ADVANCED STATISTICAL METHODS FOR ANALYZING CRASH DATASETS WITH MANY ZERO OBSERVATIONS AND A LONG TAIL: SEMIPARAMETRIC NEGATIVE BINOMIAL DIRICHLET PROCESS MIXTURE AND MODEL SELECTION HEURISTICS

A Dissertation

by

MOHAMMADALI SHIRAZI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Dominique Lord
Committee Members,	Jeffrey Hart
	Luca Quadrifoglio
	Yunlong Zhang
Head of Department,	Robin Autenrieth

December 2018

Major Subject: Civil Engineering

Copyright 2018 Mohammadali Shirazi

ABSTRACT

In this dissertation, first, a flexible model is introduced using a mixture of the Negative Binomial (NB) distribution and a random distribution characterized by Dirichlet process (DP) (referred to as NB-DP). This modeling approach aims to provide a greater flexibility to the NB distribution in order to overcome different limitations of the NB distribution, such as modeling data with many zero observations and a long (or heavy) tail. Application of the NB-DP to two observed datasets indicated that the NB-DP model offers a better performance than the NB when data are characterized by many zero observations and a long tail. In addition to a greater flexibility, the NB-DP provides a clustering by-product that allows the safety analyst to better understand the characteristics of the data or domain.

Second, a methodology is proposed to select the most-likely-true sampling distribution between potential alternatives, based on the characteristic of the data, before fitting the models. The proposed methodology employs two analytic tools: (1) Monte-Carlo Simulations and (2) Machine Learning Classifiers, to design simple heuristics to predict the label of the most-likely-true distribution for analyzing data. Next, this method was first applied to investigate when the Poisson-lognormal is preferred over the NB. The results showed that the kurtosis, skewness and percentage of zeros are the main summary statistics needed to select a distribution between these two alternatives. Then, it was investigated when the Negative Binomial Lindley (NB-L) is preferred over the NB. The results showed that the skewness, coefficient of variation, kurtosis, variance-to-mean ratio,

and the percentage of zeros are among the most important summary statistics (or predictors) required to select a logical distribution between the NB and NB-L.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to my advisor, Dr. Dominique Lord, for his encouragement, support and guidance throughout the completion of this dissertation.

I gratefully thank Dr. Luca Quadrifoglio and Dr. Yunlong Zhang for their hearty support during my doctoral study. My special thanks to Dr. Jeffrey Hart for his insightful comments and suggestions on my research.

I should next extend my sincere appreciation and gratitude to Dr. Soma Dhavala for his time, guidance, and help in many parts of this research.

Last, I would like to thank Dr. Srinivas Geedipally for his support and collaboration during my Ph.D. degree.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Dominique Lord [advisor], Dr. Yunlong Zhang and Dr. Luca Quadrifoglio from the Zachry Department of Civil Engineering, and Dr. Jeffrey Hart from the Department of Statistics. Dr. Soma Dhavala and Dr. Srinivas Geedipally collaborated in writing two papers published based on this dissertation.

Funding Sources

The support for portion of this research was provided in part by a grant from the U.S. Department of Transportation, University Transportation Centers Program to the Safety through Disruption University Transportation Center (451453-19C36). [Disclaimer: The contents of this research reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The research was funded partially by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.]

NOMENCLATURE

AADT	Annual Average Daily Traffic
ABC	Approximate Bayesian Computation
ADT	Annual Daily Traffic
AIC	Akaike Information Criteria
AUC	Area under Curve
CV	Coefficient of Variation
CURE	Cumulative Residual
DIC	Deviance Information Criterion
DP	Dirichlet Process
DT	Decision Tree
FHWA	Federal Highway Administration
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GoF	Goodness of Fit
GoL	Goodness of Logic
HSIS	Highway Safety Information System
LRT	Likelihood Ratio Test
PG	Poisson-gamma
PLN	Poisson-lognormal
MAD	Mean Absolute Deviance

MCMC	Markov Chain Monte Carlo
MSPE	Mean Squared Predictive Error
NB	Negative Binomial
NB-DP	Negative Binomial Dirichlet Process
NB-GE	Negative Binomial Generalized Exponential
NB-L	Negative Binomial Lindley
NB-TDP	Negative Binomial Truncated Dirichlet Process
RF	Random Forest
ROC	Receiver Operating Characteristics
TCDP	Truncated Centered Dirichlet Process
VMR	Variance-to-Mean Ratio

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION*

Regression models have different applications in highway safety. They can be used for estimating the number of crashes, exploring the system information, screening the variables, identifying hazardous sites and ultimately evaluating safety. As documented in Lord and Mannering (2010) and more recently in Mannering and Bhat (2014), research studies have been devoted to develop innovative and novel statistical models to estimate or predict the number of crashes and evaluate roadway safety. The statistical models specifically deal with unique characteristics that are associated with crash data. As such, heterogeneous crash data can often be characterized with high-dispersion, long (or heavy) tail and many observations with the value zero. These unique characteristics inspired researchers to propose new distributions and models that aimed to overcome the limitations associated with the most commonly used model in highway safety literature, the negative binomial (NB) model (also known as the Poisson-gamma model).

---

* Part of this chapter is reprinted with permission from Shirazi, M., Lord, D., Dhavala, S. S., Geedipally, S. R. (2016). A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention, 91, 10-18. Copyright [2016] by Elsevier. DOI: https://doi.org/10.1016/j.aap.2016.02.020 ; and, Shirazi, M., Dhavala, S. S., Lord, D., Geedipally, S. R. (2017). A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the negative binomial Lindley (NB-L) is preferred over the negative binomial (NB). Accident Analysis & Prevention, 107, 186-194. Copyright [2017] by Elsevier. https://doi.org/10.1016/j.aap.2017.07.002

This dissertation, first, contributes to crash data modeling by presenting a class of flexible models using a mixture of the NB and a random distribution characterized by Dirichlet process (DP) to analyze count/crash data (referred to as NB-DP in this dissertation). The goal of this modeling approach is providing a greater flexibility to the NB distribution to model data with many zero responses and a long tail. Then, this dissertation is continued with a discussion on selection of a sampling distribution. A methodology is presented to select the "most-likely-true" (or heuristics to be exact) sampling distribution between potential alternatives, based on characteristic of data. So far, in crash data analysis, the selection of sampling distributions and models have usually been accomplished at the post-modeling phase, using measures such as Goodness of Fit (GoF) statistics or statistical metrics such as the likelihood ratio test (LRT). These metrics are neither easy to compute nor practically doable on some instances when many alternatives exist and/or when the analyst deals with big data or datasets with a large number of zero responses. In addition, and most importantly, these metrics typically do not consider characteristics of data or the "logic" behind the model (Goodness-of-Logic or GoL, as illustrated by Miaou and Lord, 2003) in their model recommendations. The proposed approach in this dissertation, instead, targets designing heuristics for Model Selection that consider the characteristics of data to come up with the model recommendation.

This chapter is divided into three parts. First, the research problem is described. Second, the research objectives are documented. Third, the dissertation outline is presented.

1.1 Research Problem

This section is divided into two parts. First, the motivation behind the introduction of the NB-DP is described. Second, the motivation for Model Selection heuristics is documented.

1.1.1. Modeling

Recent research has shown that the NB model can be significantly affected by datasets characterized by a long (heavy) tail (Zou et al., 2015). According to Guo and Trivedi (2002), the NB regression model cannot properly capture the long tail because a negligible probability is assigned to large counts. A long tail can be caused by the data generating process itself (i.e., including observations with very large counts), or it can also be attributed to datasets that have excess zero observations. In this case, the long tail is created by shifting the overall sample mean closer to zero, which increases the spread of the observations (Lord and Geedipally, 2018).

Over the last few years, a new series of models that mixes the NB distribution with other distributions have been introduced to analyze such datasets. The NB-Lindley (NB-L) (Zamani and Ismail, 2010; Lord and Geedipally, 2011; Geedipally et al., 2012) and the NB-generalized exponential (NB-GE) (Vangala et al., 2015) generalized linear models (GLMs) are two examples of such models. Research studies show that these models perform better than the NB model when data are characterized by many zero observations or have a long tail.

Looking closely at these statistical models, it would become apparent that a recurring theme in all these models (even NB itself) is to consider a mixing distribution at

the heart of the generative model to provide more flexibility in modeling. For example, one can see the NB as a mixture of the Poisson and gamma distributions or the NB-L as a mixture of the NB and the Lindley distributions; even the Lindley distribution itself is a mixture of two gamma distributions. There are primarily three major ingredients for eliciting such mixtures, which offer a greater degree of flexibility in model construction:

1. The mixing weights: the mixing weights determine the relative weight of the individual mixing components.
2. The shape and characteristics of the mixing components or the constituent members of the mixtures, and
3. The level: in the context of hierarchical/multi-level modeling, at which level, the mixture distribution is elicited.

A transportation safety analyst might have a preference to choose or rather not to choose a particular mixture. In all cases, the analyst is required to make certain assertions about the mixture components. One way to retain the modeling flexibility and yet not be overly concerned about the assertions is to express the uncertainty explicitly by considering a random mixing distribution. The Dirichlet process, a widely used prior in Bayesian nonparametric literature, allows such representation (Antoniak, 1974; Escobar and West, 1995). One way to think about the DP is as an infinite mixture distribution, where the number of unique components and the component characteristics themselves can be learned from the data. Taking this motivation in mind, in this research, instead of a fixed shaped (or standard) distribution, a random distribution defined by the DP is mixed with the NB distribution to provide more modeling flexibility in dealing with the

heterogeneous count data and handling the NB limitations when data are characterized with a heavy tail and many zero observations. The NB-DP modeling framework is introduced and applied to analyze two crash datasets collected in Indiana and Michigan.

1.1.2. Model Selection

As noted above, there has been a phenomenal growth in introducing novel distributions and models to analyze crash data over the last decade (see Lord and Mannering, 2010; Mannering and Bhat, 2014). Selecting the most appropriate and logically sound sampling distribution among all these alternatives plays a crucial role in modeling and further systematic safety analyses or evaluations, and has always been a subject of interest to safety scientists or researchers. So far, the comparison of distributions (or models) has usually been accomplished during the post-modeling phase - once data are fitted to all competitive alternatives, using measures such as the Goodness-of-Fit (GoF) statistics or the Likelihood Ratio Test (LRT). However, such metrics are neither easy to compute nor practically doable on some instances when many alternatives exist and/or when the analyst deals with big data or datasets with many zero observations. In addition, and most importantly, these metrics do not provide any intuitions into why one distribution is preferred over another or the logic behind the Model Selection (Goodness-of-Logic, as illustrated by Miaou and Lord, 2003). In this dissertation, these issues are addressed by proposing a methodology to design heuristics for Model Selection, based on characteristics of data, without any post-modeling inputs.

The methodology proposed in this study can be motivated first by looking at the characteristics of the Poisson and NB distributions. The analyst can choose between the Poisson and NB distributions just by looking at the mean (μ) and variance (σ^2) of the data, before fitting the distributions or models. A general rule of thumb is that, when data show a sign of over dispersion (i.e., when $\sigma^2/\mu > 1$), the analyst can move from 'Poisson' to 'NB'. In this case, the variance-to-mean-ratio (VMR) serves as a "heuristic" for Model Selection and the VMR greater than one as a "switching" point. Second, the research problem can be motivated by looking at the characteristics of the NB and NB-L distributions. Both of these distributions can handle over dispersion; however, the NB-L distribution is preferred when data are characterized by many zeros and/or have a heavy (or long) tail (Lord and Geedipally, 2011). Although we know the NB-L distribution performs better when data are skewed, it is not clear at what 'point', the analyst should shift from the 'NB' to the 'NB-L'. In other words, it is not explicitly clear, for example, what the skewness of data should be to prefer the NB-L distribution over the simple NB distribution. Is skewness the only measure to look at while deciding so? We develop a systematic approach to answer such questions.

The problem statement for the selection of sampling distributions can now be introduced: what are the "switching" points to move from one distribution to another when two or more competitive distributions are available? Can we predict the model to be used based on *characteristics* of the data, reflected in its summary statistics, to find the 'most-likely-true' sampling distribution *before* fitting the model? In this dissertation, this topic is addressed by introducing a methodology that provides heuristics to select the 'most-

likely-true' sampling distribution among its competitors, based on *characteristics* of data, reflected into certain summary statistics, *before* fitting the competitive models based on their distributions.

1.2. Research Objectives

The objectives of this dissertation are described below:

First, the NB-DP model is introduced and its characteristics are documented and discussed. The model is introduced based on the Bayesian hierarchical modeling scheme using a mixture of the NB distribution and a random distribution characterized by the DP (referred to as NB-DP).

Second, application of the NB-DP model to analyze data with many zero observations and a heavy tail is investigated. Two datasets, one collected in Indiana and the other in Michigan are used to accomplish this objective.

Third, a methodology is proposed to design heuristics to decide between two or more competitive distributions based on characteristics of data in terms of the summary statistics. The designed heuristics can come up with the model recommendation only based on characteristics of data, without any post modeling efforts or inputs.

Fourth, the proposed methodology is applied to investigate the "switching" points and designing heuristics to select the 'most-likely-true' distribution between (1) the Negative Binomial and Poisson-lognormal (PLN) distributions, and (2) the Negative binomial and Negative Binomial Lindley distributions to model crash or other safety related data.

1.3. Dissertation Outline

The outline of this dissertation is as follows:

Chapter II describes and documents the characteristics of the NB-DP modeling framework. The modeling approach is introduced and its advantage in providing greater flexibility is discussed. Then, it is described how the NB-DP model can be used to cluster data; next, the implementation of the model in a statistical software is discussed.

Chapter III covers the modeling results of applying the NB-DP GLM (with lognormal base distribution) to analyze two datasets, one collected in Indiana and the other one in Michigan. The modeling results are compared with the NB and NB-L GLMs.

Chapter IV documents a methodology to design heuristics for Model Selection based on characteristics of data. The motivations behind the proposed approach is described in detail. The characteristics of the proposed method and detailed algorithm is presented and discussed. Last, the benefits and advantages of the approach are discussed in greater details.

Chapter V formulates heuristics to select a sampling distribution between the NB and PLN, and between the NB and NB-L distributions based on selected summary statistics of data.

Chapter VI concludes the dissertation. It summarizes the key discussion points of the research performed in this work and provides avenues for further research.

CHAPTER II

NB-DP GENERALIZED LINEAR MODEL*

This chapter documents and describes the characteristics of the NB-DP modeling framework. The chapter is divided into five parts. First, a background section is devoted to review and document the characteristics of the NB and NB-L GLMs and the Dirichlet process. Second, the NB-DP modeling framework is documented and discussed. Third, the NB-DP added advantage to cluster data is discussed. Then, the implementation of the model in a statistical software is described. Last, a brief summary of the chapter is provided.

2.1. Background

This section is divided into three parts. In the first section, the characteristics of the NB GLM is documented and reviewed. The second part documents the characteristics of the NB-L GLM. In the third section, the DP and its characteristics are described.

2.1.1. NB GLM

The NB distribution can be formulated given two different parameterizations (Geedipally et al., 2012): (1) a mixture of the Poisson and gamma distributions, or (2) a sequence of

* Part of this chapter is reprinted with permission from Shirazi, M., Lord, D., Dhavala, S. S., Geedipally, S. R. (2016). A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention, 91, 10-18. Copyright [2016] by Elsevier. DOI: https://doi.org/10.1016/j.aap.2016.02.020

independent Bernoulli trials. The probability mass function (pmf) of the negative binomial distribution is defined as follows:

$$P(Y = y|\,\phi, p) = \frac{\Gamma(\phi + y)}{\Gamma(\phi)\Gamma(y+1)}(p)^{\phi}(1-p)^{y}\;;\; 0 < p < 1, \phi > 0 \tag{1}$$

where p = failure probability in each trial and ϕ = inverse dispersion parameter. The long term mean response of observations of the negative binomial distribution is equal to:

$$\mu = \frac{(1-p)\phi}{p} \tag{2}$$

Taking Equation (2) into account, the parameter p can be reparametrized as a function of the mean response of the observation (μ) and the inverse dispersion parameter (ϕ) as,

$$p = \frac{\phi}{\mu + \phi} \tag{3}$$

Given Equations (1) and (3) into account, the pmf of the NB distribution can be structured with the following notation (i.e., as a Poisson-gamma model) which is the common notation that is used in the context of crash data regression modeling.

$$NB(y|\mu, \phi) \equiv p(Y = y|\,\phi, \mu) = \frac{\Gamma(\phi + y)}{\Gamma(\phi)\Gamma(y+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(1 - \frac{\phi}{\mu+\phi}\right)^{y} \tag{4}$$

In context of the NB GLM regression for crash data, the long-term mean response of the NB would have a log-linear relationship with covariates as follows:

$$\ln(\mu) = \beta_0 + \sum_{j=1}^{d} \beta_j X \tag{5}$$

where β_j = j^{th} regression coefficient, X = d-dimensional observed covariates, and d= number of covariates.

2.1.2. NB-L GLM

The NB-L model (Geedipally et al, 2012) is defined using a mixture of the NB and Lindley distributions as follows:

$$P(Y = y|\ \phi, \mu, \theta) = \int NB(y|\phi, v\mu)\text{Lindley}(v|\theta)dv \tag{6}$$

The pmf of the Lindley distribution is defined as:

$$\text{Lindley}(v|\theta) = \frac{\theta^2}{\theta + 1}(1 + v)e^{-\theta v} \qquad \theta > 0, v > 0 \tag{7}$$

The Lindley distribution is a mixture of two gamma distributions as follows:

$$v \sim \frac{1}{1 + \theta}\text{gamma}(2, \theta) + \frac{\theta}{1 + \theta}\text{gamma}(1, \theta) \tag{8}$$

Therefore, the NB-L model can be written as the following hierarchical model:

$$y \sim NB(y|\phi, v\mu) \tag{9-a}$$

$$v \sim \text{gamma}(1 + z, \theta) \tag{9-b}$$

$$z \sim \text{Bernoulli}(\frac{1}{1 + \theta}) \tag{9-c}$$

$$\ln(\mu) = \beta_0 + \sum_{j=1}^{d} \beta_j X \tag{9-d}$$

Geedipally et al. (2012) showed that the NB-L model performs better than the NB model when data have many zero observations or characterized by a long tail.

2.1.3. Dirichlet Process

There has been a phenomenal growth in theory, inference and applications concerning the DP and its related processes in the last decade; recent monographs on Bayesian nonparametric devoting significant portion on the DP and related processes is a testimony to that effect (Hjort et al., 2010; Mitra and Muller, 2015). On the application side, the DP has been applied in numerous fields ranging from network modeling (Ghosh et al., 2010) to Bioinformatics (Dhavala et al., 2010; Argiento et al., 2015) to Psychometrics (Miyazaki and Hoshino, 2009) to name a few. In particular, the application of the DP to account for over-dispersion in count data has been considered in Mukhopadhyay and Gelfand (1997) and Carota and Parmigiani (2002), with Binomial and Poisson based likelihoods.

Traditionally, the Bayesian parametric inference mechanism considers a parametric distribution $F_0(.|\theta)$, where θ is a finite vector of parameters, as a prior for the unknown parameter. However, constraining the model within specific parametric families could limit the scope of the inference. To overcome this difficulty, in context of the Bayesian nonparametric (or semiparametric) modeling, a random prior distribution is considered for the parameter as opposed to choosing a prior distribution from a known parametric family. The prior is placed over infinite-dimension space of distribution functions. In that sense, it gives more flexibility to the parameter inference mechanism by providing a wide range of prior distributions.

