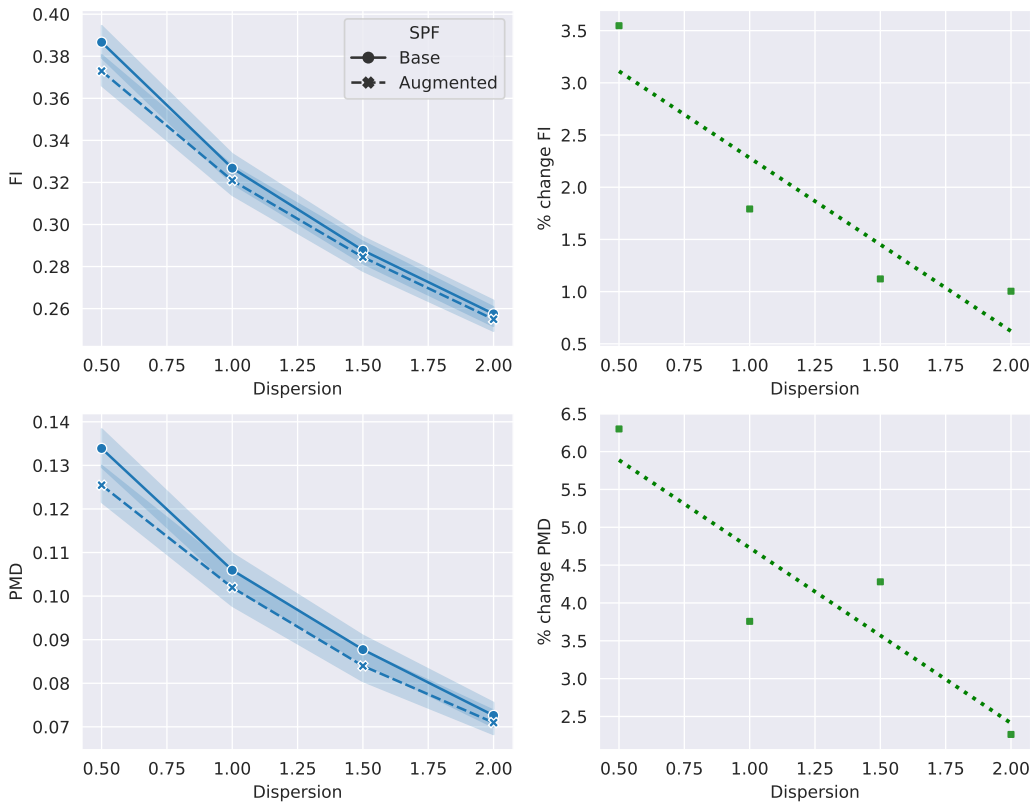


Figure 6.11: FI (left) and PMD (right) results

Unlike other metrics, which increase as dispersion increases, FI and PMD decrease as dispersion increases. Furthermore, while FI and PMD values for Augmented-SPFs are lower on average than Base-SPFs for all dispersion values, the difference becomes smaller as dispersion increases, with only the difference at dispersion = 0.5 being statistically significant (p-value 0.05). This is due, once again, to the fact that when dispersion is high, most hotspots have higher crash counts with larger weight than prediction (E1. 6.8), making the impact of SPF prediction less important.

All the results confirmed that the Augmented-SPFs outperform the Base-SPFs in almost all performance measures. In addition, it was found that the magnitude of dispersion parameter appears to have an important impact on the benefit of data augmentation. When the dispersion is high, data augmentation provides more improvements in terms of crash frequency predictions, and when the dispersion is low, the improvements are mostly in terms of the accuracy of EB estimates and accuracy of hotspot identification.

6.5 Limitations and Future Studies

The analysis presented in this chapter showed that crash data augmentation using a CGAN model can provide statistically significant improvements for different scenarios in terms of model fitting and network screening outcomes. Although the results are encouraging, there are some limitations in the current work that should be addressed in future studies:

- Our analysis using the WA data showed that data augmentation can be effective when the sample size is between 25% and 75% of the recommended sample size for the sample mean. However, this optimal range could be a function of the sample mean. In particular, there is likely some minimum number of sites required to train a function CGAN. For example, the recommended sample size for a sample mean of 5 is 200 ([8]). Our current analysis suggests that the proposed CGAN data augmentation method can provide benefits for a sample ratio of larger than 25% but even assuming 30% would provide only 60 sites of observed data which may not be sufficient for developing a functional CGAN. Further studies are required to examine the minimum sample size necessary to train a CGAN for a range of crash data characteristics.

- When the dispersion parameter is large, data augmentation is less beneficial in terms of network screening outcomes. This result is based on the fact that less weight is placed on the SPF prediction in the EB method for higher dispersion values, leaving less potential for improvement through data augmentation. Other network screening

methods, such as full Bayes, may not provide the the same result, necessitating more investigation.

