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)

Mohammadali Shirazi*

PhD. Candidate, Zachry Department of Civil Engineering

Texas A&M University, College Station, TX 77843, United States

Email: alishirazi@tamu.edu

Soma Sekhar Dhavala, Ph.D.

Chief Data Scientist

Perceptron Learning Solutions Pvt Ltd., , Bengaluru, India

Email: soma.dhavala@gmail.com

Dominique Lord, Ph.D.

Professor, Zachry Department of Civil Engineering

Texas A&M University, College Station, TX 77843, United States

Email: d-lord@tamu.edu

Srinivas Reddy Geedipally, Ph.D.

Associate Research Engineer,

Texas A&M Transportation Institute, Arlington, TX 76013, United States

Email: srinivas-g@tti.tamu.edu

Submission Date (Revised): May 24, 2017

*Corresponding Author

ABSTRACT

Safety analysts usually use post-modeling methods, such as the Goodness-of-Fit statistics or the Likelihood Ratio Test, 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 one 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 or tests do not provide any intuitions into why a specific distribution (or model) is preferred over another (Goodness-of-Logic). This paper ponders into these issues by proposing a methodology 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 easy heuristics to predict the label of the 'most-likely-true' distribution for analyzing data. The proposed methodology was applied to investigate when the recently introduced Negative Binomial Lindley (NB-L) distribution is preferred over the Negative Binomial (NB) distribution. Heuristics were designed to select the 'most-likely-true' distribution between these two distributions, given a set of prescribed summary statistics of data. The proposed heuristics were successfully compared against classical tests for several real or observed datasets. Not only they 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 the NB-L is preferred over the NB - or vice versa - when modeling data.

1. INTRODUCTION

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. 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 research, we address these topics, and contribute to the crash data modeling 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.

The research in this study was motivated first by looking at the characteristics of the Poisson and Negative Binomial (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 Negative Binomial Lindley (NB-L) (Zamani and Ismail, 2010; Lord and Geedipally, 2011; Geedipally et al., 2012) 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 comparison between the NB and NB-L distributions is used as a case study to illustrate the proposed methodology. As discussed in Lord and Mannering (2010), the NB distribution, despite all its limitations, still remains the most common sampling distribution used by safety modelers or practitioners, due to its simplicity. The analyst, however, should be cautious about the NB shortcomings when modeling crash data. As such, the NB distribution does not perform well when data are characterized by excess number of zero responses or have a long (or heavy) tail. The NB-L distribution attempts to overcome such issues by mixing the NB with the Lindley distribution. Although the NB-L, or other advanced distributions, may have a better performance than the NB, they come at a cost of a more complicated modeling and consuming more computational resources. In practice, it can be argued that the NB distribution should generally be good until a certain point, at which we may need to switch to a better but more complex distribution, such as the NB-L. An important question should now be asked: At what point should a more complex distribution such as the NB-L be used instead of the NB? Model Selection heuristics will be proposed to address this question.

The idea of the paper 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? The objectives of this study consequently are: (1) document a methodology to design heuristics to decide between two or more competitive distributions, based on summary statistics of data; (2) apply the methodology to investigate the "switching" points (or heuristics, to be exact) to select the 'most-likely-true' distribution between the NB and NB-L distributions to model crash or other safety related data.

2. METHODOLOGY

At the heart of the proposed methodology lies a paradigm shift in how Model Selection is both viewed and treated. We view Model Selection as a classification problem - that is, given a set of discriminating features of the data, we like to predict the 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. 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 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 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 Fig. 1.

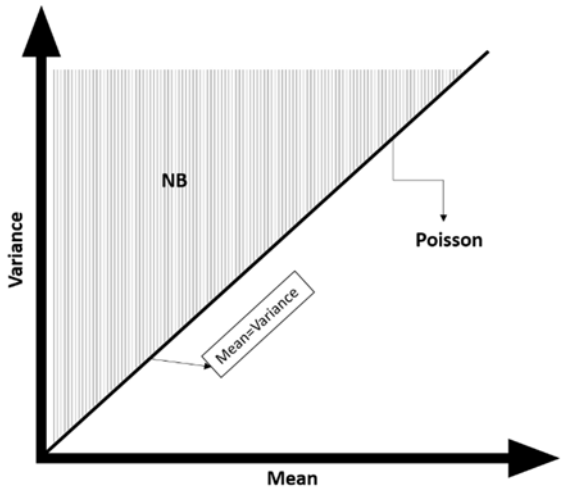

Fig. 1: Classifying the NB and Poisson Distributions Based on the Mean and Variance of the Population.