The DP (Ferguson, 1973; Ferguson, 1974) is a stochastic process that is usually used as a prior in Bayesian nonparametric (or semiparametric) modeling. Escobar and West (1998)

define the DP as a random probability measure over the space of all probability measures. In that sense, the DP is considered as a distribution over all possible distributions; that is, each draw from the DP is itself a distribution.

Let $A_1, A_2, .., A_r$ be any finite measurable partitions of the parameter space (Θ). Let us assume τ be a positive real number and $F_0(.|\theta)$ be a continuous distribution over Θ. Then, $F(.) \sim DP(\tau, F_0(.|\theta))$ if and only if (Escobar and West, 1998):

$$\big(F(A_1), F(A_2), \dots, F(A_r)\big) \sim \text{Dirichlet}\big(\tau F_0(A_1|\theta), \tau F_0(A_2|\theta), \dots, \tau F_0(A_r|\theta)\big) \qquad (10)$$

where τ is defined as the precision (or concentration) parameter and $F_0(.|\theta)$ as the base (or baseline) distribution. Note that based on the Dirichlet distribution properties, for each partition $A \subset \Theta$, we have:

$$E\big(F(A)\big) = F_0(A|\theta)$$

$$\text{var}\big(F(A)\big) = \frac{F_0(A|\theta)\big(1 - F_0(A|\theta)\big)}{1 + \tau}$$

Therefore, the base distribution $F_0(.|\theta)$ and the precision parameter τ play significant roles in the DP definition. The expectation of the random distribution $F(.)$ is the base distribution $F_0(.|\theta)$. Likewise, the precision parameter τ controls the variance of the random distribution around its mean. In other words, τ measures the variability of the target distribution around the base distribution. As $\tau \to \infty$, we would have $F(.) \to F_0(.|\theta)$ while, on the other hand, as $\tau \to 0$, the random distribution $F(.)$ would deviate further away from $F_0(.|\theta)$.

Equation (10) defines the DP indirectly through the marginal probabilities assigned to finite number of partitions. Therefore, it gives no intuition on realizations of $F(.) \sim DP(\tau, F_0(.|\theta))$. To simulate random distributions from the DP, however, Sethuraman (1994) introduced a straightforward stick-breaking constructive representation of this process as follows:

$$\gamma_k|\tau \sim \text{Beta}(1,\tau), \qquad k = 1,2,\ldots \tag{11-a}$$

$$\psi_k|\,\theta \sim F_0(.|\theta), \quad k = 1,2,\ldots \tag{11-b}$$

$$p_k = \gamma_k \prod_{k'<k} (1-\gamma_{k'}), \quad k = 1,2,\ldots \tag{11-c}$$

$$F(.) \sim DP(\tau, F_0(.|\theta)) \equiv \sum_k p_k \delta_{\psi_k} \tag{11-d}$$

where δ_{ψ_k} indicates a degenerate distribution with all its mass at ψ_k. This construction, metaphorically, can be considered as breaking a unit length of stick iteratively (Ishwaran and James, 2001). First, the stick is broken at a random proportion γ_1; an atom is generated from the base distribution (ψ_1) and is assigned to the length of the stick that was just broken (p_1). Then, recursively, the remaining portions of the stick are broken at new proportions ($\gamma_2, \gamma_3, ..$); new atoms are generated from the base distribution ($\psi_2, \psi_3, ..$) and are assigned to each broken length of the remaining sticks ($p_2, p_3, \ldots$).

Given the stick-breaking construction of the DP (Equation 11), the mean and variance of $v \sim F(.)$ can be calculated as follows (Yang et al., 2010):

$$E(v|p,\psi) = \mu_{DP} = \sum_k p_k \psi_k \tag{12}$$

$$\mathrm{var}(v|p,\psi) = v_{DP} = \sum_{k} p_k \psi_k^2 - \left(\sum_{k} p_k \psi_k\right)^2 \quad (13)$$

As indicated in Equation (11), theoretically, the stick-breaking construction of the DP includes infinite components (so called clusters); however, practically, the model can be approximated with its truncated version (TDP) by considering an upper bound on the number of components (M) as follows (Ishwaran and James, 2001; Ishwaran and Zarepour, 2002):

$$\gamma_k|\tau \sim \mathrm{Beta}(1,\tau)\,, \qquad k = 1,2,..,M \quad (14\text{-a})$$

$$\psi_k|\,\theta \sim F_0(.|\theta), \quad k = 1,2,\ldots.,M \quad (14\text{-b})$$

$$p_k = \gamma_k \prod_{k'<k} (1-\gamma_{k'}), \quad k = 1,2,\ldots,M \quad (14\text{-c})$$

$$F(.) \sim \mathrm{TDP}\big(\tau, M, F_0(.|\theta)\big) \equiv \sum_{k=1}^{M} p_k \delta_{\psi_k} \quad (14\text{-d})$$

So far, several research studies have tried to estimate the required number of components (or clusters) (M) in the truncated version of the DP (Ishwaran and James, 2001; Ohlssen et al., 2007). As a key point, first, the analyst needs to keep in mind that the number of mass points (M) in the TDP is correlated to the value of the precision parameter (τ). Theoretically, as the value of τ increases, the number of clusters that are shared by data points increases; hence, a larger value for the parameter M is required. Second, the model needs to be approximated to the level that it can be assumed that the effect of neglected clusters remains negligible ($1 - \sum_{k=1}^{M} p_k \approx \varepsilon$). Given these two rationales into account, Ohlssen et al. (2007) showed that the maximum number of clusters

can be approximated by Equation (15) as a function of τ and the desired ε- accuracy as follows:

$$M \approx 1 + \frac{\log(\varepsilon)}{\log\left(\frac{\tau}{1+\tau}\right)} \tag{15}$$

Once the model is approximated to M clusters, p_M needs to be modified using Equation (16) to make the model identifiable (i.e.: $\sum_{k=1}^{M} p_k = 1$):

$$p_M = 1 - \sum_{k=1}^{M-1} p_k \tag{16}$$

2.2. NB-DP GLM

The NB-DP class of models can be motivated, first, by looking at the NB model as a mixture of the Poisson and gamma distributions. As an extension of the Poisson model, the Poisson-gamma was developed assuming that the Poisson parameter is measured with a random error; this random error itself is gamma distributed. The Poisson-gamma mixture is thought to be a better alternative to accommodate possible over-dispersion in data (Hilbe, 2011). Second, it can be motivated by looking at the NB-L model as a mixture of the negative binomial and the Lindley distributions. The NB-L model can overcome the NB limitations when data are over-dispersed and have many zeros. Essentially, as discussed in Chapter I, although mixture models are providing better alternatives, they assume the shape and density of the distributions to be fixed. However, we can obtain even more flexibility by assuming that the mixing distribution itself is random. Given this

motivation in mind, this dissertation plans to develop a model using a mixture of the NB and a random distribution characterized by the DP.

The NB-DP distribution is defined as a mixture of the NB distribution and a random distribution characterized by the DP with a precision parameter τ and a base distribution $F_0(.|\theta)$ as follows:

$$p\big(Y = y|\mu, \phi, \tau, F_0(.|\theta)\big) = \int NB(y|v\mu, \phi)\, dF\big[v|DP\big(\tau, F_0(.|\theta)\big)\big] \tag{17}$$

The structure used to mix the NB distribution and the random distribution $F(.)$ is similar to the one that was used to introduce the mixture of the negative binomial and Lindley distribution (Geedipally et al., 2012). In this study, however, instead of the Lindley distribution, the NB distribution is mixed with a random distribution characterized by the DP to provide a more flexible model in order to better estimate the long term mean response of the negative binomial. Nonetheless, since the involved integration in NB-DP model does not have a closed form, the model cannot (or difficult) to be used with the format shown in Equation (17) to regress the count data. In order to solve this difficulty, the model was reformulated using the Bayesian hierarchical scheme as follows:

$$y_i|v_i\mu_i, \phi \sim NB(v_i\mu_i, \phi) \tag{18-a}$$

$$v_i \sim F(.) \tag{18-b}$$

$$F(.) \sim DP(\tau, F_0(.|\theta)) \tag{18-c}$$

In context of the GLM regression for crash data, the long-term mean response of the NB-DP would have a log-linear relationship with covariates as follows:

$$\ln(\mu_i) = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \tag{19}$$

where β_j = j^{th} regression coefficient, x_{ij} = d-dimensional observed covariates and d= number of covariates. Also, as noted in Section 2.1.3, the distribution of $DP(\tau, F(.|\theta))$ can be approximated by its truncated construction $TDP(\tau, M, F(.|\theta))$. Consequently, the NB-TDP model framework can be seen as a hierarchical Bayesian model described below:

$$y_i|v_i\mu_i, \phi \sim NB(v_i\mu_i, \phi) \tag{20-a}$$

$$\gamma_k|\tau \sim Beta(1, \tau)\,, \qquad k = 1,2,..,M \tag{20-b}$$

$$\psi_k|\,\theta \sim F_0(.|\theta), \quad k = 1,2,\ldots.,M \tag{20-c}$$

$$p_k = \gamma_k \prod_{k'<k} (1-\gamma_{k'}), \quad k = 1,2,\ldots,M \tag{20-d}$$

$$v_i \sim F\,(.) \tag{20-e}$$

$$F(.) \sim TDP\big(\tau, M, F_0(.|\theta)\big) \equiv \sum_{k=1}^{M} p_k \delta_{\psi_k} \tag{20-f}$$

$$\ln(\mu_i) = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \tag{20-g}$$

The model in Equation (20) is referred to as a modeling framework in this dissertation since the base distribution $F_0(.|\theta)$ can have any desired distributions such as (1) lognormal, (2) skew-lognormal (3) Lindley, or (4) Generalized Exponential, etc. Hence, the framework can be used for a wide range or class of modeling approaches.

The model described above can be thought in context of the Generalized Linear Mixed Model (GLMM) (Booth et al., 2003), where the mixed effects or frailty terms (v_i) are given a random distribution characterized by the DP with a precision parameter τ and a base distribution $F(.|\theta)$. One simple way to think about it is that if the precision parameter was infinite or very large, the distribution of mixed effects (v_i) would be very close to the base distribution (i.e., simply v_i would follow the base distribution). The precision τ, however, controls how much we know about the base distribution and in that sense the DP provides a random distribution to better accommodate the dispersion in data.

2.3. Clustering by NB-TDP GLM

In addition to providing a greater flexibility, there is an added advantage of an in-built clustering algorithm in the model (Equation 20). This unique clustering by-product is based on how sites shared the mixed effect mass points. In other words, each mass point can be considered as a cluster. The clustering advantage can be used for different purposes, such as detecting groups of units with unusual results (detecting outliers), examining the characteristics of clusters to develop crash modification factors or to implement an appropriate countermeasure, or sources of dispersion, as described above.

In order to benefit from the clustering by-product, the hard-clustering information (i.e., the information about which two data points shared the same mass point or cluster) should be recorded at each iteration of the Markov Chain Monte Carlo (MCMC) sampling. Let Z_{mn}^{q} be the component of the association matrix which is 1 if the data points "m" and "n" belong to the same cluster and 0 otherwise in the q-th MCMC sample. By definition,

Z is symmetric and $Z_{mm} = 1$. Now, the information in matrix Z can be used to elicit the clustering properties and perform further post-processing analyses (Ohlssen et al, 2007). For instance, the likelihood that site "m" and site "n" fall into the same cluster can be found by taking an average of Z_{mn}^{q} over all MCMC outputs. As another example, the matrix Z can be used to identify outliers. For this purpose, the variable W_m^q is defined as $W_m^q = \sum_{n=1}^{N} Z_{mn}^{q}$. The variable W_m^q shows the size of the cluster that the site "m" belonged to at the q-th iteration of the MCMC. Now, the mean of the cluster size can be found by taking an average of W_m^q over all MCMC outputs. Then, choosing a threshold (say 3 for example), the potential outliers can be detected.

2.4. Implementation of the NB-TDP GLM

Given an appropriate choice for the DP base distribution, all stages of the model (Equation 20) would involve only standard distributions. Therefore, the model can be implemented in a software program, such as WinBUGS (Spiegelhalter et al., 2003; Ohlssen et al, 2007) to estimate the coefficients. Based on how the Bayesian model was parameterized and the definition of the Dirichlet process, the base distribution is a non-negative distribution that the analyst believes the frailty terms on average could follow a priori. In this dissertation (in Chapter III) a lognormal distribution is used as the DP base distribution (i.e., $\ln(v_i) \sim N(\mu_b, \sigma_b^2)$). However, as disused in Section 2.2, a wide range of distributions can be used instead of the base distribution such as the Lindley or Generalized Exponential distributions.

Likewise, the analyst must make sure that the NB-DP model is identifiable (i.e., $\text{median}(v_i) = 1$) to eliminate possible correlation between the intercept (β_0) and frailty terms (v). This issue can be overcome, initially, by dropping the intercept from the model ($\beta_0 = 0$); then, after the MCMC convergence, the log-median of the mixed effects can be used instead of the intercept. In Chapter III, another intuitive method is discussed to overcome the identifiability issue using the truncated centered Dirichlet process (TCDP) method based on (Yang et al., 2010) idea to constrain the mean and variance of the Dirichlet process.

2.5. Chapter Summary

This chapter documented the development of the NB-DP (or NB-TDP to be exact) GLM. This model mixes the NB distribution with a random distribution characterized by the DP. The model can be thought in context of the Bayesian hierarchical modeling framework, where the mixed effects are given a flexible distribution. In fact, each draw from the DP is a distribution and, in that sense, instead of being constrained to a particular shape or distribution, a range of distributions is considered as a prior for mixed effects. In that regard, it provides more flexibility for the model to capture the variation in the data as well as handling issues, such as a heavy tail or many zero observations. In addition to a greater flexibility, the NB-DP model groups the data points into finite number of clusters. The clustering information can provide further insights for the transportation safety analyst, such as a better understanding of the data at hand, identify safety issues and decide

on countermeasures. The next chapter describes the application of the NB-DP using crash data.

CHAPTER III

APPLICATION OF THE NB-DP GLM*

In this chapter, the performance of the NB-DP (or NB-TDP to be exact) GLM is evaluated using two datasets, one collected in Indiana and the other one in Michigan. This chapter is divided into Four parts. First, the characteristics of two observed datasets used for the analysis are described. Second, the applications of the NB-DP to these datasets to analyze crash data are documented and discussed. Third, a few remarks about implementation of NB-DP are discussed. In the end, a brief summary is provided.

3.1. Data Description

This section documents the statistics of the datasets that were used in this chapter. The datasets were used to compare the performance of the NB-TDP GLM with NB and NB-L GLMs. The first subsection briefly describes the summary statistics of the Indiana dataset. The second subsection summarizes the characteristics of the Michigan dataset. Both datasets are characterized by high dispersion and have a heavy tail.

---

* Part of this chapter is reprinted with permission from Shirazi, M., Lord, D., Dhavala, S. S., Geedipally, S. R. (2016). A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention, 91, 10-18. Copyright [2016] by Elsevier. DOI: https://doi.org/10.1016/j.aap.2016.02.020

3.1.1. Indiana data

The Indiana data contain crash, average daily traffic (ADT) and geometric design data collected for the duration of five-years from 1995 to 1999 at 338 rural interstate road sections in Indiana. This dataset has been extensively used by others (Anastasopoulos et al., 2008; Washington et al., 2011; Geedipally et al., 2012). Out of 338 highway segments in this dataset, 120 of them did not experience any crash (approximately 36% of sites are reported with zero crashes). Table 1 shows the summary statistics of the variables of this dataset (Geedipally et al., 2012; Shirazi et al., 2016b). The complete list of variables can be found in Washington et al. (2011). The Indiana dataset is characterized by a heavy tail that is caused by the data generating process of the data (i.e., the dataset includes observations with very large values).

Table 1. Characteristics of the Indiana Data (Reprinted with Permission from Shirazi et al., 2016b).

Variable	Min	Max	Avg.	Std. dev
No. of crashes (5 years)	0	329	16.97	36.30
Average daily traffic in 5 years (ADT)	9,942	143,422	30,237.6	2,8776.4
Minimum friction on the road segment (5-year period) (FRICTION)	15.9	48.2	30.51	6.67
Pavement type (1 if asphalt, 0 if concrete) (PAVEMENT)	0	1	0.77	0.42
Median width (feet) (MW)	16	194.7	66.98	34.17
Presence of the median barrier (1 if present, 0 if absent) (BARRIER)	0	1	0.16	0.37
Interior rumble strips (RUMBLE)	0	1	0.72	0.45
Segment length (miles) (L)	0.009	11.53	0.89	1.48

3.1.2. Michigan data

The Michigan dataset includes 3,397 randomly selected (10% of the original dataset) rural two-lane highways segments in Michigan that contained single-vehicle crashes occurred in 2006; this sample was selected because of the WinBUGS memory limitation. The original dataset was collected from the Federal Highway Administration's (FHWA) Highway Safety Information System (HSIS). The dataset was used previously in Qin et al. (2004) to introduce the zero-inflated models and in Geedipally et al. (2012) to develop the NB-L GLM. In this dataset, about 70% of segments did not experience any crash. The summary statistics of the data used in this research are shown in Table 2.

Table 2. Characteristics of the Michigan Data (Reprinted with Permission from Shirazi et al., 2016b).

Variable*	Min	Max	Avg.	Std. dev.
Number of Crashes (1 year)	0	40	0.717	1.782
Annual average daily traffic (AADT)	250	19,990	4,531.77	3,290.66
Segment length (miles) (L)	0.001	4.323	0.18	0.33
Shoulder width (feet) (SW)	0	12	8.46	2.80
Lane width (feet) (LW)	8	15	11.25	0.79
Speed limit (mph) (SPEED)	25	55	52.49	6.34

*Randomly selected 10% of the original dataset.

3.2. Modeling Results

This section documents the detailed results of the application of the NB-TDP GLM to the Indiana and Michigan datasets. In this section, The NB-TDP modeling results is also compared with the NB and the NB-L GLMs. To fully specify the NB-TDP model, a normal prior was chosen for β and μ_b, a gamma prior for ϕ, and a uniform prior for σ_b^2 and τ. Moreover, given Equation (15), if we assume $\varepsilon = 0.01$ and set the upper bound of the uniform prior that is considered for precision parameter τ to 5, the parameter M would approximately be equal to 27. Hence, to round up, we set M=30. The MCMC was performed with three different chains each with 30,000 iterations. The first 15,000 samples of each chain were regarded as burn-in samples and discarded from the MCMC outputs. The chains were diagnosed using the Gelman-Rubin convergence statistic as well as the visual observations of the history plots. All chains mixed well and the Gelman Rubin statistic was almost 1 for all parameter estimates.

3.2.1. Indiana data

In all models, the segment length was considered as an offset; thus, it is assumed that the number of crashes increases linearly as the segment length increases. Table 3 presents the modeling results for the Indiana data for the NB, NB-L and NB-TDP GLMs. Given the GoF statistics shown in this table, the NB-TDP model showed a better fit compared to other GLMs. A key point to compare different models together based on GoF measures, however, is to consider their complexities. The Deviance Information Criterion (DIC)

statistics penalize the model complexity in its estimates; hence, a more reliable option to employ when models are characterized by different complexities (as it is in our case).

Table 3. Modeling Results for the Indiana Data (Reprinted with Permission from Shirazi et al., 2016b).

