Characteristics Based Heuristics to Select a Logical Distribution between the Poisson-Gamma and the Poisson-Lognormal for Crash Data Modelling

Mohammadali Shirazi, Ph.D.*
Research Fellow
Department of Civil and Environmental Engineering
University of Michigan, Ann Arbor, MI, 48109, United States
Email: shirazim@umich.edu

Dominique Lord, Ph.D.
Professor
Zachry Department of Civil Engineering
Texas A&M University, College Station, TX 77843, United States
Tel. (979) 458-3949
Email: d-lord@tamu.edu

*Corresponding Author

ABSTRACT

Several studies have shown that the Poisson-lognormal (PLN) offers a better alternative compared to the Poisson-gamma (PG) when data are skewed while the PG is a more reliable option otherwise. However, it is not explicitly clear when the analyst needs to shift from the PG to the PLN – or vice versa. In addition, so far, the comparison has usually been accomplished using the goodness-of-fit statistics or statistical tests. Such metrics rarely give any intuitions into why a specific distribution or model is preferred over another. This paper addresses these topics by (1) designing characteristics-based heuristics to select a distribution between the PG and PLN, and (2) prioritizing the most important summary statistics to select a distribution between these two options. The results show that the kurtosis and percentage-of-zeros of data are among the most important summary statistics needed to distinguish between these two options.

Keywords: Model Selection, Characteristics Based Heuristics, Classification, Poisson-gamma, Poisson-lognormal

INTRODUCTION

Crash data modelling plays a pivotal role in most safety analyses or evaluations. Over last few decades, safety scientists have placed significant efforts in introducing novel distributions or models to study crash data (Lord and Mannering, 2010; Mannering and Bhat, 2014). However, among all potential modelling alternatives, the Poisson-gamma (PG) (also known as negative binomial) and Poisson-lognormal (PLN) distributions still remain as the most popular and commonly used sampling distributions in the eyes of safety analysts and practitioners (Lord and Mannering, 2010), mostly due to their simplicity. Both of these distributions are classified as a member of the Poisson-mixture family distributions, in which the Poisson distribution is mixed with another distribution (known as a mixing distribution) to overcome the Poisson limitations in accounting over dispersion or heterogeneity in data. In these mixture settings, it is assumed that the Poisson parameter is randomly distributed by a logical mixing distribution. In the case of the PG mixture, the Poisson parameter is distributed using a gamma distribution, while in the case of the PLN distribution, the Poisson parameter is distributed using a lognormal distribution.

Although both 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 PG to handle over dispersion and a better option for modelling skewed data. In a more detailed examination of these two alternatives, Khazraee, Johnson and Lord (2018) state 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. They also found that the existing GoF measures cannot adequately select one model over the other, although the PLN fits the data better at the tail end of the distribution whereas the

PG fits the data better near the zero count data. The comparison of the PG and PLN models is not limited to the safety literature. For example, in a research study 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 PG 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 PG is a more suitable option for the opposite circumstances. However, it is not explicitly clear when the analyst may need to switch from the PG 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 paper 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 PG and PLN distributions, and recognizing the most important summary statistics to make a Model Selection decision between these two sampling distributions.

Recently, Shirazi et al. (2017a) introduced a systematic methodology to design heuristics to select the 'most-likely-true' and logical distribution among potential alternative distributions to model count data. The authors demonstrated the application of the methodology by designing heuristics to select a model between the negative binomial (NB) and negative binomial-Lindley (NB-L) (Lord and Geedipally, 2011; Geedipally, Lord, Dhavala 2012; Shirazi et al. 2016a) distributions. The proposed NB vs. NB-L heuristics recently has been successfully examined in a study by Shaon et al. (2018). The methodology described in Shirazi et al.'s study (2017a) is used as a benchmark to address

our topic. As noted by Shirazi et al. (2017a), when designed, such heuristics have notable advantages to typical Model Selection metrics, such as:

- Unlike the goodness of fit (GoF) metrics or typical statistical tests, these heuristics examine the characteristics of data – addressing the classical issue of goodness of logic (GoL)[1] - for model recommendation.
- They can be used before fitting the distributions since only the characteristics of the 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.