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[1]. 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-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.

[1] In Section 4, we show that there are strong correlations between the decision based on the VMR heuristic and the LRT statistic.

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 Fig. 1) 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 the 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 is 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 to be 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 shown in Eq.1.

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

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.

3. SIMULATION DESIGN

The first task in our 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? 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 $\phi \sim \text{Uniform}[0.1,10]$ and $\mu \sim \text{Uniform}[0.1,20]$.

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 diseases 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 $\mathbf{S}_c$[2] denote the vector of controlled factors in our experiment. The vector $\mathbf{S}_c$ may include summary statistics, such as the 'mean' or 'variance' of the data. Let $f^{trg}(\mathbf{S}_c)$ denote the 'target' (or desired) density for the collected factors. Likewise, let $f_j^{obs}(\mathbf{S}_c)$ denote the 'observed'

[2] Vectors are shown with bold notations.

multivariate empirical (or kernel) density of the controlled factors simulated from the j-th distribution. Then, the importance weight ($\mathbf{W}_j$) of the simulated datasets can be expressed as:

$$\mathbf{W}_j = \frac{f^{trg}(\mathbf{s}_c)}{f_j^{obs}(\mathbf{s}_c)} \quad (2)$$

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 the 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 get up weighed while for others it may be down weighted.

4. APPLICATION OF THE METHODOLOGY

This section is divided into two subsections. In the first part, the proposed methodology is validated by finding the switching points (i.e.: Model-Selection heuristics) between the Poisson and NB distributions, using the DT classifier, and comparing the results with the theoretical expectations (the VMR heuristic). In the second part, the methodology is employed to find heuristics for Model Selection between the NB and NB-L distributions, using the DT and RF classifiers.

4.1 Poisson vs. NB

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

$$\text{Poisson}(\lambda) \equiv P(Y = y|\,\lambda) = \frac{\lambda^y \times e^{-\lambda}}{y!} \quad (3)$$

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.

The NB distribution is a mixture of the 'Poisson' and 'gamma' distributions. The pmf of the NB distribution is defined as follows:

$$\text{NB}(\phi, p) \equiv P(Y = y|\,\phi, p) = \frac{\Gamma(\phi + y)}{\Gamma(\phi)\Gamma(y+1)}(1-p)^{\phi}(p)^y \quad (4)$$

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.

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

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

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)[3], kurtosis[4] (K), percentage-of-zeros (Z), quantiles (Q) in 10% increments, the 10%, 20%, 30% and 40% inter-quantiles (IQRs), and the range (R). Next, a DT classifier was used to classify the 22-dimensional predictor space that is created by the given summary statistics between the Poisson and NB distributions. Fig. 2 shows the results of the classification. As shown in this figure, the proposed heuristic is empirically found to be close to our theoretical expectations.

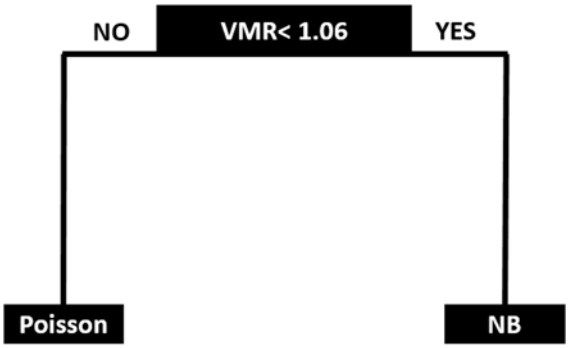

Fig. 2: Heuristic for Model Selection between the Poisson and NB Distributions Using the DT Classifier

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

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

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

(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 1. The sensitivity[5] and specificity[6] 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.