Variable	NB		NB-L		NB-TDP	
	value	Std. dev	value	Std. dev	value	Std. dev
Intercept (β_0)	-4.779	0.979	-3.739	1.115	-7.547	1.227
Ln(ADT) (β_1)	0.7219	0.091	0.630	0.106	0.9832	0.1168
Friction (β_2)	-0.02774	0.008	-0.0275	0.011	-0.01999	0.008
Pavement (β_3)	0.4613	0.135	0.4327	0.217	0.3942	0.152
MW (β_4)	-0.0050	0.001	-0.0062	0.002	-0.00468	0.002
Barrier (β_5)	-3.195	0.234	-3.238	0.326	-8.035	1.225
Rumble (β_6)	-0.4047	0.131	-0.3976	0.213	-0.378	0.150
$\alpha = 1/\varphi$	0.934	0.118	0.238	0.083	0.301	0.085
DIC[a]	1900		1701		**1638[d]**	
MAD[b]	6.91		6.89		**6.63**	
MSPE[c]	206.79		195.54		**194.5**	

[a] Deviance Information Criterion.
[b] Mean Absolute Deviance (Oh et al., 2003).
[c] Mean Squared Predictive Error (Oh et al., 2003).
[d] Bold values show a better GoF.

It is worth pointing out that the DIC for flexible models needs to be calculated with some cautions as it may give rise to bi-modal marginal distributions for the estimates (Ohlssen et al., 2007). For this reason, WinBUGS does not calculate the DIC automatically for flexible models. However, similar to what was experienced in Ohlssen et al. (2007), only a few bimodal distributions were identified for the estimates; hence, the DIC measure for this model can also be calculated outside of WinBUGS. The approach discussed in Geedipally et al. (2014) for estimating the DIC was used in this research. As it is indicated in Table 3, for this dataset, the NB-TDP model showed a better DIC between the analyzed models.

For all models, the 95% posterior credible region of none of the parameters includes zero; hence, all included variables are statistically significant. In addition, all coefficients have the same and intuitively reasonable sign. However, the estimated coefficient for each model is not necessary the same. In particular, as a key covariate to predict the number of crashes, different models estimated different ADT coefficients. The ADT coefficient is below 1 based on the NB and NB-L modeling results; it is, however, almost 1 based on the NB-TDP modeling results. Therefore, as the ADT increases, the number of crashes increases at a decreasing rate given the NB and NB-L estimate while almost linearly given the NB-TDP estimate. The Cumulative Residual (CURE) plot can be used to investigate this observation in detail. The cumulative residual plot estimates how well the proposed model fits data regarding key covariates (Hauer and Bamfo 1997). A better fit, then, occurs once this plot oscillates more closely around zero. For a better comparison, the CURE plot is usually adjusted to make the final cumulative value to be

zero. Figure 1 presents the adjusted CURE plot with respect to the ADT covariate (a key variable to estimate the number of crashes). Figure 1 shows that, with respect to the ADT covariate, both NB-L and NB-TDP models fit the Indiana data better than the NB model.

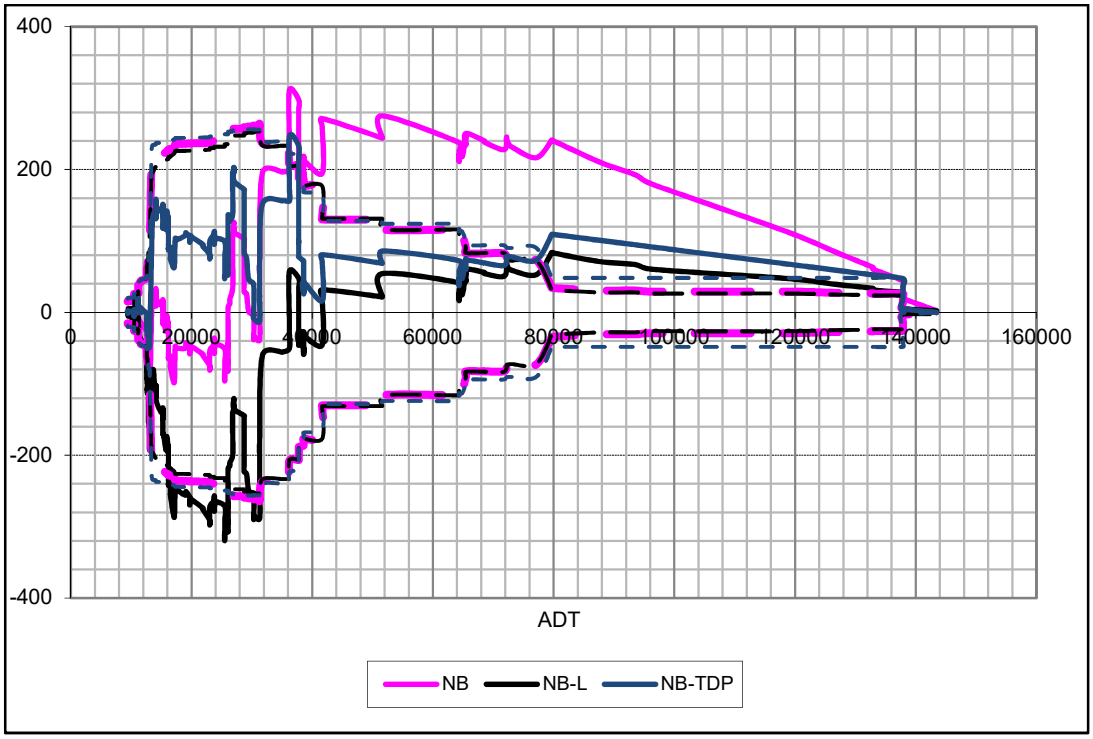

Figure 1. CURE Plots for the Indiana Dataset for the ADT Variable (Reprinted with Permission from Shirazi et al., 2016b).

As discussed in Chapter II, as a by-product of the NB-TDP GLM, data can be classified into finite number of clusters. This clustering property is based on how different sites share the mixed effect mass points (v). In order to benefit from the advantage of clustering, the partitioning information matrix needs to be recorded at each iteration of the MCMC, as discussed above. The matrix can be used to investigate similarities between

sites especially with regard to recognizing unobserved variables (note: in our model, the DP was elicited on mixed effects), or identifying safety issues and deploying countermeasures.

Let the 338 sites in the Indiana dataset be marked in descending order of ADT values in numbers from 1 to 338. Figure 2 shows the heatmap representation of the partitioning matrix for the top 10 sites with the highest ADT values. The figure shows the likelihood that site "X" and "Y" fall into same cluster. For simplicity, the probabilities were rounded to the first decimal. A higher likelihood will be represented by a darker shade in the map. As observed in this figure, for instance, with relatively high probability (~60%), site "1" falls into the same cluster as site "2", site "3" or several more. This information can offer insights to identify potential unobserved variables or safety issues and decide on appropriate countermeasures for the site "1". On the other hand, the probability that site "1" falls into the same cluster as site "9" or site "10" is very small (~10%); hence, there are very few similarities between these sites. In short, the heatmap can be extended to the entire network and be plotted in a 338×338 dimension matrix, which can provide a great visual tool to investigate similarities or dissimilarities between sites, at least with regard to identifying unobserved variables or safety issues. It is worth pointing out that the NB-TDP GLM, on average, classified the Indiana data into approximately 10 clusters (note: the posterior estimation of the precision parameter τ is equal to 2.01).

Site	1	2	3	4	5	6	7	8	9	10
1	1.0	0.6	0.6	0.6	0.6	0.2	0.6	0.6	0.1	0.1
2	0.6	1.0	0.6	0.6	0.6	0.2	0.6	0.6	0.1	0.1
3	0.6	0.6	1.0	0.6	0.6	0.2	0.6	0.6	0.1	0.1
4	0.6	0.6	0.6	1.0	0.6	0.2	0.6	0.6	0.1	0.1
5	0.6	0.6	0.6	0.6	1.0	0.2	0.6	0.6	0.1	0.1
6	0.2	0.2	0.2	0.2	0.2	1.0	0.2	0.2	0.6	0.6
7	0.6	0.6	0.6	0.6	0.6	0.2	1.0	0.6	0.1	0.1
8	0.6	0.6	0.6	0.6	0.6	0.2	0.6	1.0	0.1	0.1
9	0.1	0.1	0.1	0.1	0.1	0.6	0.1	0.1	1.0	0.6
10	0.1	0.1	0.1	0.1	0.1	0.6	0.1	0.1	0.6	1.0

Figure 2. Heatmap Representation of the Partitioning Matrix for the Top 10 Sites with the Highest ADT Values in the Indiana Dataset (Reprinted with Permission from Shirazi et al., 2016b).

3.2.2 Michigan Data

The functional form that was used in Qin et al. (2004) and Geedipally et al. (2012) to analyze the original dataset is used here in order to compare the models adequately. Unlike the Indiana data, the segment length was considered as a covariate in models (i.e., it is not an offset) similar to the original 2004 paper. However, as shown in Table 4, the coefficient of the segment length is almost 1 for all models; hence, the number of crashes increases

almost linearly once the segment length increases. Table 4 shows the rest of the modeling results. The sign for all the coefficients (those that are statistically significant) are the same as those found in Qin et al. (2004) and were left as is to be consistent with their work. For this dataset, unlike the Indiana data, different models estimated relatively similar coefficient values.

Table 4. Modeling Results for the Michigan Data (Reprinted with Permission from Shirazi et al., 2016b).

Variable	NB		NB-L		NB-TDP	
	value	**Std. dev**	**value**	**Std. dev**	**value**	**Std. dev**
Intercept (β_0)	-3.581	0.6353	-3.508	0.6789	-4.222	0.6711
Ln(ADT) (β_1)	0.4521	0.03935	0.4491	0.04217	0.4739	0.04045
Ln(L) (β_2)	0.942	0.02659	0.940	0.02909	0.968	0.02835
SW(β_3)	*0.00425*[a]	*0.0137*	*0.00491*	*0.0144*	*0.00400*	*0.0141*
LW (β_4)	0.018	0.03664	0.018	0.03916	0.034	0.03878
Speed (β_5)	0.018	0.006298	0.018	0.006629	0.022	0.006836
$\alpha = 1/\varphi$	0.6165	0.0617	0.0262	0.0202	0.0303	0.0209
DIC[b]	6223		**5796**[e]		5984	
MAD[c]	0.682		0.689		**0.665**	
MSPE[d]	1.635		1.641		**1.635**	

[a] Italic means not statistically significant at the 5% level.
[b] Deviance Information Criterion.
[c] Mean Absolute Deviance (Oh et al., 2003).
[d] Mean Squared Predictive Error (Oh et al., 2003).
[e] Bold values show a better GoF.

Table 4 shows that the NB-TDP model fits data slightly better than the NB and NB-L models based on the MAD and MSPE GoF measures. Given the DIC measure, both NB-L and NB-TDP models (as a class of multi-parameter models) fit the Michigan data better than the NB model; as discussed above, the DIC is a better measure of fit for complex hierarchical models than GoF measures based on the model errors since it penalizes the model complexity. The posterior estimate of the precision parameter τ for this dataset is equal to 3.29 and data on average were classified into 21.34 clusters. Note that intuitively it is expected that the crash data be grouped into more clusters once the number of sites in the dataset increases.

For this dataset, the DIC estimate for the NB-L model is better than the NB-TDP model. This is due the fact that, first, the NB-L mixture with its fixed distribution is specifically designed to accommodate data with many zeros (i.e., the NB-L distribution has a large density at zero). The NB-TDP model, on the other hand, provides more flexibility to capture the variation in data. Unlike the heavy tail in Indiana data which was characterized by high variation in dataset causing by large as well as small numbers of zero values (the range is 329 with ~36% zeros), the heavy tail in the Michigan dataset is characterized by a large number of zero values (the range is 40 with ~70% zeros). Second, we assumed a uniform distribution for the precision parameter and set the number of NB-TDP mass points to 30. In this case, the precision parameter can adapt to the data, and these data can be grouped up to 30 clusters. For cases when the safety analyst would like to attain a better fit, the precision parameter can be centered to larger values and the NB-TDP can be truncated with a larger number of mass points (clusters). The latter approach,

however, can be problematic to implement in WinBUGS due to its limitations; hence, the analyst should try other alternatives. Some alternative approaches to inference the Dirichlet process are discussed in Section 3.3.

3.3. Discussion

The application of the NB-DP or NB-TDP merits important discussion points. Recall that we have proposed a multi-level hierarchical model to account for over-dispersion and elicited a DP prior on the mixed effects to provide modeling flexibility. One of the critical choices we made was to truncate the Dirichlet process to have finite number of components. Statistical inference in such complex models is facilitated by employing simulation techniques, such as the MCMC. We coded the truncated model in WinBUGS to estimate the model's coefficients (i.e., infer the parameters). There are several aspects that need to be discussed with building the model and the subsequent analysis undertaken in this work, namely, truncation and inference, centering and scaling of the Dirichlet process prior for identifiability and better convergence, and the clustering property.

There are two major tasks involved in Bayesian model building: model elicitation and inference. Traditionally, except in very limited cases, Bayesian modeling in general and Bayesian nonparametric in particular, rely on MCMC for inference, as the models are generally non-tractable. One of the earliest approaches to inference under the full DP representation was due to the seminal work by Escobar and West (1995), followed by several others (Escobar and West, 1998; MacEachern and Muller, 1998). Inference in more complex models, however, was made possible due to samplers, such as the slice

sampling method (Griffin and Walker, 2011; Kalli et al. 2011). Another interesting avenue was considered by approximating the DP with a finite sum representation (Ishwaran and Zarepour, 2002). The advantage with the finite sum based approximation is that, the resulting model is much simpler and often can be fitted using standard software programs, such as WinBUGS. Consequently, the analyst can focus on trying several different models without worrying about writing a new sampler or debugging. However, such approximation comes at a cost: where to truncate? Fortunately, heuristics are available (Ishwaran and James, 2001; Ohlssen et al., 2007) to provide reasonable results, which may work very well in practice, as was the case in this study. However, the same benefit of finite sums representation can be achieved even without truncation, as it is the core idea behind retrospective sampling (Papaspiliopoulos and Roberts, 2008). In this case, a price that one needs to pay is that a significant amount of effort is required in designing and developing the samplers, as opposed to focusing more on model building.

Another very useful approach to approximate inference, the Variational Inference, tries to approximate the true posterior with its closest parametric counterpart that is much more tractable analytically (Blei and Jordan, 2006). In fact, off late, approximate inferences as opposed to exact inferences are becoming popular, such as the Approximate Bayesian Computation framework (Beaumont et al., 2002; Pudlo et al., 2014) and the emerging methods under the umbrella of Big Data (Neiswanger et al., 2013; Bardenet et al. 2014; Quiroz et al., 2015). The approximate inference methods can also be found in the frequentist literature (see Bhat, 2014). The exact approaches to inference can be carried out when the analyst has reasonable understanding of the domain (or data) with respect to

model elicitation. The motivation for choosing approximate inferential methods is speed and agility, either in model building or fitting or both. In subsequent work in this area, we will focus on efficient inference mechanisms that exploit the model characteristics.

An important challenge we faced in this work was the parameterization and identifiability. As discussed briefly earlier, the intercept term and the mean of the mixed effects are correlated. An alternative approach to solve the identifiability issue as well as to obtain a better convergence properties, is to model the mixed effects v_i with the TCDP with constrained variance using the idea proposed in Yang et al. (2010), instead of simple truncated Dirichlet process. The TCDP model given the precision parameter τ and lognormal base distribution is structured as:

$$\ln(\mathrm{v_i}) \sim \mathrm{TCDP}\left(\tau, \mathrm{M}, \mathrm{N}\left(\mu_\mathrm{b}, \sigma_\mathrm{b}^2\right)\right)$$

If and only if

$$\gamma_\mathrm{k} \sim \mathrm{Beta}(1, \tau)\,, \qquad \mathrm{k} = 1,2,..,\mathrm{M} \tag{21-a}$$

$$\psi_\mathrm{k} | \mu_\mathrm{b}, \sigma_\mathrm{b}^2 \sim \mathrm{N}\left(\mu_\mathrm{b}, \sigma_\mathrm{b}^2\right), \quad \mathrm{k} = 1,2,\ldots.,\mathrm{M} \tag{21-b}$$

$$\mathrm{p_k} = \gamma_\mathrm{k} \prod_{\mathrm{k}'<\mathrm{k}} (1 - \gamma_{\mathrm{k}'}), \quad \mathrm{k} = 1,2,\ldots,\mathrm{M} \tag{21-c}$$

$$\omega_\mathrm{k} \sim \sum_{\mathrm{k}=1}^{\mathrm{M}} \mathrm{p_k} \delta_{\psi_\mathrm{k}} \tag{21-d}$$

$$\ln(\mathrm{v_i}) = \frac{\omega_\mathrm{k} - \mu_\mathrm{DP}}{\sqrt{\mathrm{V_{DP}}}} \tag{21-e}$$

where μ_DP and $\mathrm{V_{DP}}$ are defined in Equations (3) and (4) respectively. Therefore,

$$\mathrm{median}(\mathrm{v}) \approx 1$$

Using the TCDP, not only the median of the mixed effect would be approximately to 1, but we also control the variance of the DP to provide a better convergence. Although the TCDP model has a nice interpretation and showed very good convergence properties, its implementation in WinBUGS is very time-consuming for large-scale datasets due to WinBUGS coding limitations.

Finally, one of defining characteristics of the DP is that it allows for ties in the observations as the DP is a discrete distribution almost surely. Consequently, during each iteration of the MCMC, the mixed effects are partitioned into clusters. This property of the DP is exploited to post-process MCMC samples to obtain clustering information (Medvedovic and Sivaganesan, 2002). The clustering information thus obtained can be used to gain further insights about the problem at hand (for example, which two sites are clustered together). In this regard, the NB-DP offers great opportunities for analyzing crash data in various different ways. Another utility of the clustering information is to detect outliers. For example, if one defines an outlier as belong to a cluster with no more two members in it, then in that regard, singleton clusters can be defined as outliers and can be inspected for potential risk factors. In fact, the notion of outlier can be handled much more formally, as is done in Heinzl and Tutz (2013). Indeed, a rich class of models exist in Bayesian nonparametric, such as the Product Partition Models, when inference on the partitions is of primary interest (Mitra and Muller, 2015).

3.4. Chapter Summary

In this chapter, the NB-DP was applied to two datasets that were characterized with a heavy tail and many zero observations. The results were compared with the NB and NB-L models. The results showed that the NB-DP offered much greater flexibility and a better fit compared to the NB model. Although the NB-L might work better with the dataset with many zeros, the NB-DP is actually more flexible to capture the dispersion in data, especially when the highly dispersed dataset has a heavy tail, but smaller percentage of zero observations.

As a closing note to this chapter, it must be noted that the primary goal in selecting a competitive model should not be based only on GoF measures. In addition to the GoF, the transportation safety analyst should examine other issues such as the data generating process, the relationship between variables and if the proposed model is logically or theoretically sound. The later characteristics are referred to as "Goodness-of-Logic" in Miaou and Lord (2003). The next chapter describes the characteristics of a methodology to design heuristics to select the most likely true sampling distribution to model crash datasets.

CHAPTER IV

MODEL SELECTION HEURISTICS: METHODOLOGY*

This chapter is divided into five subsections. First, a brief introduction about the motivation for developing characteristics-based heuristics is provided. Second, the proposed methodology is documented and its characteristics is described. Third, the Monte-Carlo Simulation task that is a key step in designing heuristics is discussed. In the fourth section, a few remarks about the proposed methodology are covered. In the fifth section, a brief summary of the proposed methodology is provided.