The objectives of this paper consequently are: (1) provide simple guidelines or heuristics to select a logical distribution between the PG and PLN sampling distributions, given a set of summary statistics of data, and (2) determine and prioritize the most important characteristics of data, reflected into the summary statistics, to make a decision between these two distributions. The objectives are accomplished by applying the two-steps, i.e.: (1) Monte Carlo simulations and (2) Classifications, systematic methodology described by Shirazi et al. (2017a).

---

[1] The goodness of logic terminology was first used in the work of Miaou and Lord (2003). The term implies that researchers and analysts should not solely select a model over another based on goodness of fit measures alone, but that they also need to look at the logic behind the selection of the "best model." More specifically, the model should appropriately characterize the crash generation process via the selected distribution, the functional form linking the number of crashes to the explanatory variables and how it relates to the boundary conditions.

MIXED-POISSON FAMILY MODELS

Both of the PG 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 PG and PLN are two common models used to analyze crash data in safety literature (Lord and Mannering 2010; Lord and Miranda-Moreno 2008; Aguero-Valverde and Jovanis 2008; Aguero-Valverde 2013). The characteristics of the PG and PLN distributions are described in this section.

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

$$\text{Poisson}(\lambda) \equiv P(Y = y|\,\lambda) = \frac{\lambda^y \times e^{-\lambda}}{y!} \tag{1}$$

where the mean (m), variance (VAR) and variance-to-mean ratio (VMR) of the observations are equal to:

$$E(y) = m = \lambda \tag{2a}$$

$$V(y) = VAR = \lambda \tag{2b}$$

$$VMR(y) = VMR = 1 \tag{2c}$$

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

$$y|\,\lambda \sim \text{Poisson}(\lambda) \tag{3a}$$

$$\lambda|\mu, \phi \sim \text{gamma}(\phi, \frac{\phi}{\mu}) \tag{3b}$$

The above mixture would result in a closed form NB distribution. The pmf of the NB distribution is defined as follows:

$$\mathrm{NB}(\phi,\mu) \equiv \mathrm{P}(\mathrm{Y}=\mathrm{y}|\,\phi,\mu) = \frac{\Gamma(\phi+\mathrm{y})}{\Gamma(\phi)\Gamma(\mathrm{y}+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{\mathrm{y}} \quad (4)$$

where μ = mean response of observations, and ϕ = inverse dispersion parameter. The mean (m), variance (VAR) and variance-to-mean ratio (VMR) of the PG distribution are defined as:

$$\mathrm{E}(\mathrm{y}) = \mathrm{mean} = \mu \quad (5a)$$

$$\mathrm{V}(\mathrm{y}) = \mathrm{VAR} = \mu + \frac{\mu^2}{\phi} \quad (5b)$$

$$\mathrm{VMR}(\mathrm{y}) = \mathrm{VMR} = 1 + \frac{\mu}{\phi} \quad (5c)$$

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

$$\mathrm{y}|\,\lambda \sim \mathrm{Poisson}(\lambda) \quad (6a)$$

$$\log(\lambda)|\upsilon,\sigma^2 \sim \mathrm{normal}(\upsilon,\sigma^2) \quad (6b)$$

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

$$\mathrm{E}(\lambda) = \mu_\lambda = \mathrm{e}^{\nu+\sigma^2} \quad (7a)$$

$$\mathrm{Var}(\lambda) = \mathrm{V}_\lambda = \frac{\mathrm{e}^{\sigma^2-1}}{\mathrm{e}^{2\upsilon+\sigma^2}} \quad (7b)$$

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

$$\mathrm{E}(\mathrm{y}) = \mathrm{m} = \mu_\lambda \quad (8a)$$

$$\mathrm{V}(\mathrm{y}) = \mathrm{VAR} = \mu_\lambda + \mathrm{V}_\lambda \quad (8b)$$

$$\mathrm{VMR}(\mathrm{y}) = \mathrm{VMR} = 1 + \frac{\mathrm{V}_\lambda}{\mu_\lambda} \quad (8c)$$

MODEL SELECTION HEURISTICS

Shirazi et al. (2017a) documented and discussed a systematic framework to design simple characteristics-based heuristics to predict the label of the most-likely-true distribution to model the data under analysis. In such perspective, the Model Selection problem is treated as a classification problem. The key 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. For more information on rationales behind this Model Selection perspective and detailed steps of the methodology, the readers are raftered to the work of Shirazi et al. (2017a).