Table 1: Poisson vs. NB: Confusion Matrix Based on the Results of the Decision-Tree Classifier.

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

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-likelihood ratio test between these two distributions. The LRT statistic can be derived using Eq. (5):

$$\text{LRT} = -2 \times \text{LN}\left(\frac{\text{Likelihood under the "Poisson" distribution}}{\text{Likelihood under the "NB" distribution}}\right) \quad (5)$$

[5] Sensitivity=TP/(TP+FN)
[6] Specificity=TN/(TN+FP)

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 Fig. 3. This figure 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.

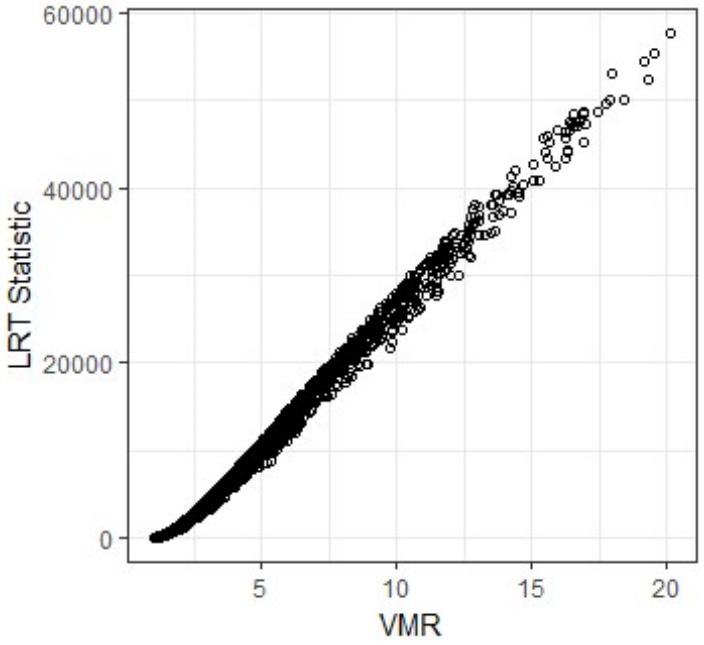

Fig. 3: Poisson vs. NB: Correlation between the Decisions Based on the VMR and the LRT Statistic.

4.2 NB Vs. NB-L

The pdf of the Lindley distribution (Lindley, 1958) is defined as:

$$\text{Lindley}(v|\theta) = \frac{\theta^2}{\theta+1}(1+v)e^{-\theta v} \qquad \theta > 0, v > 0 \tag{6}$$

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

$$y \sim NB(\phi, p = 1 - e^{-\lambda}) \text{ and } \lambda \sim Lindley(\theta) \quad (7)$$

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

$$\lambda \sim \frac{1}{1+\theta} Gamma(2, \theta) + \frac{\theta}{1+\theta} Gamma(1, \theta) \quad (8)$$

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

$$y \sim NB\left(y \middle| \phi, p = 1 - e^{-\lambda}\right) \quad (9\text{-a})$$

$$\lambda \sim Gamma(1 + z, \theta) \quad (9\text{-b})$$

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

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

$$\mu = \phi\left(\frac{\theta^3}{(\theta+1)(\theta-1)^2} - 1\right) \quad (10)$$

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 Section 2 to design Model Selection heuristics to select the 'most likely true' distribution for modeling crash data between these two distributions.

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:

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

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

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:

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

$$\frac{1}{1+\theta} \sim \text{Uniform } (0,0.5)\ ^{7}; \text{ for NB-L}$$

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

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, 22 summary statistics were recorded: 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) in 10% increments, the 10%, 20%, 30%, and 40% inter-quantiles (IQRs) and the range (R).

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 classification. Fig. 4 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[8]. As shown in Fig. 4, 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.

[7] 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.

[8] 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.