4.1. Introduction

Safety scientists usually use post-modeling methods, such as the Goodness-of-Fit (GoF) statistics or the Likelihood Ratio Test (LRT), to decide between two or more competitive distributions or models. Such metrics require all competitive distributions to be fitted to the data before any comparisons can be accomplished. Given the continuous growth in introducing new statistical distributions, choosing the best distribution using such post-modeling methods is not a trivial task, in addition to all theoretical or numerical issues the analyst may face during the analysis. Furthermore, and most importantly, these measures

* Part of this chapter is reprinted with permission from Shirazi, M., Dhavala, S. S., Lord, D., Geedipally, S. R. (2017). A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the negative binomial Lindley (NB-L) is preferred over the negative binomial (NB). Accident Analysis & Prevention, 107, 186-194. Copyright [2017] by Elsevier. https://doi.org/10.1016/j.aap.2017.07.002

or tests do not provide any intuitions into why a specific distribution (or model) is preferred over another or what is often referred to as Goodness-of-Logic (LoG) (Miaou and Lord, 2003).

In this chapter, a methodology is proposed to design heuristics for Model Selection based on the *characteristics* of data*,* in terms of descriptive summary statistics, *before* fitting the models. The proposed methodology employs two analytic tools: (1) Monte-Carlo Simulations and (2) Machine Learning Classifiers, to design simple heuristics to predict the label of the 'most-likely-true' distribution for analyzing data. Not only these heuristics are easy to use and do not need any post-modeling inputs, but also, using these heuristics, the analyst can attain useful information about why one distribution is preferred over another when modeling data.

4.2. Methodology

At the heart of the proposed methodology lies a paradigm shift in how Model Selection is both viewed and treated. We can view Model Selection as a classification problem - that is, given a set of discriminating features of the data, we like to predict a model that must have produced the observed data. It becomes a binary classification problem when the number of alternatives is two. This way of looking at Model Selection as a classification problem was first introduced, according to the authors' knowledge, by Pudlo et al. (2015), in the context of Approximate Bayesian Computation (ABC).

Learning both the discriminating function and its arguments have traditionally been based on GoF or other Model Selection criteria such as the LRT, Akaike Information

Criteria (AIC) and the likes. The discriminating function in such methods, which favor one model to the other, is often a simple comparator. A benefit of viewing the Model Selection as a classification problem is that we can take computational approach to learning a complex discriminating function based on simple descriptive statistics of the data.

To clarify the strategy, let us assume the analyst is interested in choosing between the Poisson and Negative binomial (NB) distributions, based on the population 'mean' and 'variance'. We like to come up with a function that maps these two statistics to a label: '0' for Poisson and '1' for NB. The choice of the labels is completely arbitrary. The 'mean' and 'variance' of the population would create a two-dimensional (a flat plane) predictor space (Ω) for making decisions. Now, the analyst's task is to partition the predictor space and assign a label to each partition. We know that if the population VMR is greater than one (VMR>1), we may choose the NB distribution and if it is equal to one (VMR=1), the Poisson distribution will be the preferred sampling distribution to use. Hence, the predictor space (Ω) can be classified between the Poisson and NB distributions in a way that is shown in Figure 3.

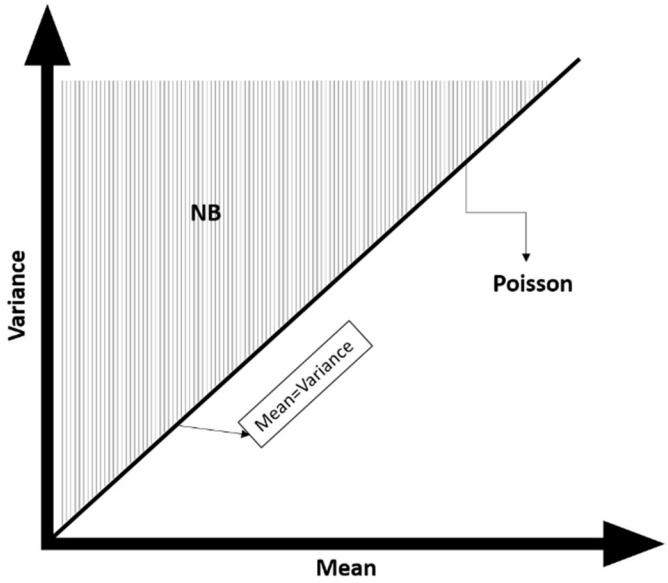

Figure 3. Classifying the Poisson and NB Distributions Based on the Mean and Variance of the Population (Reprinted with Permission from Shirazi et al., 2017b).

The decision based on the VMR statistic, in this case, serves as a heuristic to select the 'most-likely-true' sampling distribution between the Poisson and NB distributions. It does not require fitting the models, estimating the model parameters, computing the test statistics, etc. It simply uses the descriptive statistics to arrive at a model recommendation*. When working with data, the 'population' VMR essentially is replaced with its 'sample' counterpart ($\widehat{VMR}$) and the decision based on observed data will be essentially the analyst best guess. Like any Model-Selection decisions, there is a chance that the decision based on a sample version of the VMR may be incorrect; this uncertainty can be quantified in terms of standard classifier performance metrics, such as false-

* In Chapter V, it is shown that there are strong correlations between the decision based on the VMR heuristic and the LRT statistic.

positive-rate, Area under the Curve (AUC), and many others (Hastie et al., 2001; James et al., 2013).

In the case of 'Poisson' vs. 'NB', we knew, theoretically, how the two-dimensional predictor space should be partitioned between the Poisson and NB distributions; however, what if such insight was not available to us? In the absence of readily available analytical insights to guide Model Selection, we resort to computational approaches. It will be assumed that the distributions under consideration can be classified by 'm' summary statistics. These summary statistics would create an 'm-dimensional' predictor space; then, the analyst can benefit from two analytic tools, (1) Monte-Carlo Simulations, and (2) Machine Learning Classifiers, to partition the assumed m-dimensional predictor space between the competitive distributions.

Let us assume $\{A_1, A_2, \ldots, A_r\}$ and $\{S_1, S_2, \ldots, S_m\}$, respectively, denote a set of 'r' competitive distributions and 'm' types of summary statistics. We need to partition the m-dimensional predictor space that is created by the 'm' summary statistics, between all these 'r' distributions. Using Monte-Carlo Simulations, it is possible to simulate numerous datasets (say 100,000 datasets) from each of these 'r' distributions (or models) indexed by a label and record the assumed 'm' summary statistics for each. Next, a Machine Learning Classifier is trained to classify each simulated dataset to predict a model label. In the Machine Learning parlance, summary statistics are the features, the label (model) is the target. Each pair of the feature set and the target constitute a record. A Machine Learning Classifier learns a function that maps the features to a target, based on ground truth available in terms of the records.

There are several classifier methods, such as Logistic Regression, Support-Vector Machines, Decision Trees, Random Forests and many others (see Hastie et al., 2001; James et al., 2013) to accomplish the classification task. Decision Trees (DT) (Breiman et al., 1984) provide a very intuitive partitioning of the predictor space (similar to the one shown in Figure 3) but could be less accurate compared to, say, Random Forests (RF) (Breiman, 2001). A classifier in the context of this study, essentially, uses the simulation data to build a predictive tool (or heuristics) to estimate the label of the 'most-likely-true' distribution for each partition of the predictor space.

Let 'N' denote the number of datasets simulated from each distribution and 'n' denote the size of each dataset. Let $S_{A_j,i,m}$ denote the m-th summary statistic that was recorded for the i-th dataset simulated from the distribution A_j. The detailed steps of the proposed methodology are described below:

Step 1: Simulation- Preparation of Training Data

1.1 Define the experiment boundaries such that the simulated datasets reflect the characteristics of the data found in practice.

1.2 Repeat the following steps for 'N' iterations:

1.2.1 Simulate the parameters of all competitive distributions {A_1, A_2, …, and A_r} from a prior distribution.

1.2.2 Simulate a dataset of size 'n' from each competitive distribution within the experiment boundaries, given the parameters simulated in Step 1.2.1.

1.3 Compute and Record all the ‘m’ desired summary statistics for all datasets simulated in Step 1.2.

1.4 Outline the vector **Y** (distribution labels) and matrix **X** (summary statistics) as follows.

$$\mathbf{Y}=\begin{bmatrix}'A_1'\\ \vdots\\ 'A_1'\\ 'A_2'\\ \vdots\\ 'A_2'\\ \vdots\\ \vdots\\ \vdots\\ 'A_r'\\ \vdots\\ 'A_r'\end{bmatrix} \propto \mathbf{X}=\begin{bmatrix} S_{A_1,1,1} & S_{A_1,1,2} & S_{A_1,1,3} & \cdots & \cdots & \cdots & S_{A_1,1,m-1} & S_{A_1,1,m}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ S_{A_1,N,1} & S_{A_1,N,2} & S_{A_1,N,3} & \cdots & \cdots & \cdots & S_{A_1,N,m-1} & S_{A_1,N,m}\\ S_{A_2,1,1} & S_{A_2,1,2} & S_{A_2,1,3} & \cdots & \cdots & \cdots & S_{A_2,1,m-1} & S_{A_2,1,m}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ S_{A_2,N,1} & S_{A_2,N,2} & S_{A_2,N,3} & \cdots & \cdots & \cdots & S_{A_2,N,m-1} & S_{A_2,N,m}\\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots\\ S_{A_r,1,1} & S_{A_r,1,2} & S_{A_r,1,3} & \cdots & \cdots & \cdots & S_{A_r,1,m-1} & S_{A_r,1,m}\\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots\\ S_{A_r,N,1} & S_{A_r,N,2} & S_{A_r,N,3} & \cdots & \cdots & \cdots & S_{A_r,N,m-1} & S_{A_r,N,m}\end{bmatrix}$$

Step 2: Classification

Run a classifier method, such as a ‘Decision Tree’ or a ‘Random Forest’, over the summary statistics (matrix **X**) to classify the outcome−the distribution labels (vector **Y)**, i.e. partition the predictor space that is created by summary statistics among competitive distributions.

As a closing note to this section, it should be pointed that most of the metrics that summarize the performance of a classifier can be interpreted in the classical hypothesis-

testing parlance and can be used to measure the accuracy of the proposed heuristics. For example, false-positive-rate of a classifier is the type-1 error and true-positive rate is the power. In fact, we can obtain the Receiver-Operating-Characteristics (ROC) curves for the classifier and tune the classifier to obtain a desired power and type-1 error, where possible.

4.3. Simulation Design

The first task of the proposed methodology involves simulating numerous datasets from competitive distributions. This task requires designing an experiment that should represent the characteristics of the interested context; or in other words, addressing one of the most classic inferential questions in statistics: what is the target population? First, simulated data should represent the characteristics of the target population. For example, we know that the mean of crash data usually varies between 0.1 and 20; hence, the m-dimensional predictor space can be restricted to situations when the mean of the simulated data falls into that range. Second, the experiment should be designed in a way that competitive distributions have fair representations between simulated data. Sometimes, the fair simulation issue is easy to be addressed, perhaps just by simulating data using parameters that are selected from a Uniform distribution with the most common range seen in population. For example, we know that when modeling crash data with NB, the inverse dispersion parameter (ϕ) usually varies between 0.1 and 10; also, as noted earlier, we also know that the mean of crash data often varies between 0.1 and 20. Hence, we can use this

information and simulate data from NB for situations when $\phi \sim$Uniform [0.1,10] and $\mu \sim$Uniform [0.1,20].

However, in other practical situations, it may not be straightforward to generate representative datasets. In such cases, it may be far easier to generate/simulate datasets from a reference distribution that is easy to simulate from than from a target distribution that is hard to express in the generative stage, a strategy that is widely used in importance-sampling based statistical estimation techniques. To clarify this point, for a moment, let us assume a hypothetical modeling problem. Let us assume the analyst is interested in an experiment to measure the effect of some random factors, such as the effect of smoking, on causing a disease such as cancer. In this situation, he or she may want to account for factors, such as the population age, and needs to have certain coverage. In reality, as is true with many cohort-studies, the distribution of age and other factors may not be as per the design. In that case, there is a discrepancy between the sample and the target population. However, this can be easily addressed by up weighting or down weighting the samples in accordance with their representation in the target population. Importance Sampling is one such technique that is useful when the cost of obtaining data from target population is difficult or impossible compared to another source. Similar to this example, the experiment design issue in our case can also be expressed by ensuring that the controlled factors (such as the 'mean') are equally distributed over simulated datasets that are generated from all competitive distributions. In this case, the analyst seeks to discriminate the distributions based on other factors (such as 'skewness') when one or a

few factors (controlled factors such as 'mean') are equally distributed among competitive distributions.

Let S_c denote the vector of controlled factors in our experiment. The vector S_c may include summary statistics, such as the 'mean' or 'variance' of the data. Let $f^{trg}(S_c)$ denote the 'target' (or desired) density for the collected factors. Likewise, let $f_j^{obs}(S_c)$ denote the 'observed' multivariate empirical (or kernel) density of the controlled factors simulated from the j-th distribution. Then, the importance weight (W_j) of the simulated datasets can be expressed as:

$$W_j = \frac{f^{trg}(S_c)}{f_j^{obs}(S_c)}$$

Once the importance weights are estimated, they can be incorporated into the Classifier. Most Classifier packages in R have an option to pass importance weights, so that the importance of each dataset is altered in a way such that the controlled factors are distributed according to the target density between the competitive distributions. For that matter, any target distribution, not necessarily Uniform, so long as the support of f^{obs} is at least as large as f^{trg} can be used. In other words, the dataset importance for some datasets may be up weighed while for others it may be down weighted.

4.4. Discussion

The proposed methodology develops simple heuristics to select a model based on a few characteristics of the data, described in terms of the summary statistics, without the need to fit the models. This is accomplished by learning the patterns in the data that discriminate

one model with another. Key to this approach are (1) simulating datasets that closely represent the population under consideration and (2) using the simulated data to train a classifier that learns how to discriminate different models. The Model Selection was essentially treated as a classification problem. In fact, any Model Selection problems can be recast fundamentally as a classification problem and the label attached to a model is only notional. What is different though is the way in which classification is performed between in our proposed method and any Model Selection based on test statistics such as GoF, LRT and others.

If we look carefully, two components are involved in Model Selection: (1) a test statistic and (2) decision criteria (or a rule) that maps the test statistic to a model label. In the classical approach to Model Selection, say for example based on the Likelihood Ratio Tests, one computes the LRT test statistic and if the LRT is above a certain threshold, one chooses the alternative model as opposed to the null model. The statistic used to make the decision is a very complex function of data. It requires computing the log-likelihoods under both models, which requires fitting those models to the data in the first place but the decision rule is very simple. More often than not, the distribution of the test statistic is known analytically, and the errors incurred due to the decision rule can be quantified in terms of type-1 error and power. However, in this research, we are proposing a computational approach to the Model Selection problem, with the intent to flip the complexity of each of the two tasks involved in the decision-making problem. That is, we like to keep the test statistics as simple as possible that does not require estimating models, but the decision can be as complex as it needs to be. The advantage is that, one has the

ability to explain why one model fits better than the other, unlike omnibus test statistics such as those based on LRT or Walds' tests that do not provide any intuitions to the analyst.

Separating the Model Selection task into (a) training a classifier based on summary statistics and (b) scoring a new dataset to predict the model label has another benefit, in the context of Big Data and Data Science automation. Without really fitting models and then selecting the models, we simply learn the Model Selection patterns and use those patterns to score a new dataset based on simple computations. This is particularly useful when large volumes of high velocity data have to processed and appropriate modeling techniques have to be applied. According to our knowledge, this is a small but a very important step in enabling Data Science automation.

There is one more added advantage in such heuristics. When using classical tests or GoF statistics, not only the safety scientist should concern about the statistical fit but also about the model complexity. Many classical tests or GoF metrics do not consider complexity in their estimations and cannot be used when alternatives have different complexities. The proposed heuristics, however, can be employed even when the competitive models have different complexities. This is due to treating the Model Selection as a classification problem. Under this setting, model parameters are integrated out, and Model Selection will exclusively rely on classification probabilities.

It should be pointed out that in addition to all theoretical advantages, the proposed heuristics can also be useful as easy and straightforward Model-Selection guidelines based

on characteristics of data for safety practitioners. Such characteristics-based guidelines have recently been a subject of interest in several studies in safety literature. As such, recently, guidelines based on characteristics of data have been proposed for selecting a reliable calibration sample size (see Shirazi et al., 2016a; Shirazi et al., 2017a). These kinds of guidelines are useful in better use of data and modeling resources in practice.

4.5. Chapter Summary

In this chapter, a systematic methodology was proposed to develop Model Selection tools (or heuristics, to be exact) to select a sampling distribution among its competitors given an input from selected summary statistics of data, without a need to fit the models. Unlike the most common GoF measures or statistical tests, the proposed methodology addresses the classical issue of Goodness-of-Logic and examines the characteristics of data to find the 'most-likely-true' distribution for modeling. The next chapter presents the results of the application of the methodology to design heuristics for model selection between different distributions.

CHAPTER V

MODEL SELECTION HEURISTICS: APPLICATION*

This chapter is divided into four subsections. In the first part, the proposed methodology to design heuristics is validated by finding the switching points (i.e.: Model-Selection heuristics) between the Poisson and Negative Binomial distributions using a Decision Tree classifier. The results of this part are compared to the results with the theoretical expectations (the VMR heuristic). In the second part, the methodology is applied to design Model Selection heuristics between the Negative Binomial and the Poisson-lognormal distributions, using the Decision Tree and Random Forest classifiers. In the third part, the methodology is employed to find heuristics for Model Selection between the Negative Binomial and Negative Binomial Lindley distributions, using the Decision Tree and Random Forest classifiers. Last, a brief summary is provieded.

5.1. Poisson vs. NB Heuristics

The probability mass function (pmf) of the Poisson distribution is defined as follows:

$$\text{Poisson}(\lambda) \equiv P(Y = y|\,\lambda) = \frac{\lambda^y \times e^{-\lambda}}{y!} \tag{22}$$

* Part of this chapter is reprinted with permission from Shirazi, M., Dhavala, S. S., Lord, D., Geedipally, S. R. (2017). A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the negative binomial Lindley (NB-L) is preferred over the negative binomial (NB). Accident Analysis & Prevention, 107, 186-194. Copyright [2017] by Elsevier. https://doi.org/10.1016/j.aap.2017.07.002

where λ= the average number of events per interval. Note that $\lambda = \mu = \sigma^2$ where μ and σ^2 represent the mean and the variance of the observations, respectively.

As noted in Chapter II, the NB distribution is a mixture of the 'Poisson' and 'gamma' distributions. The pmf of the NB distribution is defined as follows:

$$NB(\phi, p) \equiv P(Y = y|\ \phi, p) = \frac{\Gamma(\phi + y)}{\Gamma(\phi)\Gamma(y+1)}(1-p)^{\phi}(p)^{y} \qquad (23)$$

where $p = \frac{\mu}{\mu+\phi}$, μ =mean response of observations, and ϕ = inverse dispersion parameter.

The experiment was designed for datasets that have a mean that is between 0.1 and 20. 100,000 datasets (N=100,000), each with 5,000 data points (n=5,000), were simulated from the Poisson and NB distributions. The following Uniform distributions were used to simulate the parameters of the Poisson and NB distributions.