This section is divided into three parts. First, the detailed steps of the simulation design are described. In the second part, a Decision Tree (DT) (Breiman, Friedman and Olshen 1984) classifier is used to design simple and straightforward heuristics to select a distribution for modelling between the PG and PLN distributions. The results of this section can be used as straightforward guidelines to select a logical distribution between these two alternatives. In the third part, a Random Forest (RF) (Breiman 2001) classifier is trained to design a more accurate Model Selection tool to predict the 'most-likely-true' distribution between the PG and PLN distributions, as well as prioritizing the key summary statistics to discriminate these two distributions.

Simulation Design

Although the paper by Shirazi et al. (2017a) documented a framework to design the model selection heuristics using the characteristics of the data, this approach requires designing a solid simulation protocol that is tailored to the specific comparison that is being studied. The simulation protocol, itself, can vary substantially from one comparison to another. 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 (Shirazi et al. 2017a). 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 (Shirazi et al. 2017a). 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.

In our problem design, we ensure 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 Eqs. (9a) and (9b)[2].

$$\mathrm{m} \sim \mathrm{uniform}(0.1,20) \tag{9a}$$

$$\mathrm{VMR} \sim \mathrm{uniform}(1,25) \tag{9b}$$

Next, given Eqs. (5a) and (5c), the parameters of the PG distribution can be estimated as:

$$\mu = \mathrm{m} \tag{10a}$$

$$\phi = \frac{\mu}{\mathrm{VMR} - 1} \tag{10b}$$

Similarly, given the Eqs. (8a) and (8c), first, we have:

[2] We assumed that the mean of crash data varies from 0.1 to 20 in our simulation protocol. It is worth pointing out that there are instances that we may have a larger mean for crash data. However, in those situations, our analysis showed that the difference between using the Poisson-gamma and the Poisson- lognormal would become negligible and both will perform similarly when modelling data.

$$\mu_\lambda = m \tag{11a}$$

$$V_\lambda = (VMR - 1) \times \mu_\lambda \tag{11b}$$

Then, given the Eqs. (7a) and (7b), the parameters of the PLN distribution can be derived as:

$$\nu = \log\left(\frac{\mu_\lambda^2}{\sqrt{V_\lambda + \mu_\lambda^2}}\right) \tag{12a}$$

$$\sigma = \sqrt{\log\left(\frac{V_\lambda}{\mu_\lambda^2} + 1\right)} \tag{12b}$$

Now, it is possible to simulate a dataset with a size of n=5,000 from the PG distribution given parameters derived in Eq. 10, and from the PLN distribution given the parameters derived in Eq. 12. The above procedure can be repeated for N=100,000 iterations, for each one of these distributions. Each time, m-types of summary statistics are recorded. We used 22 type of summary statistics in our analysis. These summary statistics include the mean (μ), variance (σ^2), standard deviation (σ), variance-to-mean ratio (VMR), coefficient-of-variation (CV), skewness (skew), kurtosis (K), percentage-of-zeros (Zeros), quantiles (or percentiles) in 10% increments, the 10-th, 20-th, 30-th and 40-th inter-quantiles (or inter-percentiles), and the range (R).

The detailed steps of the simulation protocol are described as follows:

Repeat the following steps for N=100,000 iterations:

1. Simulate the mean (m) and the VMR from the Eqs. (9a) and (9b).
2. Find the parameters of the PG distribution from the Eqs. (10) and the PLN distribution from Eqs. (11) and (12).
3. Simulate a dataset with a size of 'n' given the parameters derived in Step 2, from both of the PG and the PLN distributions.