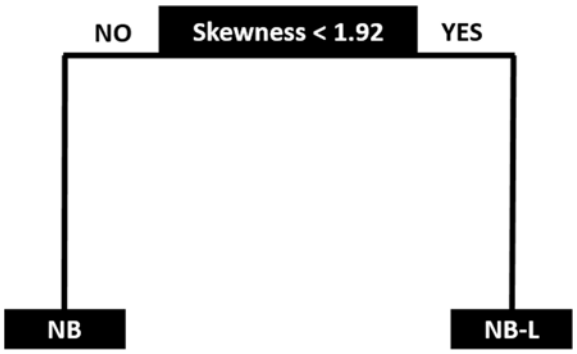

Fig. 4: Heuristic for Model Selection between the NB and NB-L Distributions. (Note: tree can be used for data with 0.1 < mean< 20 and 1 < VMR< 100)

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 2. 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.

Table 2: NB vs. NB-L: Confusion Matrix Based on the Results of the Decision-Tree Classifier.

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

Receiver-Operating-Characteristics (ROC) curves are another tool to evaluate the performance of a classifier. The ROC curves 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

curve based on the results of this classifier is shown in Fig. 5. The value of the AUC is equal to 0.941.

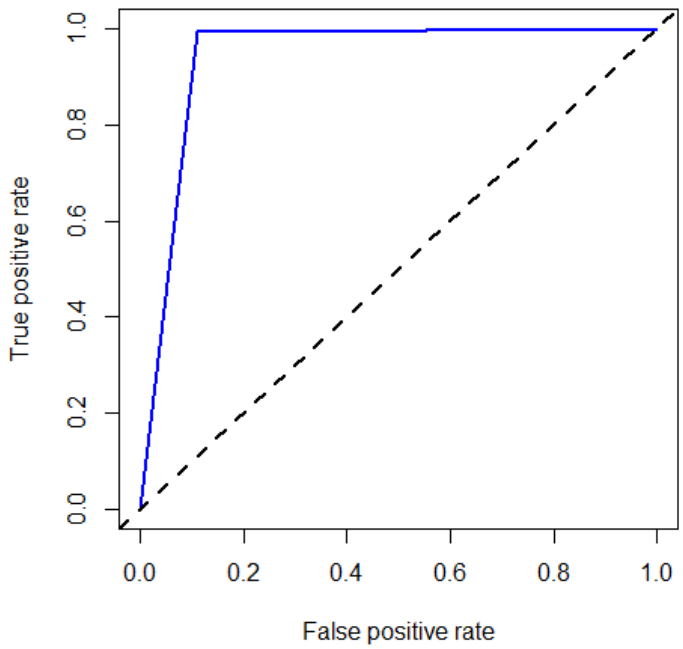

Fig. 5: NB vs. NB-L: ROC Plot Based on the Results of the Decision-Tree Classifier

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. The interested readers are referred to Hastie et al. (2001) or James et al. (2013) for more information about these two measures. Table 3 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. 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 3: NB vs. NB-L: Importance of the Predictors (Summary Statistics) in Partitioning the Predictor Space Based on the Results of the Random Forest Classifier.

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

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 paper. The confusion matrix for the results of the Random-Forest classification is shown in Table 4. 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 Fig. 6. The value of the AUC is equal to 0.999.

Table 4: NB vs. NB-L: Confusion Matrix Based on the Results of the Random-Forest Classifier.