$\mu\sim$ Uniform [0.1,20] ; for both Poisson and NB

$\phi\sim$Uniform [0.1,10] ; for NB only

For each simulated dataset, 22 summary statistics were recorded. The recorded summary statistics include the value of mean (μ), variance (σ^2), standard deviation (σ), variance-to-mean ratio (VMR), coefficient-of-variation (CV), skewness (skew)*, kurtosis† (K), percentage-of-zeros (Z), quantiles (Q) (or percentile) in 10% increments, the 10%,

* Skewness (skew) is the ratio of the third central moment (m_3) and standard deviation cubed (σ^3), i.e.: skew=$\frac{m_3}{\sigma^3}$

† Kurtosis (K) is the ratio of the fourth central moment (m_4) and the squared variance (σ^4), i.e.: K=$\frac{m_4}{\sigma^4}$

20%, 30% and 40% inter-quantiles (IQRs) (or inter-percentile), and the range (R). Next, a Decision Tree classifier was used to classify the 22-dimensional predictor space that is created by the given summary statistics between the Poisson and NB distributions. Figure 4 shows the results of the classification. As shown in this figure, the proposed heuristic is empirically found to be close to our theoretical expectations.

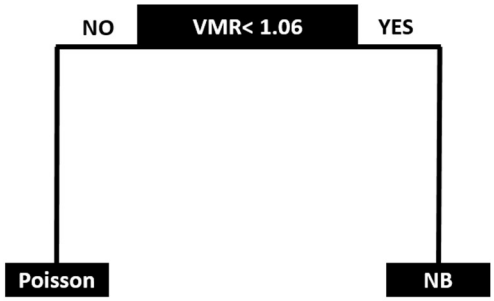

Figure 4. Heuristic for Model Selection between the Poisson and NB Distributions Using a Decision Tree Classifier (Reprinted with Permission from Shirazi et al., 2017b).

The classification problem between the Poisson and NB distributions can be seen in a binary-classification fashion. Let a dataset simulated from the NB distribution be labeled as a positive outcome (P), and a dataset simulated from the Poisson distribution as a negative outcome (N). This notation represents a test that indicates when the analyst should switch from a simple model (here 'Poisson') to a more complex model (here 'NB'). The prediction of the classifier can either be True (T) when the classifier correctly predicts the label of the model, or False (F) when the prediction is incorrect. Taking this notation into account, the confusion matrix for the results of the classification problem can be structured as shown in Table 5.

Table 5. Poisson vs. NB: Confusion Matrix Based on the Results of the Decision-Tree Classifier (Reprinted with Permission from Shirazi et al., 2017b).

Predicted	Actual	
	NB	Poisson
NB	49.46% (TP)	0.08% (FN)
Poisson	0.54% (FP)	49.92% (TN)

The sensitivity* and specificity† of the classification is equal to 99.8% and 98.9%, respectively. The overall misclassification error (FP+FN) is equal to 0.62%. A close analysis on misclassified datasets showed that misclassifications only were appeared at the boundary of the proposed heuristic when the value of the VMR is close to the threshold. No misclassifications are observed as the value of VMR deviates further away from the threshold.

The likelihood (or log-likelihood) ratio test reveals how likely data appear under the 'alternative' model than the 'null' model and is referred to the most powerful statistical test among its competitors, when some regularity conditions are met. If the value of log-likelihood ratio is greater than some threshold, the analyst can select the alternative model with a specific power and a type-1 error. Let us assume the Poisson distribution be the 'null' and the NB distribution be the 'alternative' hypothesis in constructing the log-

* Sensitivity=TP/(TP+FN)
† Specificity=TN/(TN+FP)

likelihood ratio test between these two distributions. The LRT statistic can be derived using Equation (24):

$$\text{LRT} = -2 \times \text{LN}\left(\frac{\text{Likelihood under the "Poisson" distribution}}{\text{Likelihood under the "NB" distribution}}\right) \quad (24)$$

As the value of the LRT statistic becomes larger, the analyst can reject the 'null' hypothesis (here 'Poisson') with a much greater power. Interestingly, one can see a strong correlation between the LRT statistic and the VMR heuristic. To clarify this point, the LRT statistic was plotted against the VMR, for 10,000 randomly simulated datasets from the NB distribution, and was shown in Figure 5.

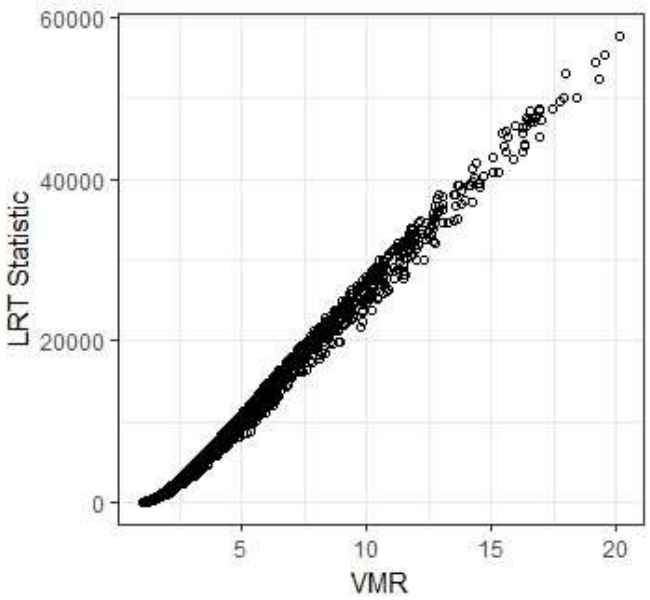

Figure 5. Poisson vs. NB: Correlation between the Decisions Based on the VMR and the LRT Statistic (Reprinted with Permission from Shirazi et al., 2017b).

Figure 5 indicates a strong correlation between the value of the VMR and the LRT statistic. In other words, the decision based on the value of the VMR heuristic closely follows the decision based on the LRT. In that regard, similar to log-likelihood test, as the

VMR gets further away from one, the analyst can reject the null model (here 'Poisson') with much greater confidence. This observation empirically establishes that VMR approximates LRT and that the approach to designing heuristics for Model Selection can reproduce well-established results.

5.2. NB vs. PLN Heuristics

This section is divided into three subsections. First: a brief background about the NB and PLN characteristics is provided. Second, Heuristics between NB and PLN are designed using the methodology described in Chapter IV. Third, the proposed heuristics are evaluated using observed data.

5.2.1. Background

Although both of the NB and PLN are appropriate when data express a sign of over dispersion, each of these distributions or models has its own positive and negative traits. As such, according to Lord and Mannering (2010), the PLN is more flexible than the NB to handle over dispersion and a better option for modeling skewed data. In a more detailed examination of these two alternatives, Khazraee (2016) states that the thick tail of the lognormal distribution, theoretically, can give the PLN a substantial boost when data are characterized by excessive large and/or unusual crash observations. The comparison of the NB and PLN models is not limited to the safety literature. In a research that was conducted to characterize the microbial counts in foods, Gonzaless-Barron and Butler (2011) showed that the PLN is a better alternative when data include observations with

large numbers, while the NB outperforms the PLN for data with small count observations, and/or those with larger amount of zero responses.

Overall, the previous studies indicate that the PLN is a better alternative for data with larger skewness, and/or data that involve large count observations but fewer zero responses, while the NB is a more suitable option for the opposite circumstances. However, it is not explicitly clear when the analyst may need to switch from the NB to the PLN - or vice versa- and/or what characteristics should be observed a priori to select a logical distribution between these two alternatives. This section addresses this topic and ponders into this issue by providing guidelines and tools (or heuristics, to be exact) to select a logical distribution between the NB and PLN distributions and recognizing the most important summary statistics to make a Model Selection decision between these two sampling distributions.

Both of the NB and PLN distributions are classified as a member of the Mixed-Poisson family distributions, where the Poisson parameter is mixed with a distribution to accommodate the over-dispersed data. The NB and PLN are two common models used to analyze crash data in safety literature (Lord and Mannering, 2010; Aguero-Valverde, and Jovanis, 2008; Lord and Miranda-Moreno, 2008; Aguero-Valverde, 2013).

As noted in Chapter II, the NB distribution can be structured as a mixture of the Poisson and gamma distributions as follows:

$$y|\lambda \sim \text{Poisson}(\lambda) \tag{25a}$$

$$\lambda|\mu,\phi \sim \text{gamma}(\phi, \frac{\phi}{\mu}) \tag{25b}$$

The mean (m), variance (VAR) and variance-to-mean ratio (VMR) of the NB distribution are defined as:

$$E(y) = m = \mu \tag{26a}$$

$$V(y) = VAR = \mu + \frac{\mu^2}{\phi} \tag{26b}$$

$$VMR(y) = VMR = 1 + \frac{\mu}{\phi} \tag{26c}$$

The PLN distribution is a mixture of the Poisson and lognormal distributions, which can be structured as the following hierarchical representation:

$$y|\,\lambda \sim \text{Poisson}(\lambda) \tag{27a}$$

$$\log(\lambda)|\upsilon, \sigma^2 \sim \text{Normal}(\upsilon, \sigma^2) \tag{27b}$$

Note that the mean (μ_λ) and variance (V_λ) of the lognormal distribution with parameters υ, σ^2 are equal to:

$$E(\lambda) = \mu_\lambda = e^{\upsilon+\sigma^2} \tag{28a}$$

$$\text{Var}(\lambda) = V_\lambda = \frac{e^{\sigma^2-1}}{e^{2\upsilon+\sigma^2}} \tag{28b}$$

Therefore, the mean (m), variance (VAR), and variance-to-mean ratio (VMR) of the PLN distribution are defined as:

$$E(y) = m = \mu_\lambda \tag{29a}$$

$$V(y) = VAR = \mu_\lambda + V_\lambda \tag{29b}$$

$$VMR(y) = VMR = 1 + \frac{V_\lambda}{\mu_\lambda} \tag{29c}$$

5.2.2. Heuristics Results

As noted in Chapter IV, simulation is a key step in designing Model Selection heuristics. It is essential to first make sure that the simulated datasets represent the characteristics of the target population, and then ensure that the alternative distributions have fair representations among simulated data. The first concern can be addressed by simulating data given the most common range observed in context population, in our case, the crash data population. The second concern can be addressed by ensuring that some summary statistics (referred to as control factors) are distributed similarly among the simulated datasets from alternative distributions (see Chapter IV). In other words, the analyst seeks to discriminate the distributions based on factors such as the 'kurtosis' and/or 'skewness', while the control factors such as the 'mean' or the 'VMR' are distributed similarly among simulated datasets.

For the problem (or simulation) design, it is assured that the 'mean' and the 'VMR' of data are uniformly distributed among the generated datasets from both of these distributions, simply, by simulating the mean (m) and the VMR from a uniform distribution with a range that is the most common observed range in crash data, as shown in Equation (30a) and Equation (30b).

$$\mathrm{m} \sim \mathrm{Uniform}(0.1,20) \tag{30a}$$

$$\mathrm{VMR} \sim \mathrm{Uniform}(1,25) \tag{30b}$$

Next, given the Equation (26a) and Equation (26c), the parameters of the NB distribution can be estimated as:

$$\mu = m \tag{31a}$$

$$\phi = \frac{m}{VMR - 1} \tag{31b}$$

Similarly, given the Equation (29a) and Equation (29c), first, we have:

$$\mu_\lambda = m \tag{32a}$$

$$V_\lambda = (VMR - 1) \times \mu_\lambda \tag{32b}$$

Then, given the Equation (28a) and Equation (28b), the parameters of the PLN distribution can be derived as:

$$\upsilon = \log\left(\frac{\mu_\lambda^2}{\sqrt{V_\lambda + \mu_\lambda^2}}\right) \tag{33a}$$

$$\sigma = \sqrt{\log\left(\frac{V_\lambda}{\mu_\lambda^2} + 1\right)} \tag{33b}$$

Now, it is possible to simulate a dataset with a size of n=5,000 from the NB distribution given parameters derived in Equation (31), and from the PLN distribution given the parameters derived in Equation (33). The above procedure can be repeated for N=100,000 iterations, for each one of these distributions. Each time, 22-types of summary statistics described in Section 5.1 was recorded. The detailed steps of the simulation protocol are described as follows:

Repeat the following steps for 'N' iterations:

1. Simulate the mean (m) and the VMR from the Equation (30a) and Equation (30b)

2. Find the parameters of the NB distribution from the Equation (31a) and Equation (31b) and the PLN distribution from Equation (33a) and Equation (33b).
3. Simulate a dataset with a size of 'n' given the parameters derived in Step 2, from both of the NB using Equation (25) and the PLN using Equation (27).
4. Record the 22 types of summary statistics described above.

A Decision Tree classifier was used as a tool to partition the 22-dimensional predictor space that is created by the simulated summary statistics, and assign a label (either the NB or the PLN) to each partition. Figure 6 shows the outcome of the Decision Tree classifier. As shown in Figure 6, the population kurtosis and the percentage-of-zeros play a substantial role in deciding between the NB and PLN distributions. As seen in this figure, overall, the PLN is recommended for situations when data are more skewed but has fewer zero responses, while the NB distribution is a better option otherwise; these results confirm the trends observed and/or reported in previous studies in the literature (see Lord and Mannering, 2010; Gonzaless-Barron and Butler, 2011 and khazraee, 2016). Unlike previous studies, however, Figure 6 provides a more perspicuous characteristics-based guidance on selecting a sampling distribution between these two alternatives.

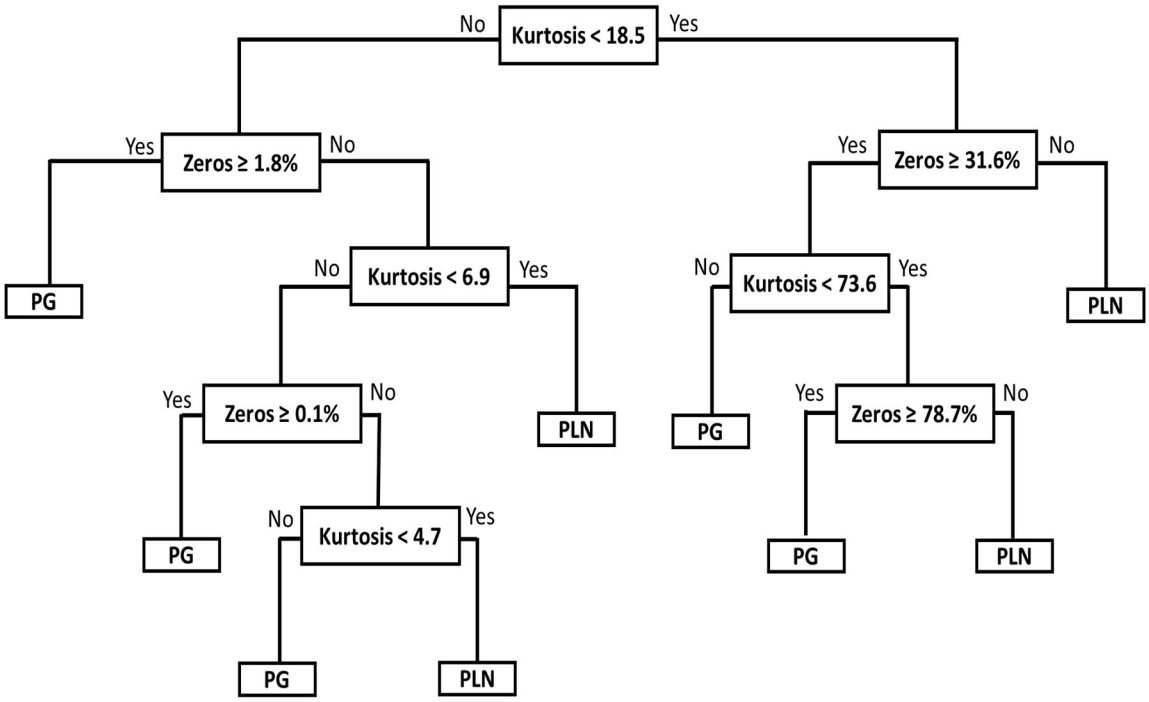

Figure 6. Heuristic for Model Selection between the NB and PLN Distributions (Note: tree can be used for data with the characteristics of 0.1 < mean < 20 and 1 < VMR < 25).

The output of a binary classifier can be either True (T) when it correctly classifies the label of the distribution, or False (F) when it misclassifies the label of the correct distribution. Let the PLN and NB distributions, respectively, be labeled as the positive (P) and negative (N) outputs of the binary classification. Such definitions represent a test when the analyst assumes the NB distribution as a base model, while he or she seeks to know when a shift to the PLN distribution is recommended. Table 6 shows the confusion matrix of the binary classification given such assumptions.

Table 6. NB vs. PLN: Confusion Matrix Based on the Results of the Decision Tree Classifier.

Predicted	Actual	
	PLN	NB
PLN	41.50% (TP)	1.18% (FN)
NB	8.50% (FP)	48.82% (TN)

The overall misclassification error is equal to 9.68% and the sensitivity and Specificity of the classification are equal to 97.24% and 85.12%, respectively. The sensitivity of the classification is very high indicating that when the outcome of the binary classifier is the PLN distribution, there is a high chance that the classifier has correctly detected the label of the distribution. However, the specificity of the classification is not as high as its sensitivity, meaning that when the outcome of the classifier is the NB distribution, there are still some chances that the output label was detected incorrectly. When the output of the classifier is the NB distribution, the analyst may consider other tests as well to decide between these two distributions and/or can decide to choose an alter tolerance threshold to decide between the NB and PLN.

Receiver-Operating-Characteristics (ROC) plots are another tool to evaluate the performance of a classifier (Hastie et al., 2001, James et al., 2013). The ROC plots are graphics that are used to display the performance of a binary classifier. The curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) by varying the discriminating threshold. The overall performance of a

classifier is measured by the area under the ROC curve which is referred to as AUC measure. We expect the AUC to be between 0.5 (an AUC=0.5 represents a decision that is made completely by chance like flipping a coin) to 1 (an AUC=1 represents a model with no misclassification errors). The greater the value of the AUC, the better the performance of the classifier. The ROC plot is shown in Figure 7 and the value of the AUC is equal to 0.93.

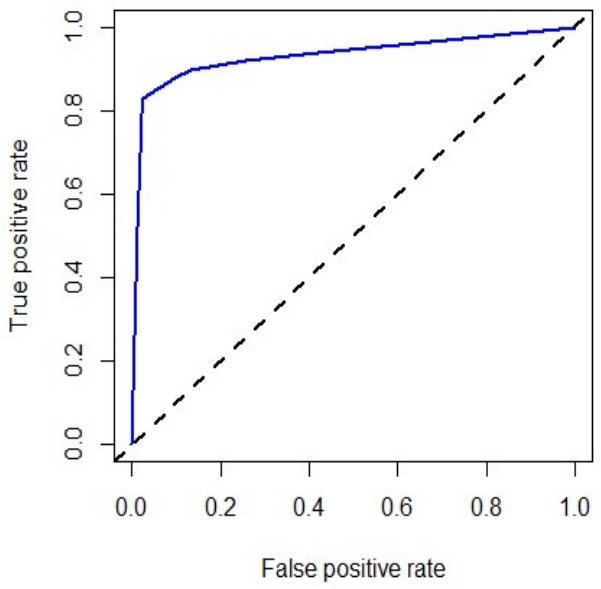

Figure 7. ROC Plot of the Classification between NB and PLN Based on the Decision Tree Results.

Although they are easy to interpret and use, decision trees may not be as accurate as other classifiers (say Random Forest) and can be non-robust (see Hastie et al., 2001, James et al., 2013). This means that a potential change in data could possibly result in altering in the final decision tree. The Random Forest classifier tries to overcome this issue

by building many trees, instead of one, to substantially improve the performance of the classification (see Hastie et al., 2001, James et al., 2013).