4. Record all the 22 types of summary statistics described above for the simulated datasets.

Decision Tree Heuristic

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 PG or the PLN) to each partition. Fig. 1 shows the outcome of the Decision Tree classifier. As shown in Fig. 1, the population kurtosis and the percentage-of-zeros play a substantial role in making a decision between the PG 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 PG distribution is a better option otherwise; these results confirm the trends observed and/or reported in previous studies in the literature (Lord and Mannering 2010; Gonzaless-Barron and Butler 2011; Khazraee, Johnson and Lord, 2018). Unlike previous studies, however, Fig. 1 provides a more perspicuous characteristics-based guidance on selecting a sampling distribution between these two alternatives. It is worth pointing out that since different types of summary statistics are independent of parameters, the same criteria is applicable for other NB parameterizations as well.

< Figure 1 >

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 PG distributions, respectively, be labelled as the positive (P) and negative (N) outputs of the binary classification. These definitions represent a test when the analyst assumes the PG distribution as a base model, while he or she seeks to

know when a shift to the PLN distribution is recommended. Table 1 shows the confusion matrix of the binary classification given such assumptions.

< Table 1 >

The overall misclassification error is equal to 9.68% and the sensitivity [Note: Sensitivity=TP/(TP+FN)] and specificity [Note: Specificity=TN/(TN+FP)] 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 very 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 PG distribution, there are still some chances that the output label was detected incorrectly. When the output of the classifier is the PG 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 PG and PLN. In the next section, we use a Random Forest classifier for a more accurate classification. Note that when the sample kurtosis and the percentage-of-zeros deviate further away from the discriminating threshold, the 'most-likely-true' label can potentially be selected with greater confidence. Although not reported here, the DT heuristic was tested for simulated test data and the misclassification error was less than 10% for the test data.

Random Forest Heuristic

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 (Hastie, Tibshirani and Friedman 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 (Hastie, Tibshirani and Friedman 2001; James et al. 2013). As a bagging method, the Random Forest classifier avoids over-fitting using a bootstrap technique. In that regard, this classifier is a more appropriate alternative comparing to boosting classifiers.

In our Random Forest classification, the number of trees was set to 100. Unlike the Decision Tree classification, the outcome of a Random Frost 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 PG and PLN distributions, without any post-modelling efforts. Table 2 shows the confusion matrix of the binary classification between the PG and PLN, based on the results of the Random Forest classifier. The misclassification error is equal to 0.01%, and the sensitivity and specificity of the classifier are almost equal to 100%. 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 2 >

As a by-product of the Random Forest classifier, the predictors (summary statistics) can be ranked by their importance. Fig. 2 and 3 show the importance of the summary statistics based on two criteria: (1) mean decrees Deviance Accuracy and (2) mean decrease Gini index (Hastie, Tibshirani and Friedman 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 PG and PLN distributions.

< Figure 2 >

< Figure 3 >

As a closing note to this section, it is worth pointing out that in scenarios when a particular covariate provides an extra variability to the model, this covariate itself can also be included in building the heuristics. However, according to Khazraee, Johnson and Lord (2018), the selection of distribution itself is also critical for model selection between Poisson-gamma and Poisson-lognormal. The heuristics developed in this paper provide insightful guidelines on the selection of the sampling distribution.

APPLICATION OF PROPOSED HEURISTICS TO 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. This dataset was utilized in several previous studies (Qin, Ivan and Ravishanker. 2004; Geedipally, Lord and Dhavala. 2013; Shirazi et al. 2016a). 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, Geedipally and Shirazi. 2016. Shirazi, Geedipally and Lord. 2017c; Geedipally, Shirazi and Lord. 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 3.

< Table 3 >

Table 4 and 5, 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-modelling 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 favoured 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 PG (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 more logical distribution to model data. For example, a large kurtosis value in both datasets plays a substantial role in choosing the PLN over the PG. It is worth pointing out that although the results of the Model Selection based on the proposed heuristics and the classic tests are the same for the two examples provided in this paper, this may not be generally the case. In addition, factors such as the sample size and unobserved heterogeneity could influence model selection decisions.