- The portion of this chapter using empirical data examined performance of CGAN data augmentation for a single network site type, namely divided 4-lane segments from urban freeways. It is recommended to confirm that the proposed CGAN data augmentation method provides beneficial performance when applied to other types of sites (intersections, roundabouts, etc.). In addition, it should be investigated if one single CGAN can model the distribution of crash data sets for multiple different network facility types.

- To identify the optimum structures suited for crash data sets with different size and characteristics, it is important to explore various architecture and loss functions for CGAN.

- Finally, the performance of CGAN should be compared to that of other generative models, such as variational auto-encoders (VAE) and Gaussian mixture models (GMM), to determine which is best suited for crash data augmentation tasks.

6.6 Conclusions and Recommendations

SPFs are commonly used in road network safety screening to identify and prioritize crash hotspots and to assess the efficiency of safety countermeasures. One of the challenges for developing reliable SPFs is dealing with crash data sets with inadequate sample sizes. Traditionally, options for addressing this problem include expanding the data set by using more years of crash counts for each site and/or by combining site types and/or combining crash severity types. However, each of these approaches has their own limitations.

In this chapter, a data augmentation method for crash frequency data based on conditional generative adversarial networks (CGAN) is proposed, which can mimic the underlying distribution of the given crash data and generate synthesized samples. Using the generated samples, the size of the original data set can be increased to the minimum recommended sample size for SPF development purposes ([8]). The method is evaluated using a real-world crash data set for 3085 urban freeway four-lane road segments in Washington State USA. The results showed that data augmentation could improve the results in terms of reducing coefficient standard error and improving the accuracy of SPF predictions. Also, based on results from simulated crash data sets, data augmentation provides greater improvement in terms of the accuracy of network screening outcomes for lower dispersion

values (i.e. less than 1). In this regard, False Identification (FI) and Poison Mean Difference (PMD) tests for Augmented-SPFs on average were improved by approximately 1-3% and 2-6% respectively in comparing to the results for Base SPFs. Also, Augmented SPFs showed lower mean absolute error (MAE) and root mean square error (RMSE) for EB estimates (1-3%), and crash frequency predictions (1-3% for RMSE and 3-6% for MAE). Finally, MAE for dispersion parameter estimation was about 20% lower on average for Augmented-SPFs.

Chapter 7

Conclusions

In road safety studies, such as in network screening, the most popular approach is of using crash predictive models (i.e. SPFs) based on generalized linear model and empirical Bayes method to estimate the long-term safety (i.e. expected crash frequency) of locations. As stated in the first chapter of this thesis, there are several challenges in this process, including a lack of objective tools to determine how frequently SPFs need to be redeveloped, how sensitive the process is to input data accuracy (e.g. AADT), limitations of generalised linear models, and the low sample size issue. The primary objectives of this thesis are therefore to address these challenges through developing cost-benefit analysis tools, non-parametric EB estimation method, and a crash frequency data augmentation method. This chapter highlights the main contributions of this thesis followed by direction for future research.

7.1 Contributions

The four contributions of this thesis are as follows:

1. **Developed a benefit-cost analysis method to investigate the sensitivity of NS results to AADT accuracy**
   Conventional road safety network screening relies on measures of historical crash data and annual average daily traffic (AADT) as a measure of exposure to develop safety performance functions. AADT is typically estimated from short-term counts and contains error which can negatively impact network screening outcomes and result in the inefficient allocation of safety improvement resources. In this thesis, we developed and proposed a simulation-based method for quantifying the monetary benefit of improving AADT accuracy. The results of applying the method under various conditions show that crash data sets with higher sample mean and dispersion parameter values are more sensitive to AADT error and consequently benefit more from improving AADT accuracy. The use of the proposed method is illustrated through the application to a real-world example, which can help jurisdictions quantify the benefits of investing in methods to increase AADT accuracy.

2. **Developed a benefit-cost analysis method for objectively deciding whether to redevelop or recalibrate SPFs**
   This research proposed a method by which jurisdictions can determine when it is economically beneficial to redevelop SPFs instead of using recalibrated old SPFs. To this end, extensive annual SPF development and network screening are implemented on three data sets from three regions. It is assumed that re-development of SPFs

using the most recently available data for the local jurisdiction provides the most accurate NS results. In order to quantify the consequence of using outdated SPFs instead of redeveloped SPFs, a benefit metric based societal costs of crash types are proposed. Then, three trend measures (based on change in AADT, crash count, time passed from last redevelopment) are introduced that are readily available for jurisdictions and can be used to estimate benefit before actually redeveloping SPFs. The data required for this method is either already available in municipalities which have locally developed SPFs or can be easily estimated. The estimation models for the method are created based on real data sets and were successfully tested on validation data sets. The method is compared with three benchmarks (i.e. always redevelop, always recalibrate, redevelop after 4 years), and outperformed the best benchmark by 13% in terms of accuracy of correct decisions.