For the Random Forest classification, the number of trees was set to 100. Unlike the Decision Tree classification, the outcome of a Random Forest classification cannot be shown graphically. However, the trained forest can be recorded and still be used as an easy characteristics-based Model Selection tool to select a distribution between the NB and PLN distributions, without any post-modeling efforts. Table 7 shows the confusion matrix of the binary classification between the NB and PLN, based on the results of the Random Forest classifier. The misclassification error is equal to 0.01% (for trained data), and the sensitivity and specificity of the classifier are almost equal to 100%. The ROC plot is shown in Figure 8 and the value of the AUC is almost equal to 1. Although not reported here, the Random Forest heuristic was tested for simulated test data and the misclassification error was less than 1.5% for the test data.

Table 7. NB vs. PLN: Confusion Matrix Based on the Results of the Random Forest Classifier.

Predicted	**Actual**	
	PLN	NB
PLN	50.00% (TP)	0.01% (FN)
NB	0.00% (FP)	49.99% (TN)

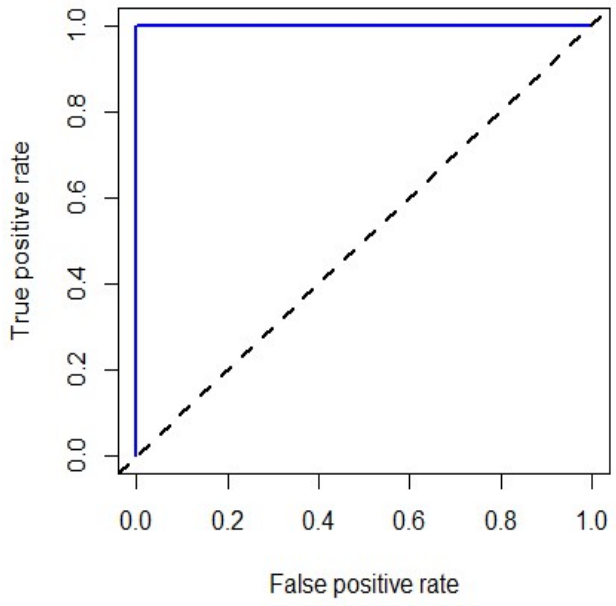

Figure 8. ROC Plot of the Classification between the NB and PLN Based the Random Forest Results.

As a by-product of the Random Forest classifier, the predictors (summary statistics) can be ranked by their importance. Figure 9 and Figure 10 show the importance of the summary statistics based on two criteria: (1) mean decrease Deviance Accuracy and (2) mean decrease Gini index (Hastie et al., 2001; James et al., 2013). As shown in these figures, kurtosis, skewness and the percentage-of-zeros are among the most important summary statistics to select a model between the NB and PLN distributions.

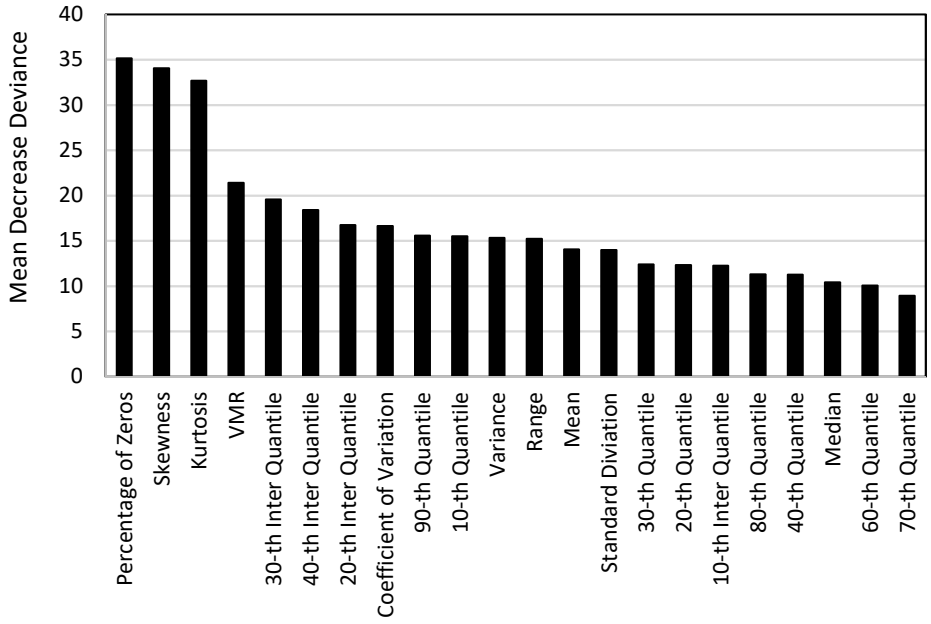

Figure 9. Importance of the Summary Statistics to Select a Distribution between the NB and PLN Based on the Mean Decrease Deviance Accuracy Given the Results of the Random Forest Classifier.

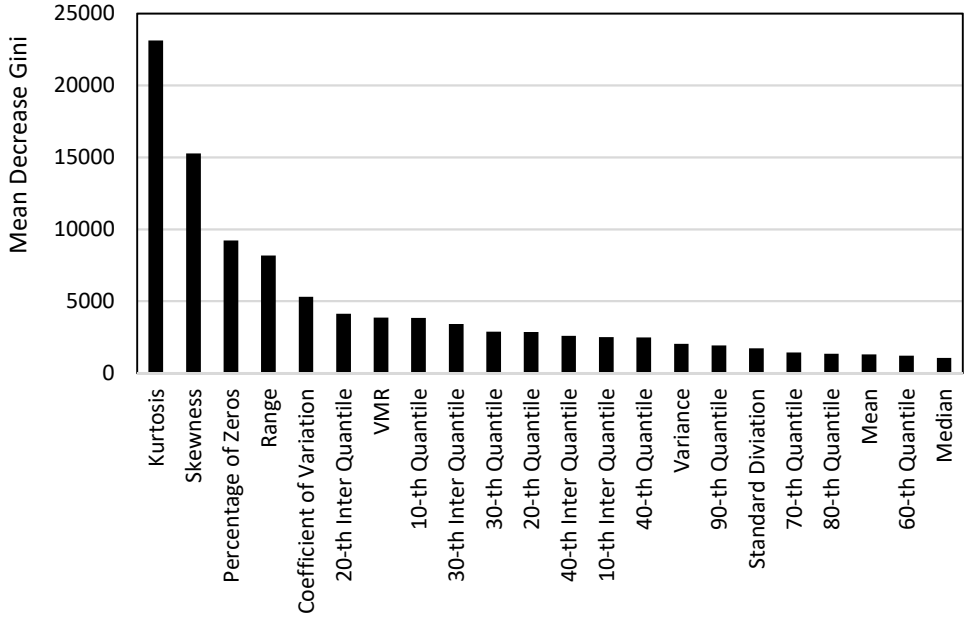

Figure 10. Importance of the Summary Statistics to Select a Distribution between the NB and PLN Based on the Mean Decrease Gini, Given the Results of the Random Forest Classifier.

5.2.3. Evaluation with Observed Data

In this section, two datasets are used to evaluate the proposed heuristics. The first dataset includes information related to single-vehicle crashes that occurred on Michigan rural two-lane highway in 2006 used in Chapter III. As noted before, this dataset was utilized in several previous studies (Qin et al, 2004; Geedipally et al., 2012; Shirazi et al, 2016b). The dataset includes 33,970 segments, and the mean, variance, VMR, kurtosis, and the percentage-of-zeros of data are equal to: 0.68, 3.15, 4.62, 123.6 and 69.7%, respectively. The second dataset contains crash data that occurred between 2012 and 2014 on Texas urban four-lane arterials. This dataset also has been used in several studies (Lord et al., 2016; Geedipally et al., 2017) in the past. The dataset includes 4,264 segments, and the mean, variance, VMR, kurtosis, and the percentage-of-zeros of data are equal to: 2.26, 45.53, 19.27, 92.8 and 56.5%, respectively. The detailed summary statistics of the two datasets are shown in Table 8.

Table 8. Summary Statistics of the Datasets Used to Evaluate the NB vs. PLN Heuristics.

Summary Statistics	Michigan Dataset	Texas Dataset
Mean	0.68	2.36
Variance	3.15	45.53
Standard Deviation (Sd.)	1.77	6.75
Variance-to-Mean-Ratio (VMR)	4.62	19.27
Coefficient-of-Variation (CV)	2.60	2.86
skewness (skew)	7.76	7.92
kurtosis (K)	123.59	92.67
Percentage-of-Zeros (Z)	69.6%	56.5%
10% Quantile	0	0
20% Quantile	0	0
30% Quantile	0	0
40% Quantile	0	0
50% Quantile (Median)	0	0
60% Quantile	0	1
70% Quantile	1	1
80% Quantile	1	3
90% Quantile	2	6
10% Inter-Quantile	1	1
20% Inter-Quantile	1	1
30% Inter-Quantile	1	3
40% Inter-Quantile	2	6
Range	61	120

Table 9 and Table 10, respectively, show the recommended models for the Michigan and Texas data based on the proposed heuristics and the log-likelihood metric. While the classical metrics require the distributions to be fitted to the data before coming up with the model recommendation, the proposed heuristics can be used without any post-modeling inputs and/or efforts. The decision based on the proposed heuristics solely rely on characteristics of data. For both datasets, the PLN distribution is the favored distribution to model data, based on the classical log-likelihood metric and the proposed heuristics. Classical metrics, such as the log-likelihood, do not give any intuitions into why the PLN is preferred to the NB (addressing the Goodness-of-Logic issue). On the other hand, the proposed heuristics come up with the model recommendation by considering the characteristics of data; hence, in this case, the analyst can select a logical distribution to model data. For example, a large kurtosis value in both datasets plays a substantial role in choosing the PLN over the NB.

Table 9. Model Selection for the Michigan Data Based on the Classical Statistical Tests and Proposed Heuristics.

Method	NB	PLN	Criteria	Favored Distribution
Log-Likelihood (LL)[1]	-36332.85	-36117.54	$LL_{PLN} > LL_{NB}$	PLN
Decision Tree Heuristic [2]	kurtosis= 123.6 zeros=69.7%		kurtosis > 73.6 zeros < 78.7%	PLN
Random Forest Heuristic[2]	Using All 22 Summary Statistics		Using the RF Heuristic	PLN

[1]Requires fitting the distributions.
[2] Do not require fitting the distributions.

Table 10. Model Selection for the Texas Data Based on the Classical Statistical Tests and Proposed Heuristics.

Method	NB	PLN	Criteria	Favored Distribution
Log-Likelihood (LL)[1]	-7462.91	-7432.35	$LL_{PLN} > LL_{NB}$	PLN
Decision Tree Heuristic [2]	kurtosis= 92.8 zeros= 56.5%		kurtosis > 73.6 zeros < 78.7%	PLN
Random Forest Heuristic[2]	Using All 22 Summary Statistics		Using the RF Heuristic	PLN

[1]Requires fitting the distributions.
[2]Do not require fitting the distributions.

5.3. NB vs. NB-L Heuristics

This section is divided into three subsections. First, the characteristics of the NB-L distribution is briefly reviewed. Second, Heuristics to select a sampling distribution between the NB and NB-L are designed using the methodology described in Chapter IV. Third, the proposed heuristics are evaluated using observed data.

5.3.1. Background

The NB-L GLM was introduced in Chapter II. Here a brief review of the NB-L distribution is provided. The pdf of the Lindley distribution (Lindley, 1958) is defined as:

$$\text{Lindley}(v|\theta) = \frac{\theta^2}{\theta+1}(1+v)e^{-\theta v} \qquad \theta > 0, v > 0 \tag{34}$$

The random variable y is distributed by the NB-L (ϕ, θ) distribution if (Zamani and Ismail, 2010; Lord and Geedipally, 2011):

$$y \sim NB(\phi, p = 1 - e^{-\lambda}) \text{ and } \lambda \sim Lindley(\theta) \quad (35)$$

The Lindley distribution, in fact, is a mixture of two gamma distributions as follows:

$$\lambda \sim \frac{1}{1+\theta} gamma(2, \theta) + \frac{\theta}{1+\theta} gamma(1, \theta) \quad (36)$$

Therefore, the NB-L distribution can be written in following hierarchical representation:

$$y \sim NB\left(y \middle| \phi, p = 1 - e^{-\lambda}\right) \quad (37\text{-a})$$

$$\lambda \sim gamma(1 + z, \theta) \quad (37\text{-b})$$

$$z \sim Bernoulli\left(\frac{1}{1+\theta}\right) \quad (37\text{-c})$$

The mean of the NB-L distribution is equal to (Zamani and Ismail, 2010):

$$\mu = \phi\left(\frac{\theta^3}{(\theta+1)(\theta-1)^2} - 1\right) \quad (38)$$

Lord and Geedipally (2011) showed that the NB-L distribution performs better than the NB distribution when data have many zeros or characterized by a heavy (or long) tail. However, it is not clear, at what point the NB-L distribution should be used instead of the NB distribution. In this section, we use the methodology described in Chapter IV to design Model Selection heuristics to select the 'most likely true' distribution for modeling crash data between these two distributions.

5.3.2. Heuristics Results

The experiment (or simulation boundaries) was designed for datasets with the following range for the 'mean' and 'VMR' of the population that is the most common

range observed in crash data. The mean of crash data was assumed to varies from 0.1 to 20 and its VMR from 1 to 100, as follows:

$$0.1 < \text{mean} < 20$$

$$1 < \text{VMR} < 100$$

100,000 datasets (N=100,000), each with 5,000 data points (n=5,000), were simulated from the NB and NB-L distributions. The following Uniform distributions were used to simulate the NB and NB-L parameters at each iteration of the simulation:

$$\mu \sim \text{Uniform}(0.1, 20); \text{ for both NB and NB-L}$$

$$\frac{1}{1+\theta} \sim \text{Uniform}(0,0.5)^{*}; \text{ for NB-L}$$

$$\phi \sim \text{Uniform}(0.1,10); \text{ for NB}$$

By simulating the mean of the NB and NB-L distributions from a Uniform distribution, we guarantee that the distribution of the 'mean' of the simulated datasets generated from both these distributions is uniformly distributed. For each simulated dataset the same 22 summary statistics described in Section 5.1 were recorded.

Two classifier methods are used in this section to partition the predictor space into regions that are most likely to be covered by either the NB or NB-L distributions. First, the Decision-Tree classifier is used for a simple and easy to interpret but less accurate

---

[*] Note that for situations when the value of θ is smaller than or close to 1, simulation from the NB-L distribution would face some numerical problems and the NB-L random variable simulator may produce data with an infinite value. The range of the Uniform distribution for simulating the $\frac{1}{1+\theta}$ parameter was chosen in way that would avoid such numerical difficulties.

classification. Figure 11 shows the results of applying the Decision-Tree method to partition the 22-dimensional predictor space between the NB and NB-L distributions. Out of 22 summary statistics used for the analysis, only the 'skewness' of the population was used by classifier in the decision tree to separate the NB-L distribution from the NB[*]. As shown in Figure 11, the tree involves only one splitting rule. Starting at the top of the tree, it is divided into two sections based on the value of 'skewness'. The observations that have a 'skewness' of less than 1.92 are assigned to the left branch and the 'NB' label is assigned to them. On the other hand, when the value of the 'skewness' is greater than 1.92, the NB-L distribution is recommended to be used.

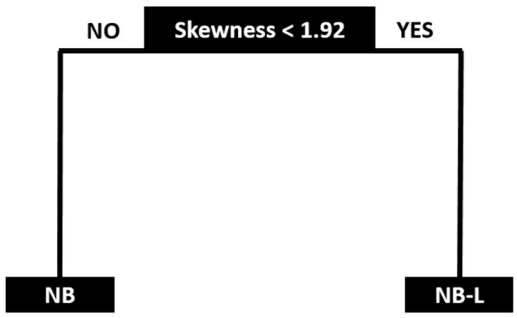

Figure 11. Heuristic for Model Selection between the NB and NB-L Distributions (Note: tree can be used for data with 0.1 < mean< 20 and 1 < VMR< 100) (Reprinted with Permission from Shirazi et al., 2017b).

[*] The skewness (20), kurtosis (19), CV (18), percentage-of-zeros (15), and VMR (14), respectively, were found to be the most important predictors to classify the 22-dimensional predictor space between the NB and NB-L distributions (Note: the number in parenthesis denotes the importance rate); However, the 'skewness' of the population was the only variable used by the classifier in the decision tree.

The classification between the NB and NB-L distributions can be seen in a binary-classification fashion. The confusion matrix for the results of the classification problem can be structured as shown in Table 11. The overall misclassification error (FP+FN) is equal to 5.90%. The value of the sensitivity and specificity of the classification is equal to 89.96% and 99.21%, respectively. The ROC curve based on the results of this classifier is shown in Figure 12. The value of the AUC is equal to 0.941.

Table 11. NB vs. NB-L: Confusion Matrix Based on the Results of the Decision-Tree Classifier (Reprinted with Permission from Shirazi et al., 2017b).

Predicted	**Actual**	
	NB-L	NB
NB-L	49.64% (TP)	5.54% (FN)
NB	0.36% (FP)	44.46% (TN)

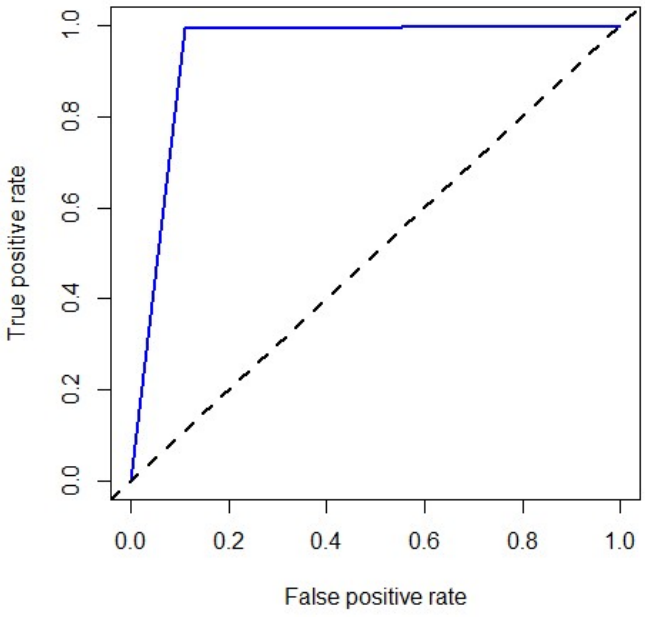

Figure 12. NB vs. NB-L: ROC Plot Based on the Results of the Decision-Tree Classifier (Reprinted with Permission from Shirazi et al., 2017b).

As noted in Section 5.2.2, although it is simple and easy to interpret or use, there are some drawbacks with the simple Decision-Tree method. Trees can be very non-robust; i.e., a change in the data can cause a large change in the final estimated tree (James et al., 2013). This issue, however, can be overcome substantially by aggregating over many decision trees instead of contracting only one, using methods like Random Forest. The Random-Forest classifier improves the performance of the simple Decision-Tree method by applying two tricks (James et al., 2013): (1) instead of using one decision tree, the Random Forest method aggregates the results of fitting 'n trees' from 'n bootstraps' of the

training data; (2) instead of using all ‘m’ predictors, only ‘p’ predictors (usually $p=\sqrt{m}$) is used at a time to form each decision tree.