< Table 4 >

< Table 5 >

SUMMARY AND CONCLUSIONS

The Poisson-gamma and Poisson-lognormal are the most popular sampling distributions used in safety analyses and evaluations as a means to analyse crash data. According to the previous research (Khazraee et al, 2018), the selection of a distribution between the PG and PLN, itself, is critical for model selection and the subsequent safety studies or

analyses. This study investigated, under what circumstances the PLN is preferred over the PG, and vice versa, based on characteristics of data, reflected in the summary statistics. A decision tree was constructed and proposed as quick guidelines 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 Fig.1 can be used by practitioners as useful guidelines for selecting a sampling distribution between the PG and PLN. We used a Random Forest classifier 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 PG and PLN. The next step should compare the PLN and NB-L, to decide when the percentage-of-zeros favours a model over the other. Further analysis in context of heuristics is also needed to consider the effect of the sample-size (Lord 2006; Shirazi, Lord and Geedipally 2016b; Shirazi, Geedipally and Lord 2017b) on proposed heuristics.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Soma Dhavala for sharing his valuable insights with us. Support for 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 paper 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 report is funded, partially or entirely, 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.]

REFERENCES

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, pp. 289-297.

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

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

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.

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.

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, pp. 258-265.

Geedipally, S. R., Shirazi, M., Lord, D. 2017. Exploring the Need for Region-Specific Calibration Factors. *Transportation Research Record: Journal of the Transportation Research Board*, (2636), pp. 73-79.

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

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

Khazraee, S. H., Johnson, V., Lord, D. 2018. Bayesian Poisson hierarchical models for crash data analysis: Investigating the impact of model choice on site-specific predictions. Accident Analysis & Prevention, 117, pp. 181-195.

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), pp.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), pp. 1738-1742.

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

Lord, D., 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), pp. 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), 2008, pp. 751-770.

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

Miaou, S. P., & Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transportation Research Record, 1840(1), 31-40.

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), pp. 183-191

Shaon, M. R. R., Qin, X., Shirazi, M., Lord, D., Geedipally, S. R. 2018. Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data. *Analytic Methods in Accident Research*, *18*, pp. 33-44.

Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R. 2016a. 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, pp. 10-18.

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

Shirazi, M., Dhavala, S. S., Lord, D., Geedipally, S. R. 2017a. 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.

Shirazi, M., Geedipally, S. R., Lord. D. 2017b. 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, 2017, pp. 303-311.

Shirazi, M., Geedipally, S. R., & Lord, D. 2017c. A procedure to determine when safety performance functions should be recalibrated. Journal of Transportation Safety & Security, 9(4), 457-469.

TABLES

TABLE 1: PG vs. PLN: Confusion Matrix Based on the Results of the Decision Tree Classifier.

Predicted	True	
	PLN	**PG**
PLN	41.50% (TP)	1.18% (FN)
PG	8.50% (FP)	48.82% (TN)

TABLE 2: PG vs. PLN: Confusion Matrix Based on the Results of the Random Forest Classifier.

Predicted	True	
	PLN	**PG**
PLN	50.00% (TP)	0.01% (FN)
PG	0.00% (FP)	49.99% (TN)

TABLE 3: Summary Statistics of the Datasets.

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-th Quantile	0	0
20-th Quantile	0	0
30-th Quantile	0	0
40-th Quantile	0	0
50-th Quantile (Median)	0	0
60-th Quantile	0	1
70-th Quantile	1	1
80-th Quantile	1	3
90-th Quantile	2	6
10-th Inter-Quantile	1	1
20-th Inter-Quantile	1	1
30-th Inter-Quantile	1	3
40-th Inter-Quantile	2	6
Range	61	120

TABLE 4: Model Selection for the Michigan Data.

Method	**PG**	**PLN**	**Criteria**	**Favored Distribution**
Log-Likelihood (LL)[1]	-36332.85 ($\phi = 0.30, \mu = 0.68$)	-36117.54 ($\upsilon = -1.48, \sigma = 1.50$)	$LL_{PLN} > LL_{PG}$	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 5: Model Selection for the Texas Data.