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

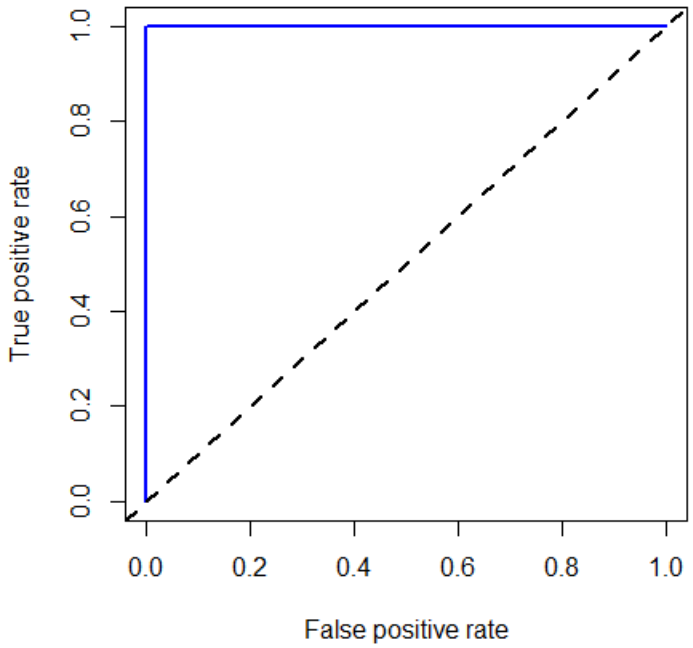

Fig. 6: NB vs. NB-L: ROC Plot Based on the Results of the Random-Forest Classifier.

5. EVALUATING THE NB vs. NB-L HEURISTICS 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 5 shows the summary statistics of these datasets.

Table 5: Summary Statistics of Datasets

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

Tables 6, 7 and 8 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 Tables 6 and 7, both classical tests and proposed heuristics favor the NB-L distribution to model the Texas

datasets. On the other hand, as shown in Table 8, for the Toronto dataset, the NB distribution is the favored distribution among these two options. 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).

Table 6: Model Selection for the Texas Divided Multi-Lane Rural Highway Segments Data Based on the Classical Statistical Tests and Proposed Heuristics.

Method	NB	NB-L	Criteria	Favored Distribution
Chi-Square (χ^2)[1]	2.73	1.68	$\chi^2_{\text{NB-L}} < \chi^2_{\text{NB}}$	NB-L
Log-Likelihood (LL)[1]	-696.1	-695.1	$\text{LL}_{\text{NB-L}} > \text{LL}_{\text{NB}}$	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 7: Model Selection for the Texas Rural Two-Lane Horizontal Curves Data Based on the Statistical Tests and Proposed Heuristics

Method	NB	NB-L	Criteria	Favored Distribution
Chi-Square (χ^2)[1]	57.47	11.68	$\chi^2_{\text{NB-L}} < \chi^2_{\text{NB}}$	NB-L
Log-Likelihood (LL)[1]	-13,557.7	-13,529.8	$\text{LL}_{\text{NB-L}} > \text{LL}_{\text{NB}}$	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 8: Model Selection for the Toronto Four-Legged Signalized Intersections Data Based on the Statistical Tests and Proposed Heuristics

Method	NB	NB-L	Criteria	Favored Distribution
Chi-Square (χ^2)[1]	74.86	615.68	$\chi^2_{\text{NB-L}} > \chi^2_{\text{NB}}$	NB
Log-Likelihood (LL)[1]	-2,988.825	-3,291.933	$\text{LL}_{\text{NB-L}} < \text{LL}_{\text{NB}}$	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.

As a closing note to this section, it should be pointed out that in addition to all theoretical advantages, the proposed heuristics can also be handy as an easy and straightforward Model-

Selection *guidelines* based on characteristics of data for safety practitioners. Such characteristics based guidelines has 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., 2017). These kinds of guidelines are useful in better use of data and modeling resources in practice.

6. DISCUSSION

Our 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 paper, 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.

In this study, we focused on fitting univariate distributions which form the sampling distributions of much complex generative models, such as the NB mixture with the Dirichlet process (NB-DP) (Shirazi et al., 2016b) or other parametric or semi-parametric 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.

7. SUMMARY AND CONCLUSIONS

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, our proposed methodology addresses the classical issue of *Goodness-of-Logic* and looks at the characteristics of data to find the 'most-likely-true' distribution for modeling. 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.

Acknowledgment

The authors would like to thank the Safe-D UTC program for the support obtained for this research. The opinions expressed by the authors in this research do not necessarily reflect those from the Safe-D UTC program.

References

Behnood, A., Roshandeh, A.M. and 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.

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.

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

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

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

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

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

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

Mannering, F. L., and 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.

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.

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.

Shirazi, M., Lord, D. and 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., and 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.

Shirazi, M., Geedipally, S.R. and Lord, D., 2017. 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.

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