The Random-Forest classifier was trained over the simulated summary statistics to partition the 22-dimensional predictor space. The number of trees in the Random-Forest method was set to 100 trees. The importance of the predictors, i.e., the importance of each summary statistics to predict the model label between the NB and NB-L distributions, was measured based on their effect in mean-decrease of two criteria: (1) Gini Index, and (2) Deviance accuracy (Hastie et al., 2001; James et al., 2013). Table 12 shows the importance of the predictors (summary statistics) to partition the 22-dimensional predictor space between the NB and NB-L distributions, based on these two criteria. Figure 13 and Figure 14 show the importance of summary statistics graphically. skewness, CV, kurtosis, VMR, and percentage-of-zeros were the top 5 predictors that decrease the Gini index the most, while skewness, kurtosis, percentage-of-zeros, 40% inter-quantile, and VMR were the top 5 most important predictors to decrease the value of the Deviance accuracy.

Table 12. NB vs. NB-L: Importance of the Predictors (Summary Statistics) in Partitioning the Predictor Space Based on the Results of the Random Forest Classifier (Reprinted with Permission from Shirazi et al., 2017b).

Predictor (Summary Statistics)[1]	Mean-Decrease Gini	Mean-Decrease Deviance
Skewness (skew)	22022.1	22.3
Coefficient-of-Variation (CV)	17958.2	15.7
kurtosis (K)	16531.2	21.5
Variance-to-Mean-Ratio (VMR)	10470.8	16.9
Percentage-of-Zeros (Z)	6759.7	20.6
10% Quantile	4750.5	10.2
Range	3913.5	10.3
20% Quantile	3337.5	11.8
Standard Deviation (Sd.)	2142.0	14.7
Variance	1866.7	14.6
40% Inter-Quantile	1710.8	18.5
90% Quantile	1305.3	15.9
30% Inter-Quantile	1150.1	13.7
30% Quantile	1109.7	8.9
40% Quantile	1041.7	8.5
Mean	879.4	13.0
80% Quantile	740.4	11.7
20% Inter-Quantile	592.3	13.2
50% Quantile (Median)	420.6	8.1
60% Quantile	378.8	7.7
70% Quantile	367.5	8.0
10% Inter-Quantile	310.5	8.8

[1] Predictors were sorted based on Mean-Decrease Gini criteria

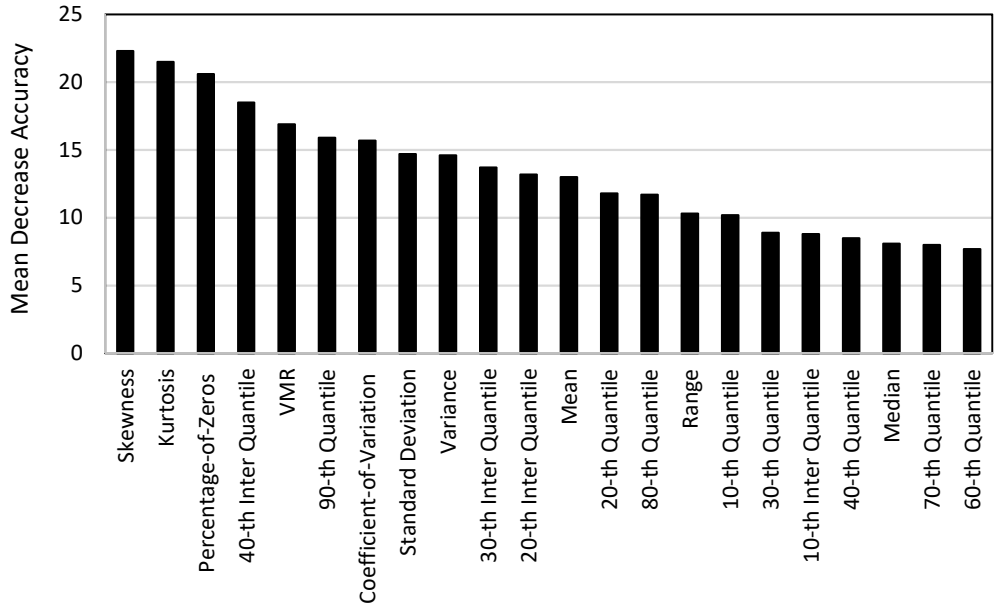

Figure 13: Importance of Summary Statistics to Select a Distribution between the NB and NB-L Based on the Mean Decrease Deviance Accuracy Given the Results of the Random Forest Classifier.

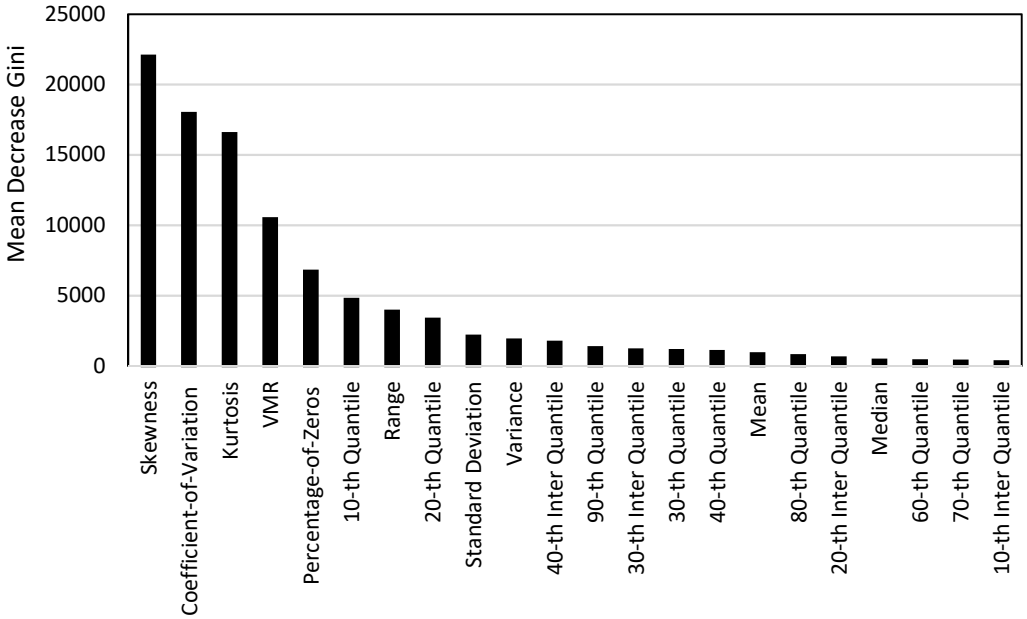

Figure 14: Importance of Summary Statistics to Select a Distribution between the NB and NB-L Based on the Mean Decrease Gini Index Given the Results of the Random Forest Classifier.

Unlike the Decision-Tree classifier, the results of the Random-Forest classifier cannot be shown graphically. However, the trained forest can be saved, and employed as a simple and convenient heuristic tool to predict the model label. This is referred to as the RF heuristic tool in this research. The confusion matrix for the results of the Random-Forest classification is shown in Table 13. The overall misclassification error (FP+FN) is equal to 0.04%. The value of the sensitivity and specificity of the classification is equal to 99.9% and 100%, respectively. Both the sensitivity and specificity of the classification are high and the proposed tool can detect the 'most-likely-true' distribution between the NB and NB-L distributions with a good precision. The ROC plot based on the results of the Random-Forest classifier is shown in Figure 15. The value of the AUC is equal to 0.999.

Table 13. NB vs. NB-L: Confusion Matrix Based on the Results of the Random-Forest Classifier (Reprinted with Permission from Shirazi et al., 2017b).

Predicted	**Actual**	
	NB-L	NB
NB-L	50.00% (TP)	0.04% (FN)
NB	0.00% (FP)	49.96 % (TN)

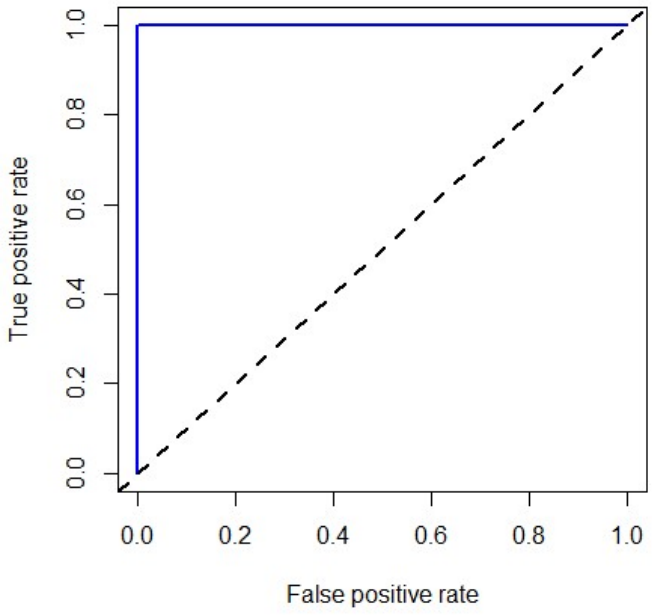

Figure 15. NB vs. NB-L: ROC Plot Based on the Results of the Random-Forest Classifier (Reprinted with Permission from Shirazi et al., 2017b).

5.3.3. Evaluation with Observed Data

The main goal of this section involves comparing the results of the Model Selection based on our proposed heuristics against the Model Selection based on traditional Test Statistics. Three datasets were used to accomplish this objective. The first dataset includes the single-vehicle fatal crashes that occurred on 1,721 divided multi-lane rural highway segments between 1997 and 2001 in Texas. The second dataset involves single-vehicle roadway departure fatal crashes that occurred on 32,672 rural two-lane horizontal curves between 2003 and 2008 in Texas. These two datasets were previously used in Lord and Geedipally (2011) to compare the NB and NB-L distributions for data with excess number of zero

responses. The third dataset involve crash data collected in 1995 at 868 four-legged signalized intersections located in Toronto, Ontario; this dataset has extensively been used in other research studies (see, Miaou and Lord, 2003; Lord et al., 2008; Lord et al., 2016). Table 14 shows the summary statistics of these datasets.

Table 14. Summary Statistics of the Datasets Used to Evaluate NB vs. NB-L Heuristics (Reprinted with Permission from Shirazi et al., 2017b).

Summary Statistics	Texas Rural Divided Multi-Lane Highway	Texas Rural Two-Lane Horizontal Curves	Toronto Four-Legged signalized Intersections
Mean	0.131	0.138	11.555
Variance	0.171	0.204	100.363
Standard Deviation (Sd.)	0.414	0.452	10.012
Variance-to-Mean-Ratio (VMR)	1.303	1.458	8.685
Coefficient-of-Variation (CV)	3.149	3.258	0.866
Skewness (skew)	3.981	5.120	1.499
kurtosis (K)	20.481	45.255	2.312
Percentage-of-Zeros (Z)	89%	89%	1.84%
10% Quantile	0	0	2
20% Quantile	0	0	4
30% Quantile	0	0	5
40% Quantile	0	0	7
50% Quantile (Median)	0	0	8
60% Quantile	0	0	11
70% Quantile	0	0	14
80% Quantile	0	0	19
90% Quantile	1	1	25
10% Inter-Quantile	0	0	4
20% Inter-Quantile	0	0	10
30% Inter-Quantile	0	0	14
40% Inter-Quantile	1	1	23
Range	4	10	54

Table 15, Table 16 and Table 17 show the Model Selection results based on the classical tests and our proposed heuristics. To estimate the Chi-square and log-likelihood, data should be fitted to both NB and NB-L distributions. The proposed heuristics, on the other hand, can be used simply *before* fitting the distributions, based on inputs from *characteristics* of data. As shown in Table 15 and Table 16, both classical tests and proposed heuristics favor the NB-L distribution to model the Texas datasets. On the other hand, as shown in Table 17, for the Toronto dataset, the NB distribution is the favored distribution between these two options.

Table 15. Model Selection for the Texas Divided Multi-Lane Rural Highway Segments Data Based on the Classical Statistical Tests and Proposed Heuristics (Reprinted with Permission from Shirazi et al., 2017b).

Method	NB	NB-L	Criteria	Favored Distribution
Chi-Square (χ^2)[1]	2.73	1.68	$\chi^2_{NB-L} < \chi^2_{NB}$	NB-L
Log-Likelihood (LL)[1]	-696.1	-695.1	$LL_{NB-L} > LL_{NB}$	NB-L
DT Heuristic[2]	-		skewness>1.92	NB-L
RF Heuristic[2]	-		Using the RF Heuristic Tool	NB-L

[1]Requires fitting the distributions.
[2] Do not require fitting the distributions.

Table 16. Model Selection for the Texas Rural Two-Lane Horizontal Curves Data Based on the Statistical Tests and Proposed Heuristics (Reprinted with Permission from Shirazi et al., 2017b).

Method	NB	NB-L	Criteria	Favored Distribution
Chi-Square (χ^2)[1]	57.47	11.68	$\chi^2_{NB-L} < \chi^2_{NB}$	NB-L
Log-Likelihood (LL)[1]	-13,557.7	-13,529.8	$LL_{NB-L} > LL_{NB}$	NB-L
DT Heuristic[2]	-		skewness>1.92	NB-L
RF Heuristic[2]	-		Using the RF Heuristic Tool	NB-L

[1]Requires fitting the distributions.
[2] Do not require fitting the distributions.

Table 17. Model Selection for the Toronto Four-Legged Signalized Intersections Data Based on the Statistical Tests and Proposed Heuristics (Reprinted with Permission from Shirazi et al., 2017b).

Method	NB	NB-L	Criteria	Favored Distribution
Chi-Square (χ^2)[1]	74.86	615.68	$\chi^2_{NB-L} > \chi^2_{NB}$	NB
Log-Likelihood (LL)[1]	-2,988.825	-3,291.933	$LL_{NB-L} < LL_{NB}$	NB
DT Heuristic[2]	-		skewness<1.92	NB
RF Heuristic[2]	-		Using the RF Heuristic Tool	NB

[1]Requires fitting the distributions.
[2]Do not require fitting the distributions.

Unlike the classical tests that do not provide any intuitions into why a specific distribution is favored to the other, using the proposed heuristics, the analyst can select a distribution that is most suitable based on the characteristics of data, reflected into the descriptive summary statistics. For instance, the value of the skewness plays an important role to select the NB-L distribution for the two Texas datasets (large skewness) and the NB distribution for the Toronto data (small skewness).

5.4. Chapter Summary

This chapter documented the application of the methodology described in Chapter IV to investigate under what circumstances the PLN is preferred over the NB, and vice versa, based on characteristics of data, reflected in the summary statistics. A decision tree was constructed and proposed as simple heuristics to select a distribution between these two alternatives. The kurtosis and percentage-of-zeros were the only summary statistics used by the classifier in the decision tree. Although Decision Tree classifiers are non-robust and potentially provide different tree splits, the results shown in Figure 6 can be used by practitioners as useful guidelines for selecting a "most-likely-true" sampling distribution between the NB and PLN. A Random Forest classifier was used to design a more accurate tool to select a distribution between these two options. As a by-product of a Random Forest classifier, the summary statistics can be ranked by their importance. Among the 22 types of summary statistics used in the analysis, kurtosis, skewness and the percentage-of-zeros were found the most important and critical summary statistics to select a model between the NB and PLN.

Next, the methodology was applied to propose heuristics to select the 'most-likely-true' distribution between the NB and NB-L distributions. First, a Decision-Tree classifier was employed to design a simple decision tree to choose between the NB and NB-L distributions. The skewness of data was the only predictor used by the classifier in the decision tree among all the 22 summary statistics that were included in the analysis to distinguish these two distributions. Next, a Random-Forest classifier was applied to design a more accurate Model Selection tool (or heuristics). skewness, CV, kurtosis, VMR, and percentage-of-zeros were among the most important summary statistics needed to choose between the NB and NB-L distributions, based on the results of the Random-Forest classifier. The next chapter documents the highlights of the research accomplished in this work and provides avenues for further research.

CHAPTER VI

SUMMARY AND FUTURE RESEARCH AVENUES

This dissertation contributed to the crash data modeling by (1) documenting characteristics of a flexible model using a mixture of the NB and a random distribution characterized by Dirichlet process, and (2) Proposing a methodology to design characteristics-based heuristics to select a sampling distribution between potential alternatives. This chapter is divided into two parts. First, the dissertation effort is summarized and the key findings are documented and discussed. Second, a few avenues for further research are explained.

6.1. Dissertation Summary

Chapter II documented the characteristics of the NB-DP (or NB-TDP to be exact) GLM framework for analyzing count/crash data. As noted in Chapter II, the recurring theme in most statistical models to analyze count/crash data include considering a mixing distribution at the heart of the generative model to obtain a greater degree of flexibility. The shape of the mixing distribution, the mixture weights, and the level that the hierarchical model is constructed are the three major ingredients used by statisticians to provide flexibility in modeling. In most mixture models, the analyst have certain assertions about the mixture ingredients. Using a random mixing distribution, however, is one way to incorporate flexibility in modeling while not overly concerned about characteristics of the mixing distribution. Dirichlet process (DP), a widely used prior in Bayesian nonparametric (semiparametric), allows such representation. Each draw from the DP is a

random distribution itself. Hence, using DP, instead of being constrained to a particular shape or distribution, a random distribution will be used at the heart of a generative model.

The proposed NB-DP model can be thought in context of the Bayesian hierarchical modeling framework, where the mixed effects in NB GLM are given a flexible distribution that follows the Dirichlet process. The NB-DP model allows a greater degree of flexibility to the model to capture the variation in the data as well as handling issues with datasets that are characterized by a long tail and/or include many zero observations. In addition to a greater flexibility, there is one more added advantage to the NB-DP (or NB-TDP to be exact). While modeling data, the NB-DP model partitions the data points into finite number of clusters. The clustering information provides further insights about the domain or data. As such, the safety scientist can obtain a better understanding about the unobserved variables, identify safety issues or decide on countermeasures.

In Chapter III, the NB-DP was applied to study two observed datasets, one collected in Indiana and the other one in Michigan. Both datasets were characterized with a long (or heavy) tail. In addition, about 36% of the locations in the Indiana dataset, and 70% of locations in the Michigan dataset did not experience any crash. The NB-DP GLM* was applied to the both datasets, and the modeling results were compared with the results obtained from the NB and NB-L GLMs. The modeling results indicated that the NB-DP offers a greater flexibility and a better fit compared to the NB model. The DIC value for the NB-DP model was better than the NB-L model when the models were used to fit the

* The model applied using a lognormal distribution for the DP base distribution.

Indiana data, while the DIC of the NB-L outperforms the NB-DP for the Michigan dataset. It was concluded that while the NB-L may work better with datasets with many zero observations, the NB-DP is more flexible to capture the dispersion in data, especially when the highly dispersed dataset is characterized by a long tail, but smaller percentage of zero observations. However, still further research is needed to better examine the NB-DP and NB-L using various other datasets. In addition, the NB-L and NB-DP should be examined when other distributions are considered instead of the DP base distribution to conclude a better comparison between the NB-L and NB-DP under different scenarios.

Chapter IV documented a novel approach to design characteristics-based heuristics to select a sampling distribution among competitive alternatives given a few selected summary statistics of data. Using this method, the Model Selection problem is treated as a classification problem. The keys to this approach are (1) simulating datasets that closely represent the population under consideration and recording the summary statistics of each dataset, and (2) training a classifier over the summary statistics to learn the patterns in the data to discriminate one distribution from another. The proposed heuristics, once designed, can come up with the model recommendation without any post modeling inputs. In addition, unlike the most common GoF statistics or statistical tests, the designed heuristics can address the classical issue of Goodness-of-Logic. In summary, the proposed heuristics have the following key characteristics:

- Unlike the Goodness of Fit (GoF) statistics or typical statistical tests, these heuristics examine the characteristics of data – addressing the classical issue of Goodness of Logic (GoL) – to provide model recommendations.