3. **Developed a non-parametric EB estimation method**
   In this research a novel non-parametric EB method for modeling crash frequency data data based on Conditional Generative Adversarial Networks (CGAN) is proposed and evaluated over a real-world crash data set. Unlike parametric approaches, there is no need for a pre-specified underlying relationship between dependent and independent variables in the proposed CGAN-EB and it is able to model any type of distribution. The proposed methodology is applied to real-world and simulated crash data sets. The performance of the proposed CGAN-EB method in terms of model fit, predictive performance and network screening outcomes is compared with the conventional approach (NB-EB) as a benchmark. The results indicate that the proposed CGAN-EB approach outperforms NB-EB in terms of prediction power and hotspot identification tests. Also, a series of simulation experiments are devised and carried out to assess the CGAN-EB performance across a wide range of conditions and compares it to the NB-EB. The simulation results show that CGAN-EB performs as well as NB-EB when conditions favor the NB-EB model (i.e. data conform to the assumptions of the NB model) and outperforms NB-EB in experiments reflecting conditions frequently encountered in practice (i.e. low sample mean crash rates, and when crash frequency does not follow a log-linear relationship with covariates). Also, temporal and spatial transferability of both approaches were evaluated using field data and both CGAN-EB and NB-EB approaches were found to have similar performance.

4. **Developed and evaluated a novel crash frequency data augmentation method**
   In road safety analysis, crash frequency models (also called safety performance functions or SPFs) are developed for specific crash and site types based on historical

crash data and site characteristics. When samples sizes are small, it is frequently not possible to develop statistically reliable SPFs. In this research, we developed and evaluated a crash frequency data augmentation method using Conditional Generative Adversarial Networks (CGANs) to address this problem. The proposed method is evaluated by comparing the performance of Base-SPF (SPFs developed using original data) and Augmented-SPF (SPFs developed using original data plus synthesized data) in terms of hotspot identification performance (i.e. False Identification (FI) and Poisson Mean Difference (PMD) tests), accuracy of estimated long-term crash means, crash frequencies, and dispersion parameters. The experiments are conducted using both real-world and simulated crash data sets. The results from real data experiments indicate that the crash frequency estimation accuracy for the test data set has improved by up to 25% when using Augmented-SPFs. Also, the standard error associated with SPF coefficients were lower for Augmented-SPFs. In addition, the simulation results show that, the Augmented-SPF can improve Base-SPF in terms of FI (up to 3%), PMD (up to 6%), MAE/RMSE of empirical Bayes estimates (up to 3%), and MAE/RMSE of dispersion parameter estimation (up to 20%).

7.2 Publications

In this section the list of peer-reviewed journal papers and conference articles emanated from this PhD research is given.

7.2.1 Journal papers

- **Zarei, M.**, Hellinga, B., & Izadpanah, P. (2022). A Benefit-Cost Based Method to Determine When Safety Performance Functions (SPFs) Should be Redeveloped for Use in Intersection Network Screening. *Transportation Research Record*, 2676(11), 239–249.
- **Zarei, M.**, & Hellinga, B. (2022). Method for Estimating the Monetary Benefit of Improving Annual Average Daily Traffic Accuracy in the Context of Road Safety Network Screening. *Transportation Research Record*.
- **Zarei, M.**, Hellinga, B., & Izadpanah, P. (2022). CGAN-EB: A Non-parametric Empirical Bayes Method for Crash Frequency Modeling Using Conditional Generative Adversarial Networks as Safety Performance Functions. *International Journal of Transportation Science and Technology*, ISSN 2046-0430.

- **Zarei, M.**, Hellinga, B., & Izadpanah, P. (2023). Application of Conditional Deep Generative Networks (CGAN) in Empirical Bayes Estimation of Road Crash Risk and Identifying Crash Hotspots *International Journal of Transportation Science and Technology*, ISSN 2046-0430.

7.2.2 Conference presentations

- **Zarei, M.**, & Hellinga, B. (2022). A quantitative method to determine when safety performance functions used for network screening should be redeveloped. *In Proceedings of the Transportation Research Board (TRB) annual meeting.*

7.3 Recommendations for Future Work

The following potential topics are suggested for the extension and continuation of the research presented in this PhD dissertation:

- **Investigating error characteristics in AADT data for the benefit cost AADT sensitivity simulation method and extend it to intersections**. In the study, it was assumed that the coefficient of variation (COV) of AADT errors is constant for all sites and that the errors are normally distributed, but further research should be done to confirm the validity of these assumptions. Additionally, if it is found that the COV of AADT errors varies as a function of the AADT or follows a different distribution, the proposed simulation method should still be used by modifying the relevant equation. Another area for future research is the application of the proposed method to intersections, where the AADT of both major and minor roads are input for model development. In this context, there may be error correlations that should be considered. Finally, the proposed method monetizes the cost of inaccurate AADT values based on the average cost of crashes and the change in Potential Safety Improvement (PSI), but does not account for variations in countermeasure costs and effectiveness. Future research should explore ways to incorporate these factors into the quantification of the value of improved AADT data accuracy.

- **Extending the benefit-cost model for SPF redevelopment to larger data set and different location types**. One potential avenue for future research is to further develop and refine the method presented in this thesis for determining when SPF redevelopment is warranted. This could involve collecting more data and testing

the method on a wider range of data sets to ensure its robustness and investigate its generalizability to other location types (e.g. road segments). Investigating the model's generalizability to other location types, such as road segments, could help to extend its applicability beyond the specific contexts in which it has been developed and tested. This could involve adapting the model to account for different road geometries and traffic characteristics and testing its performance on data sets from a variety of location types.

- **Developing one single large global CGAN as a crash predictive model that can be used for different regions and all locations types**. Developing a single large global Conditional Generative Adversarial Network (CGAN) as a crash predictive model that can be used for different regions and all locations types has the potential to greatly improve the accuracy and efficiency of crash prediction efforts. By leveraging a large, diverse dataset, such a model could capture a wide range of factors that may influence crash rates and accurately predict crashes across a variety of regions and locations types. Additionally, a global CGAN model would be more convenient to use than having to develop separate models for each region or location type, potentially saving time and resources. However, creating a global CGAN model would likely require a significant amount of data and computational resources, and there may be challenges in ensuring the model is sufficiently generalizable to different regions and locations. Further research should explore the feasibility and potential benefits of developing such a model. As a future study, it would also be interesting to investigate the application of the proposed CGAN method in observational before-after studies. This would involve using the CGAN model to predict crash rates before and after the implementation of specific road safety interventions, and comparing the results with actual observed crash rates.

- **Examining the relationship between sample size and improvement via crash data augmentation, as well as their local transferability**. In this work, we applied crash data augmentation to a few crash data sets with limited ranges of sample means, namely 0.29 for real-world crash data and 1.6 for simulated crash data. It is found that the effectiveness of data augmentation varied depending on the sample mean. This is because data sets with lower sample means have a higher recommended sample size, which means there are more samples available for training a better crash data augmentation model than in data sets with higher sample means with a similar sample ratio. In other words, when the sample size is very small, it is not possible to train an adequate augmentation model. One solution to this issue might be using transfer learning and fine-tuning an already existing crash

data augmentation model which is trained over a large, diverse dataset instead of retraining it from scratch with a small data. More research might be conducted to investigate the relationship between sample size versus performance improvement, and the feasibility of locally transferring crash data augmentation models .

References

[1] WHO. *Global status report on road safety 2018*. World Health Organization, 2018.

[2] Transport Canada. *Canadian Motor Vehicle Traffic Collision Statistics: 2018*. Technical report, Transport Canada, 2018.

[3] AASHTO. *Highway Safety Manual*, volume 1. American Association of State Highway Transportation Professionals, Washington, D.C., 2010.

[4] Dominique Lord and Simon Washington. *Safe mobility: Challenges, methodology and solutions*. Emerald Publishing, 2018.

[5] Wen Cheng and Simon Washington. New criteria for evaluating methods of identifying hot spots. *Transportation Research Record*, 2083(1):76–85, 2008.

[6] Alfonso Montella. A comparative analysis of hotspot identification methods. *Accident Analysis & Prevention*, 42(2):571–581, 2010.

[7] Simon Washington, Md Mazharul Haque, Jutaek Oh, and Dongmin Lee. Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots. *Accident Analysis & Prevention*, 66:136–146, 2014.

[8] Dominique Lord. Modeling motor vehicle crashes using poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, 38(4):751–766, 2006.

[9] Raghavan Srinivasan, Daniel Carter, and Karin Bauer. Safety Performance Function decision guide: SPF calibration vs SPF development. *Federal Highway Administration–Office of Safety Report*, 2013.

[10] Helai Huang, Hoong Chor Chin, and Md Mazharul Haque. Empirical evaluation of alternative approaches in identifying crash hot spots: Naive ranking, empirical bayes, full bayes methods. *Transportation Research Record*, 2103(1):32–41, 2009.

[11] Craig Milligan, Jeannette Montufar, Jonathan Regehr, and Bartholomew Ghanney. Road safety performance measures and aadt uncertainty from short-term counts. *Accident Analysis & Prevention*, 97:186–196, 2016.

[12] YooSeok Jung and JuSam Oh. Optimization of short-term traffic count plan to improve aadt estimation error. *Optimization*, 13(10):71–79, 2017.

[13] Karim El-Basyouny and Tarek Sayed. Safety performance functions with measurement errors in traffic volume. *Safety science*, 48(10):1339–1344, 2010.

[14] Dominique Lord and Fred Mannering. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation research part A: policy and practice*, 44(5):291–305, 2010.

[15] Guangyuan Pan, Liping Fu, and Lalita Thakali. Development of a global road safety performance function using deep neural networks. *International journal of transportation science and technology*, 6(3):159–173, 2017.

[16] Chunjiao Dong, Chunfu Shao, Juan Li, and Zhihua Xiong. An improved deep learning model for traffic crash prediction. *Journal of Advanced Transportation*, 2018, 2018.

[17] Zhuojun Jiang, Mark R McCord, and Prem K Goel. Improved aadt estimation by combining information in image-and ground-based traffic data. *Journal of transportation engineering*, 132(7):523–530, 2006.