- They can be used before fitting the distributions since only the characteristics of data, in terms of the summary statistics, are considered to come up with the model recommendation.
- They can be used as quick characteristics-based guidelines for the safety analysts or practitioners to select a model between the potential alternatives.
- The complexity of the potential alternatives is considered implicitly in such Model Selection perspective.
- They can be used as quick heuristics when the analyst deals with high velocity of big data and prompt Model Selection decisions are needed periodically.

Chapter V documented the application of the methodology described in Chapter IV to design heuristics to select a logical distribution between (1) the NB and PLN distributions, and (2) the NB and NB-L distributions. The NB and PLN distributions are the most popular and commonly used sampling distributions by safety analysts and practitioners (Lord and Mannering, 2010), mostly due to their simplicity, while the NB-L is a promising distribution to model crash data especially when the datasets are characterized by a long tail or many zero observations. The following points summarizes the results and the key findings:

- **NB vs PLN**: A decision tree was constructed to select a logical distribution between the NB and PLN. The results are shown in Figure 6. Although Decision-Tree classifiers are non-robust and may result in different tree splits in different experiments, Figure 6 can be used by safety analysts as useful characteristics-based

guidelines to select a sampling distribution between the NB and PLN. The overall results indicated that the PLN distribution should be used when data are more skewed but have less percentage of zero observations, while the NB distribution is likely a true distribution otherwise. Next, a Random Forest classifier was used to design a better heuristic. Although the results of a Random Forest classifier cannot be shown graphically, the trained forest can be saved and be used as characteristics-based heuristics to decide between the NB and PLN. The Random Forest classification indicated that between the 22 types of summary statistics used in the analysis, kurtosis, skewness and the percentage-of-zeros are among the most critical summary statistics to choose a sampling distribution between the NB and PLN.

- **NB vs. NB-L**: A Decision-Tree classifier was employed to design a simple decision tree to choose a distribution between the NB and NB-L. Figure 11 indicates the results. The skewness of data was the only summary statistics used by the classifier to discriminate these two distributions. Next, a Random-Forest classifier was applied to design a more accurate Model Selection tool (or heuristics) between these two distributions. The Random Forest classification indicated that the skewness, CV, kurtosis, VMR, and the percentage-of-zeros are among the most important summary statistics (or predictors) required to select a logical distribution between the NB and NB-L.

6.2. Future Research Avenues

In this section, a few potential avenues for further research are explained. This section is divided into three parts. The first part describes the detailed steps of designing a simulation study to explore the performance of various models, under different scenarios that characterizes the mean, variance, and percentage of zeros of data. The second part explores an alternative NB-DP model when the DP base distribution follows a Lindley distribution. The third part describes a few avenues to extend the research for the Model Selection heuristics.

6.2.1. Simulation Analysis

Simulated data are often used to evaluate the performance of different modeling approaches under different scenarios. Since the analyst has a better control over the input and output of analysis, simulation studies, often, result in better or much reliable conclusions. In addition, the analyst can explore and analyze a wider range of scenarios. In most simulation studies in highway safety, a few positive independent variables are simulated from a known distribution (such as the lognormal distribution). Next, the crash data are simulated from a given distribution (such as the NB distribution) for a range of scenarios. Then, the simulated data can be altered to obtain the desired data needed for each analysis. After preparing the required data, alternative models are used to model the simulated data. The modeling results, then, are evaluated based on different metrics such as GoF statistics.

In this study, it was shown that the NB-TDP model outperforms the NB model when data has many zeros observations and/or is characterized by a long tail, using two observed datasets. The experiment with two observed data indicated that the NB-L and NB-DP perform similarly when data include many zero observations while NB-DP can be a better alternative when data are characterized by larger variation or include a few large or unusual numbers that could cause a long tail. However, it is not clear under what conditions (e.g. number of zeros, mean, or dispersion), the NB-DP outperforms the NB-L. A simulation study can be designed to investigate the answer to this question. Potential scenarios to investigate are described below:

- Low mean (μ=0.5), moderate mean (μ=5) and high mean (μ=10),
- Low, moderate and high dispersion.
- Different percentage of zero responses.

The following simulation protocol can be used to simulate data and evaluate the performance of different models.

Step 1: Simulating the Original Simulated Dataset

1.1 Fit an NB GLM to a known dataset (say the Indiana dataset). Record the estimated coefficients for variables.

1.2 Set the size of the original simulated dataset to a large number (say 50,000).

1.3 Simulate a few independent variables (use the lognormal and Bernoulli distributions for simulating continuous and binary variables, respectively).

1.4 Adopt the model intercept to reach the desired mean for simulated data.

1.5 Find the mean of the NB distribution at each site (μ_i) given the estimated coefficients in Step 1.1, simulated data in Step 1.3 and adopted intercept in Step 1.4.

1.6 Set the value of inverse dispersion parameter (φ) to desired value. φ=0.5, 2 and 5 respectively denote high, medium, and low dispersion.

1.7 Simulate a dataset with 50,000 data points from the NB GLM using simulated μ_i in Step 1.5 and φ in Step 1.6.

Step 2: Split the original Data into two Datasets one with all zeros and the other with no zero observation

2.1 Put the data points with zero observation in D_1 and data points with observations that are greater than zero in D_2.

Step 3: Sampling: generate data with the desired zero percentage.

3.1 Set the size of the test dataset to N=1,000.

3.2 Set the percentage of zeros to Z (%). (Z=20%, 30%, 40%, 50%, 60%, 70% 80%, 90%).

3.3 Randomly sample N× Z% data points from D_1 dataset and N× (100-Z)% data points from D_2 dataset.

3.4 Merge data sampled in Step 3.3 together, and shuffle the combined dataset.

3.5 Run NB//NB-L and NB-DP for the dataset generated in Step 3.4 and record the DIC and other GoF metrics.

3.6 Evaluate the models based on the value of the different metrics collected in Step 3.5

6.2.2. NB-DP with Lindley Base Distribution

As noted in Chapter II, the model in Equation (20) can be referred to as a modeling framework, since different distributions can be considered instead of $F_0(.|\theta)$. In this dissertation, a lognormal distribution was used instead of $F_0(.|\theta)$ assuming that the frailty terms on average follow a lognormal distribution a priori. Given the superior performance of the NB-L GLM when data have many zero observations, one interesting option to explore is to consider a Lindley distribution instead of $F_0(.|\theta)$. Equation (40) indicates this model.

$$y_i|v_i\mu_i, \phi \sim NB(v_i\mu_i, \phi) \tag{40-a}$$

$$\gamma_k|\tau \sim Beta(1, \tau)\,, \qquad k = 1,2,..,M \tag{40-b}$$

$$\psi_k|\,\theta \sim \text{Lindley}\,(\theta), \quad k = 1,2,\dots.,M \tag{40-c}$$

$$p_k = \gamma_k \prod_{k'<k} (1-\gamma_{k'}), \quad k = 1,2,\dots,M \tag{40-d}$$

$$v_i \sim F\,(.) \tag{40-e}$$

$$F(.) \sim \text{TDP}(\tau, M, \text{Lindley}\,(\theta)) \equiv \sum_{k=1}^{M} p_k \delta_{\psi_k} \tag{40-f}$$

$$\ln(\mu_i) = \beta_0 + \sum_{j=1}^{d} \beta_j\, x_{ij} \tag{40-g}$$

The simulation protocol explained in the previous section can be used to explore advantages of this model compared to the model that the lognormal distribution is assumed as the DP base distribution.

6.2.3. Further Research in Model Selection Heuristics

In this dissertation, we proposed a method to select a logical distribution between potential alternatives to model crash data. There are a few avenues to improve or extend the proposed approach:

- As noted in Chapter I, substantial efforts have been placed to propose various distributions and models to model crash data over the last decade (Lord and Mannering, 2010; Mannering and Bhat, 2014). In this dissertation, we proposed heuristics to select a logical distribution between the NB and PLN as well as the NB and NB-L. In the future, it is worth to extend the proposed methodology to design heuristics for other common distributions documented in Lord and Mannering (2010).
- In this study, the proposed Model Selection approach was focused on univariate distributions, which form the sampling distributions of much complex generative models, such as the NB mixture with the Dirichlet process (NB-DP) or other parametric or semiparametric generalized linear models (GLMs). “How can we incorporate the covariates into the Model Selection problem?” would be a relevant to help in applying the above procedure in GLM scenarios. If any distributional assumptions on the covariates are made, then it is plausible to extend the present work by augmenting the summary statistics of the dependent variable with the independent variables. However, model misspecification and issues like heterogeneity (Mannering et al., 2016; Behnood et al. 2014; Shirazi et al., 2016b)

could be difficult to handle, but would be an interesting avenue to explore. The key to succeed in such settings involves recognizing and including relevant summary statistics, not only about observations but also the covariates, as well as the interactions between them. For instance, the correlation between covariates and the response variable is deemed to be a key factor (Shirazi et al., 2017b).

- In this dissertation, the effect of the sample size on proposed heuristics was ignored assuming that the sample-size is large. However, the size of the dataset can be a critical factor itself to select one distribution over another. As such, Lord and Mannering (2010) suggested using the PLN distribution over the NB when data are characterized by small sample size and sample mean, due to the potential biased estimation for the NB dispersion parameter. Further analysis in context of heuristics is needed to consider the effect of the sample-size (Lord, 2006, Shirazi et al., 2016a, Shirazi et al., 2017a) on proposed heuristics.

REFERENCES

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record, 2061, 55–63.

Aguero-Valverde, J., 2013. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. Accident Analysis & Prevention, 50, 289-297.

Anastasopoulos, P.C., Tarko, A.P., Mannering, F.L., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accident Analysis & Prevention 40 (2), 768-775.

Antoniak, C.E., 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. Annals of Statistics 2 (6), 1152-1174.

Argiento, R., Guglielmi, A., Hsiao, C., Ruggeri, F., Wang, C., 2015. Modelling the association between clusters of snps and disease responses. . In R. Mitra and P. Mueller (Eds.), Nonparametric Bayesian Methods in Biostatistics and Bioinformatics. Springer.

Bardenet, R., Doucet, A., Holmes, C., 2014. Towards scaling up markov chain monte carlo: An adaptive subsampling approach. In Proceedings of the 31st International Conference on Machine Learning, 405-413.

Beaumont, M.A., Zhang, W.Y., Balding, D.J., 2002. Approximate bayesian computation in population genetics. Genetics 162 (4), 2025-2035.

Behnood, A., Roshandeh, A.M., Mannering, F.L., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Analytic Methods in Accident Research, 3, pp.56-91.

Bhat, C.R., 2014, The Composite Marginal Likelihood (CML) Inference Approach with Applications to Discrete and Mixed Dependent Variable Models. Foundations and Trends in Econometrics, 7(1), pp. 1-117

Blei, D.M., Jordan, M.I., 2006. Variational inference for dirichlet process mixtures. Bayesian Analysis 1 (1), 121-143.

Booth, J.G., Casella, G., Friedl, H., Hobert, J.P., 2003. Negative binomial loglinear mixed models. Statistical Modelling 3 (3), 179-191.

Breiman, L, Friedman, J. H., Olshen, R. A.; Stone, C. J., 1984. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Carota, C., Parmigiani, G., 2002. Semiparametric regression for count data. Biometrika 89 (2), 265-281.

Dhavala, S.S., Datta, S., Mallick, B.K., Carroll, R.J., Khare, S., Lawhon, S.D., Adams, L.G., 2010. Bayesian modeling of mpss data: Gene expression analysis of bovine salmonella infection. Journal of the American Statistical Association 105 (491), 956-967.

Escobar, M.D., West, M., 1995. Bayesian density-estimation and inference using mixtures. Journal of the American Statistical Association 90 (430), 577-588.

Escobar, M.D., West, M., 1998. Computing nonparametric hierarchical models. Practical Nonparametric and Semiparametric Bayesian Statistics, 1-22.

Ferguson, T.S., 1973. A bayesian analysis of some nonparametric problems. The Annals of Statistics, 209-230.

Ferguson, T.S., 1974. Prior distributions on spaces of probability measures. . The annals of statistics, 615-629.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-lindley generalized linear model: Characteristics and application using crash data. Accident Analysis & Prevention 45, 258-265.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2014. A caution about using deviance information criterion while modeling traffic crashes. Safety Science 62, 495-498.

Geedipally, S. R., Shirazi, M., Lord, D., 2017. Exploring the Need for Having Region-Specific Calibration Factors Transportation Research Record, in press.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2014. Bayesian Data Analysis, CRC Press.

Ghosh, P., Gill, P., Muthukumarana, S., Swartz, T., 2010. A semiparametric bayesian approach to network modelling using dirichlet process prior distributions. Australian & New Zealand Journal of Statistics 52 (3), 289-302.

Gonzales-Barron, U., Butler, F., 2011. A comparison between the discrete Poisson-gamma and Poisson-lognormal distributions to characterise microbial counts in foods. Food Control, 22(8), pp. 1279-1286.

Griffin, J.E., Walker, S.G., 2011. Posterior simulation of normalized random measure mixtures. Journal of Computational and Graphical Statistics 20 (1), 241-259.

Guo, J.Q., Trivedi, P.K., 2002. Flexible parametric models for long-tailed patent count distributions*. Oxford bulletin of economics and statistics. 64 1, 63-82.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning. 2001. NY Springer.

Hauer, E., Bamfo, J., 1997. Two tools for finding what function links the dependent variable to the explanatory variables. In: Proceedings of the ICTCT 1997 Conference, Lund, Sweden.

Heinzl, F., Tutz, G., 2013. Clustering in linear mixed models with approximate dirichlet process mixtures using em algorithm. Statistical Modelling 13 (1), 41-67.

Hilbe, J.,2011. Negative binomial regression. 2nd Ed. Cambridge : Cambridge University Press, Cambridge.

Hjort, N., Holmes, C., Muller, P., Walker, S., 2010. Bayesian nonparametrics. Cambridge University Press, Cambridge, UK.

Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96 (453), 161-173.

Ishwaran, H., Zarepour, M., 2002. Exact and approximate representations for the sum dirichlet process. Canadian Journal of Statistics 30 (2), 269-283.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning (Vol. 6). New York: springer.

Kalli, M., Griffin, J.E., Walker, S.G., 2011. Slice sampling mixture models. Statistics and Computing 21 (1), 93-105.

Khazraee S.H., 2016. Full Bayesian Poisson-hierarchical models for crash data analysis: investigating the impact of model choice on site-specific predictions. PhD Dissertation. Department of Civil Engineering, Texas A&M University, College Station, Texas.

Lindley, D. V., 1958. Fiducial distributions and Bayes' theorem. Journal of the Royal Statistical Society. Series B (Methodological), 102-107.

Lord, D., 2006. 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.

Lord, D., Geedipally, S. R., 2011. The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. Accident Analysis & Prevention, 43(5), 1738-1742.

Lord, D., Geedipally, S.R., 2018. Safety Prediction with Datasets Characterised with Excess Zero Responses and Long Tails. In Safe Mobility: Challenges, Methodology and Solutions (pp. 297-323). Emerald Publishing Limited.

Lord, D., Geedipally, S.R., Shirazi, M., 2016. Improved Guidelines for Estimating the Highway Safety Manual Calibration Factors. ATLAS-2015-10.

Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis & Prevention 40 (3), pp. 1123-1134.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research-Part A 44 (5), 291-305.

Lord, D., Miranda-Moreno, L.F, 2008. Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter of Poisson-gamma Models for Modeling Motor Vehicle Crashes: A Bayesian Perspective. Safety Science, 46 (5), pp. 751-770.

Maceachern, S.N., Muller, P., 1998. Estimating mixture of dirichlet process models. Journal of Computational and Graphical Statistics 7 (2), 223-238.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research 1, 1-22.

Mannering, F. L., Shankar, V., and Bhat, C. R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research, 11, 1-16.

Medvedovic, M., Sivaganesan, S., 2002. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 18 (9), 1194-1206.

Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus Empirical Bayes. Transportation Research Record 1840, 31–40.

Mitra, R., Muller, P., 2015. Nonparametric bayesian methods in biostatistics and bioinformatics. Springer-Verlag, New York.

Miyazaki, K., Hoshino, T., 2009. A bayesian semiparametric item response model with dirichlet process priors. Psychometrika 74 (3), 375-393.

Mukhopadhyay, S., Gelfand, A.E., 1997. Dirichlet process mixed generalized linear models. Journal of the American Statistical Association 92 (438), 633-639.

Neiswanger, W., Wang, C., Xing, E., 2013. Asymptotically exact, embarrassingly parallel mcmc. stat 1050 (19).

Oh, J., Lyon, C., Washington, S., Persaud, B., Bared, J., 2003. Validation of fhwa crash models for rural intersections - lessons learned. Transportation Research Record 1840, 41-49.

Ohlssen, D.I., Sharples, L.D., Spiegelhalter, D.J., 2007. Flexible random-effects models using bayesian semi-parametric models: Applications to institutional comparisons. Statistics in Medicine 26 (9), 2088-2112.

Papaspiliopoulos, O., Roberts, G.O., 2008. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. Biometrika 95 (1), 169-186.

Peng, Y., D. Lord, and Y. Zou (2014) Applying the Generalized Waring model for investigating sources of variance in motor vehicle crash analysis. Accident Analysis & Prevention 73, 20-26.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2015). Reliable ABC model choice via random forests. Bioinformatics, 32(6), 859-866.

Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis and Prevention 36 (2), 183-191.

Quiroz, M., Villani, M., Kohn, R., 2015. Speeding up mcmc by efficient data subsampling. Riksbank Research Paper Series 121.

Shirazi, M., Lord, D., Geedipally, S.R., 2016a. Sample-size guidelines for recalibrating crash prediction models: recommendations for the Highway Safety Manual. Accident Analysis & Prevention, 93, pp.160-168.

Shirazi, M., Lord, D., Dhavala, S. S., Geedipally, S. R., 2016b. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention, 91, 10-18. Copyright [2016] by Elsevier. DOI: https://doi.org/10.1016/j.aap.2016.02.020

Shirazi, M., Geedipally, S.R., Lord, D., 2017a. A Monte-Carlo simulation analysis for evaluating the severity distribution functions (SDFs) calibration methodology and determining the minimum sample-size requirements. Accident Analysis & Prevention, 98, pp.303-311.

Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017b. A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the Negative Binomial Lindley (NB-L) is preferred over the Negative Binomial (NB). Accident Analysis & Prevention, 107, pp.186-194. Copyright [2017] by Elsevier. DOI: https://doi.org/10.1016/j.aap.2017.07.002

Sethuraman, J., 1994. A constructive definition of dirichlet priors. Statistica Sinica 4 (2), 639-650.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lun, D., 2003. Winbugs version 1.4.1 user manual. MRC Biostatistics Unit, Cambridge.

Teh, Y. W., 2010. Dirichlet Processes. Encyclopedia of Machine Learning. Springer

Vangala, P., Lord, D., Geedipally, S. R., 2015. Exploring the application of the Negative Binomial-Generalized Exponential model for analyzing traffic crash data with excess zeros. Analytic Methods in Accident Research 7, 29-36.

Washington, S., Karlaftis, M., Mannering, F., 2011. Statistical and econometric methods for transportation data analysis,. Chapman and Hall/CRC, Boca Raton, FL.

Yang, M.A., Dunson, D.B., Baird, D., 2010. Semiparametric bayes hierarchical models with mean and variance constraints. Computational Statistics & Data Analysis 54 (9), 2172-2186.

Zamani, H., Ismail, N., 2010. Negative binomial-Lindley distribution and its application. Journal of Mathematics and Statistics, 6(1), 4-9.

Zou, Y., Wu, L., Lord, D., 2015. 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.