[18] Satish Sharma, Pawan Lingras, Fei Xu, and Peter Kilburn. Application of neural networks to estimate aadt on low-volume roads. *Journal of Transportation Engineering*, 127(5):426–432, 2001.

[19] Ezra Hauer. *Observational before/after studies in road safety. estimating the effect of highway and traffic engineering measures on road safety.* 1997.

[20] Ali Khodadadi, Ioannis Tsapakis, Subasish Das, Dominique Lord, and Yingfeng Li. Application of different negative binomial parameterizations to develop safety performance functions for non-federal aid system roads. *Accident Analysis & Prevention*, 156:106103, 2021.

[21] Mehdi Hosseinpour, Sina Sahebi, Zamira Hasanah Zamzuri, Ahmad Shukri Yahaya, and Noriszura Ismail. Predicting crash frequency for multi-vehicle collision types using multivariate poisson-lognormal spatial model: A comparative analysis. *Accident Analysis & Prevention*, 118:277–288, 2018.

[22] Dominique Lord, Srinivas Reddy Geedipally, and Seth D Guikema. Extension of the application of conway-maxwell-poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis: An International Journal*, 30(8):1268–1276, 2010.

[23] Dibakar Saha, Priyanka Alluri, Eric Dumbaugh, and Albert Gan. Application of the poisson-tweedie distribution in analyzing crash frequency data. *Accident Analysis & Prevention*, 137:105456, 2020.

[24] Yajie Zou, Dominique Lord, Yunlong Zhang, and Yichuan Peng. Comparison of sichel and negative binomial models in estimating empirical bayes estimates. *Transportation research record*, 2392(1):11–21, 2013.

[25] Mohammad Zarei, Bruce Hellinga, and Pedram Izadpanah. CGAN-EB: A non-parametric empirical bayes method for crash frequency modeling using conditional generative adversarial networks as safety performance functions (accepted). *International journal of transportation science and technology*, 2022.

[26] Xiugang Li, Dominique Lord, Yunlong Zhang, and Yuanchang Xie. Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention*, 40(4):1611–1618, 2008.

[27] Traffic monitoring guide. *Federal Highway Administration, US Department of Transportation, October 2016*, 2016.

[28] Ioannis Tsapakis and William H Schneider. Use of support vector machines to assign short-term counts to seasonal adjustment factor groups. *Transportation Research Record*, 2527(1):8–17, 2015.

[29] Brent Selby and Kara M Kockelman. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*, 29:24–32, 2013.

[30] Yongze Song, Xiangyu Wang, Graeme Wright, Dominique Thatcher, Peng Wu, and Pascal Felix. Traffic volume prediction with segment-based regression kriging and

its implementation in assessing the impact of heavy vehicles. *Ieee transactions on intelligent transportation systems*, 20(1):232–243, 2018.

[31] William HK Lam and Jianmin Xu. Estimation of aadt from short period counts in hong kong—a comparison between neural network method and regression analysis. *Journal of Advanced Transportation*, 34(2):249–268, 2000.

[32] Mostafa H Tawfeek and Karim El-Basyouny. Estimating traffic volume on minor roads at rural stop-controlled intersections using deep learning. *Transportation research record*, 2673(4):108–116, 2019.

[33] Xu Zhang and Mei Chen. Enhancing statewide annual average daily traffic estimation with ubiquitous probe vehicle data. *Transportation Research Record*, 2674(9):649–660, 2020.

[34] Hyun-ho Chang and Seung-hoon Cheon. The potential use of big vehicle gps data for estimations of annual average daily traffic for unmeasured road segments. *Transportation*, 46(3):1011–1032, 2019.

[35] Shawn Turner, Pete Koeneman, et al. Using mobile device samples to estimate traffic volumes. Technical report, Minnesota. Dept. of Transportation. Research Services & Library, 2017.

[36] Robert Krile, Jeremy Schroeder, and Steven Jessberger. Assessing roadway traffic count duration and frequency impacts on annual average daily traffic estimation: Assessing accuracy issues related to short-term count durations. Technical report, United States. Federal Highway Administration, 2016.

[37] Craig Milligan, Jeannette Montufar, Jonathan Regehr, and Bartholomew Ghanney. Road safety performance measures and aadt uncertainty from short-term counts. *Accident Analysis & Prevention*, 97:186–196, 2016.

[38] Michael J Maher and Ian Summersgill. A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis & Prevention*, 28(3):281–296, 1996.

[39] Karim El-Basyouny and Tarek Sayed. Safety performance functions with measurement errors in traffic volume. *Safety science*, 48(10):1339–1344, 2010.

[40] Anusha Musunuru and Richard J Porter. Applications of measurement error correction approaches in statistical road safety modeling. *Transportation research record*, 2673(8):125–135, 2019.

[41] Royce A Francis, Srinivas Reddy Geedipally, Seth D Guikema, Soma Sekhar Dhavala, Dominique Lord, and Sarah LaRocca. Characterizing the performance of the conway-maxwell poisson generalized linear model. *Risk Analysis: An International Journal*, 32(1):167–183, 2012.

[42] Yajie Zou, Lingtao Wu, and Dominique Lord. Modeling over-dispersed crash data with a long tail: examining the accuracy of the dispersion parameter in negative binomial models. *Analytic Methods in Accident Research*, 5:1–16, 2015.

[43] Xin Ye, Ke Wang, Yajie Zou, and Dominique Lord. A semi-nonparametric poisson regression model for analyzing motor vehicle crash data. *PloS one*, 13(5):e0197338, 2018.

[44] Bhagwant Persaud, Craig Lyon, and Thu Nguyen. Empirical bayes procedure for ranking sites for safety investigation by potential for safety improvement. *Transportation research record*, 1665(1):7–12, 1999.

[45] Mohammad Zarei, Bruce Hellinga, and Pedram Izadpanah. A quantitative method to determine when safety performance functions used for network screening should be redeveloped. In *Proceedings of the Transportation Research Board (TRB) annual meeting*, 2022.

[46] Paul De Leur, Laura Thue, and Brian Ladd. Collision cost study update final report. Technical report, 2019.

[47] Forrest M Council, Eduard Zaloshnja, Ted Miller, Bhagwant Naraine Persaud, et al. Crash cost estimates by maximum police-reported injury severity within selected crash geometrics. 2005.

[48] Ezra Hauer. *The art of regression modeling in road safety*, volume 38. Springer, 2015.

[49] G Bahar and Ezra Hauer. User's guide to develop highway safety manual safety performance function calibration factors. *National Cooperative Highway Research Program*, 2014.

[50] Raghavan Srinivasan and Karin Bauer. Safety Performance Function decision guide: Developing jurisdiction-specific SPFs. *Federal Highway Administration–Office of Safety Report*, 2013.

[51] Ziad Sawalha and Tarek Sayed. Transferability of accident prediction models. *Safety science*, 44(3):209–219, 2006.

[52] Raghavan Srinivasan, Michael Colety, Geni Bahar, Brent Crowther, and Matt Farmen. Estimation of calibration functions for predicting crashes on rural two-lane roads in arizona. *Transportation research record*, 2583(1):17–24, 2016.

[53] Ahmed Farid, Mohamed Abdel-Aty, and Jaeyoung Lee. A new approach for calibrating safety performance functions. *Accident Analysis & Prevention*, 119:188–194, 2018.

[54] Mohammadali Shirazi, Srinivas Reddy Geedipally, and Dominique Lord. A procedure to determine when safety performance functions should be recalibrated. *Journal of Transportation Safety & Security*, 9(4):457–469, 2017.

[55] Ezra Hauer, Jake Kononov, Bryan Allery, and Michael S Griffith. Screening the road network for sites with promise. *Transportation Research Record*, 1784(1):27–32, 2002.

[56] Jiří Ambros, Veronika Valentová, and Jiří Sedoník. Developing updatable crash prediction model for network screening: case study of czech two-lane rural road segments. *Transportation research record*, 2583(1):1–7, 2016.

[57] Yajie Zou, John E Ash, Byung-Jung Park, Dominique Lord, and Lingtao Wu. Empirical bayes estimates of finite mixture of negative binomial regression models and its application to highway safety. *Journal of Applied Statistics*, 45(9):1652–1669, 2018.

[58] Xin Ye, Ke Wang, Yajie Zou, and Dominique Lord. A semi-nonparametric poisson regression model for analyzing motor vehicle crash data. *PloS one*, 13(5):e0197338, 2018.

[59] Dominique Lord, Simon P Washington, and John N Ivan. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, 37(1):35–46, 2005.

[60] Di Yang, Kun Xie, Kaan Ozbay, Zifeng Zhao, and Hong Yang. Copula-based joint modeling of crash count and conflict risk measures with accommodation of mixed count-continuous margins. *Analytic Methods in Accident Research*, 31:100162, 2021.

[61] Andy H Lee, Mark R Stevenson, Kui Wang, and Kelvin KW Yau. Modeling young driver motor vehicle crashes: data with extra zeros. *Accident Analysis & Prevention*, 34(4):515–521, 2002.

[62] Karin M Bauer and Douglas W Harwood. Safety effects of horizontal curve and grade combinations on rural two-lane highways. *Transportation research record*, 2398(1):37–49, 2013.

[63] Jake Kononov, Craig Lyon, and Bryan K Allery. Relation of flow, speed, and density of urban freeways to functional form of a safety performance function. *Transportation research record*, 2236(1):11–19, 2011.

[64] Chunjiao Dong, Chunfu Shao, Juan Li, and Zhihua Xiong. An improved deep learning model for traffic crash prediction. *Journal of Advanced Transportation*, 2018, 2018.

[65] Qiang Zeng, Helai Huang, Xin Pei, and SC Wong. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic methods in accident research*, 10:12–25, 2016.

[66] Helai Huang, Qiang Zeng, Xin Pei, SC Wong, and Pengpeng Xu. Predicting crash frequency using an optimised radial basis function neural network model. *Transportmetrica A: transport science*, 12(4):330–345, 2016.

[67] Guangyuan Pan, Liping Fu, and Lalita Thakali. Development of a global road safety performance function using deep neural networks. *International journal of transportation science and technology*, 6(3):159–173, 2017.

[68] Gyanendra Singh, Mahesh Pal, Yogender Yadav, and Tushar Singla. Deep neural network-based predictive modeling of road accidents. *Neural Computing and Applications*, pages 1–10, 2020.

[69] CP Obite, NP Olewuezi, GU Ugwuanyim, and DC Bartholomew. Multicollinearity effect in regression analysis: A feed forward artificial neural network approach. *Asian Journal of Probability and Statistics*, pages 22–33, 2020.

[70] Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. *arXiv preprint arXiv:1911.12116*, 2019.

[71] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[72] Qili Chen, Guangyuan Pan, Wenbai Chena, and Peiliang Wu. A novel explainable deep belief network framework and its application for feature importance analysis. *IEEE Sensors Journal*, 2021.

[73] Athanasios Theofilatos, Cong Chen, and Constantinos Antoniou. Comparing machine learning and deep learning methods for real-time crash prediction. *Transportation research record*, 2673(8):169–178, 2019.

[74] Pei Li, Mohamed Abdel-Aty, and Jinghui Yuan. Real-time crash risk prediction on arterials based on lstm-cnn. *Accident Analysis & Prevention*, 135:105371, 2020.

[75] Shile Zhang, Mohamed Abdel-Aty, Yina Wu, and Ou Zheng. Modeling pedestrians' near-accident events at signalized intersections using gated recurrent unit (gru). *Accident Analysis & Prevention*, 148:105844, 2020.

[76] Mahdi Rezapour, Sahima Nazneen, and Khaled Ksaibati. Application of deep learning techniques in predicting motorcycle crash severity. *Engineering Reports*, 2(7):e12175, 2020.

[77] Ming Zheng, Tong Li, Rui Zhu, Jing Chen, Zifei Ma, Mingjing Tang, Zhongqiang Cui, and Zhan Wang. Traffic accident's severity prediction: A deep-learning approach-based cnn network. *IEEE Access*, 7:39897–39910, 2019.

[78] Zubayer Islam, Mohamed Abdel-Aty, Qing Cai, and Jinghui Yuan. Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention*, 151:105950, 2021.

[79] Qing Cai, Mohamed Abdel-Aty, Jinghui Yuan, Jaeyoung Lee, and Yina Wu. Real-time crash prediction on expressways using deep generative models. *Transportation research part C: emerging technologies*, 117:102697, 2020.

[80] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[81] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[82] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[83] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[84] Karan Aggarwal, Matthieu Kirchmeyer, Pranjul Yadav, S Sathiya Keerthi, and Patrick Gallinari. Conditional generative adversarial networks for regression. *ArXiv190512868 Cs Stat.(10)*, 2019.

[85] Yuxuan Zhang, Senzhang Wang, Bing Chen, Jiannong Cao, and Zhiqiu Huang. Trafficgan: Network-scale deep traffic prediction with generative adversarial nets. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[86] Yuanyuan Chen, Yisheng Lv, and Fei-Yue Wang. Traffic flow imputation using parallel data and generative adversarial networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(4):1624–1630, 2019.

[87] Wen Cheng and Simon Washington. New criteria for evaluating methods of identifying hot spots. *Transportation Research Record*, 2083(1):76–85, 2008.

[88] Wen Cheng and Simon P Washington. Experimental evaluation of hotspot identification methods. *Accident Analysis & Prevention*, 37(5):870–881, 2005.

[89] Xiaobo Qu and Qiang Meng. A note on hotspot identification for urban expressways. *Safety Science*, 66:87–91, 2014.

[90] Dominique Lord, Xiao Qin, and Srinivas Geedipally. *Highway safety analytics and modeling*. Elsevier, 2021.

[91] Ximiao Jiang, Mohamed Abdel-Aty, and Samer Alamili. Application of poisson random effect models for highway network screening. *Accident Analysis & Prevention*, 63:74–82, 2014.

[92] Yajie Zou, Dominique Lord, Yunlong Zhang, and Yichuan Peng. Comparison of sichel and negative binomial models in estimating empirical bayes estimates. *Transportation research record*, 2392(1):11–21, 2013.

[93] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[94] A Colin Cameron and Pravin K Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of econometrics*, 46(3):347–364, 1990.

[95] François Chollet et al. Keras: The python deep learning library. *ascl*, pages ascl–1806, 2018.

[96] Sagar Sharma. Activation functions in neural networks. *Towards Data Science*, 6, 2017.

[97] Dabal Pedamonti. Comparison of non-linear activation functions for deep neural networks on mnist classification task. *arXiv preprint arXiv:1804.02763*, 2018.

[98] William Young, Amir Sobhani, Michael G Lenné, and Majid Sarvi. Simulation of safety: A review of the state of the art in road safety simulation modelling. *Accident Analysis & Prevention*, 66:89–103, 2014.

[99] Byung-Jung Park, Dominique Lord, and Chungwon Lee. Finite mixture modeling for vehicle crash data with application to hotspot identification. *Accident Analysis & Prevention*, 71:319–326, 2014.

[100] Hao Yu, Pan Liu, Jun Chen, and Hao Wang. Comparative analysis of the spatial analysis methods for hotspot identification. *Accident Analysis & Prevention*, 66:80–88, 2014.

[101] AK Sharma and VS Landge. Zero inflated negative binomial for modeling heavy vehicle crash rate on indian rural highway. *International Journal of Advances in Engineering & Technology*, 5(2):292, 2013.

[102] Bhagwant Persaud, Bo Lan, Craig Lyon, and Ravi Bhim. Comparison of empirical bayes and full bayes approaches for before–after road safety evaluations. *Accident Analysis & Prevention*, 42(1):38–43, 2010.

[103] James A Bonneson, Srinivas Geedipally, Michael P Pratt, and Dominique Lord. Safety prediction methodology and analysis tool for freeways and interchanges. Technical report, 2021.

[104] Michael P Pratt, Srinivas R Geedipally, Bryan Wilson, Subasish Das, Marcus Brewer, and Dominique Lord. Pavement safety-based guidelines for horizontal curve safety. Technical report, 2018.

[105] Zubayer Islam, Mohamed Abdel-Aty, Qing Cai, and Jinghui Yuan. Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention*, 151:105950, 2021.

[106] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

[107] Hongliang Ding, Yuhuan Lu, NN Sze, Tiantian Chen, Yanyong Guo, and Qinghai Lin. A deep generative approach for crash frequency model with heterogeneous imbalanced data. *Analytic methods in accident research*, 34:100212, 2022.

[108] Mohammad Zarei, Bruce Hellinga, and Pedram Izadpanah. Benefit–cost-based method to determine when safety performance functions should be redeveloped for use in intersection network screening. *Transportation Research Record*, June 2022.

[109] Rongjie Yu, Mohammed Quddus, Xuesong Wang, and Kui Yang. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. *Accident Analysis & Prevention*, 120:304–310, 2018.

[110] Chengcheng Xu, Pan Liu, Wei Wang, and Zhibin Li. Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis & Prevention*, 47:162–171, 2012.

[111] Shaofeng Sun, Bei Zhou, and Shengrui Zhang. Analysis of factors affecting injury severity in motorcycle involved crashes. In *CICTP 2020*, pages 4207–4219. 2020.

[112] Amir Bahador Parsa, Homa Taghipour, Sybil Derrible, and Abolfazl Kouros Mohammadian. Real-time accident detection: coping with imbalanced data. *Accident Analysis & Prevention*, 129:202–210, 2019.

[113] Pei Li, Mohamed Abdel-Aty, Qing Cai, and Cheng Yuan. The application of novel connected vehicles emulated data on real-time crash potential prediction for arterials. *Accident Analysis & Prevention*, 144, 2020.

[114] Mahama Yahaya, Xinguo Jiang, Chuanyun Fu, Kamal Bashir, and Wenbo Fan. Enhancing crash injury severity prediction on imbalanced crash data by sampling technique with variable selection. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 363–368. IEEE, 2019.

[115] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[116] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[117] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[118] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 475–482. Springer, 2009.

[119] Tomasz Maciejewski and Jerzy Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *2011 IEEE symposium on computational intelligence and data mining (CIDM)*, pages 104–111. IEEE, 2011.

[120] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[121] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[122] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*, 2020.

[123] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[124] Yi Lin, Linchao Li, Hailong Jing, Bin Ran, and Dongye Sun. Automated traffic incident detection with a smaller dataset based on generative adversarial networks. *Accident Analysis & Prevention*, 144:105628, 2020.

[125] Zhijun Chen, Jingming Zhang, Yishi Zhang, and Zihao Huang. Traffic accident data generation based on improved generative adversarial networks. *Sensors*, 21(17):5767, 2021.

[126] Alex Kuefler, Jeremy Morton, Tim Wheeler, and Mykel Kochenderfer. Imitating driver behavior with generative adversarial networks. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 204–211. IEEE, 2017.

[127] Henrik Arnelid, Edvin Listo Zec, and Nasser Mohammadiha. Recurrent conditional generative adversarial networks for autonomous driving sensor modelling. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1613–1618. IEEE, 2019.

[128] Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training. *ACM Computing Surveys (CSUR)*, 54(8):1–49, 2021.

[129] GR Wood. Confidence and prediction intervals for generalised linear accident models. *Accident Analysis & Prevention*, 37(2):267–273, 2005.

[130] John E Ash, Yajie Zou, Dominique Lord, and Yinhai Wang. Comparison of confidence and prediction intervals for different mixed-poisson regression models. *Journal of Transportation Safety & Security*, 13(3):357–379, 2021.

[131] Arijit Chakrabarti and Jayanta K Ghosh. Aic, bic and recent advances in model selection. *Philosophy of statistics*, pages 583–605, 2011.