Method	**PG**	**PLN**	**Criteria**	**Favored Distribution**
Log-Likelihood (LL)[1]	-7462.91 ($\phi = 0.23, \mu = 2.36$)	-7432.35 ($\upsilon = -0.82, \sigma = 1.95$)	$LL_{PLN} > LL_{PG}$	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

FIGURES

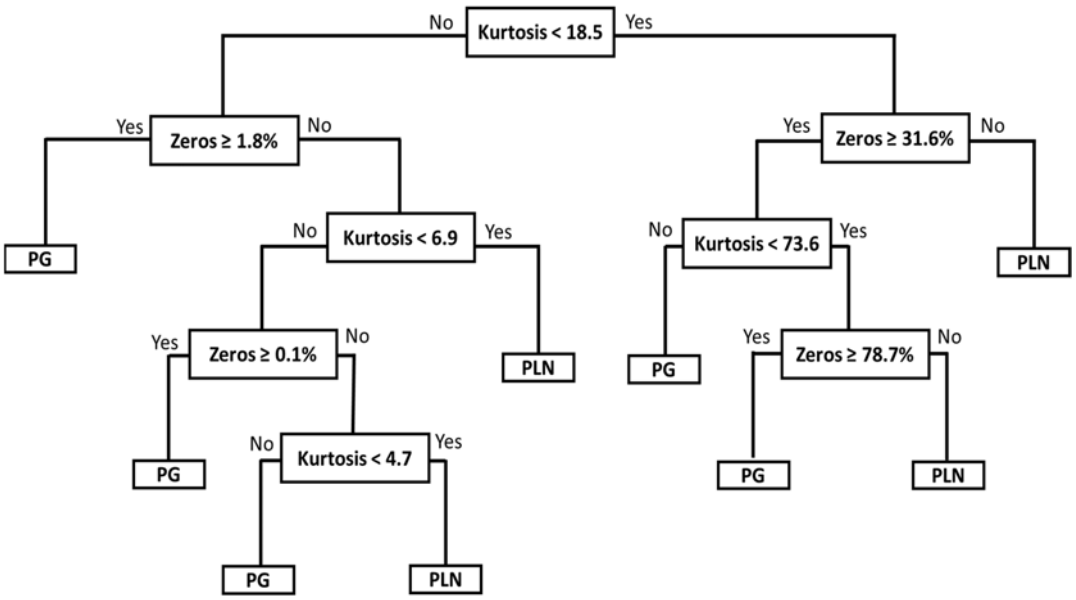

FIGURE 1: Characteristics Based Heuristic to select a Model between the PG and PLN Distributions (Tree can be used for data with the characteristics of 0.1 < mean < 20 and 1 < VMR < 25).

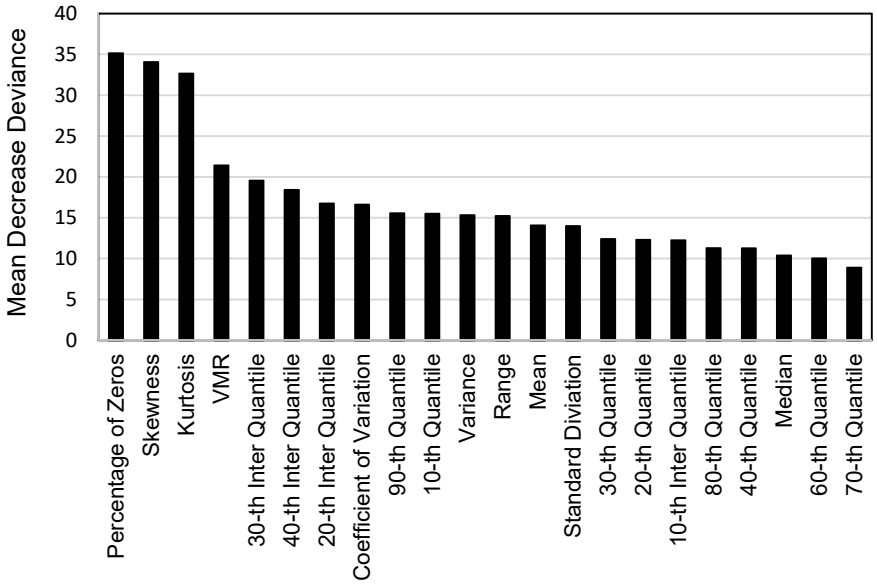

FIGURE 2: Importance of the Summary Statistics to Select a Distribution between the PG and PLN Based on the Mean Decrease Deviance Accuracy Given the Results of the Random Forest Classifier.

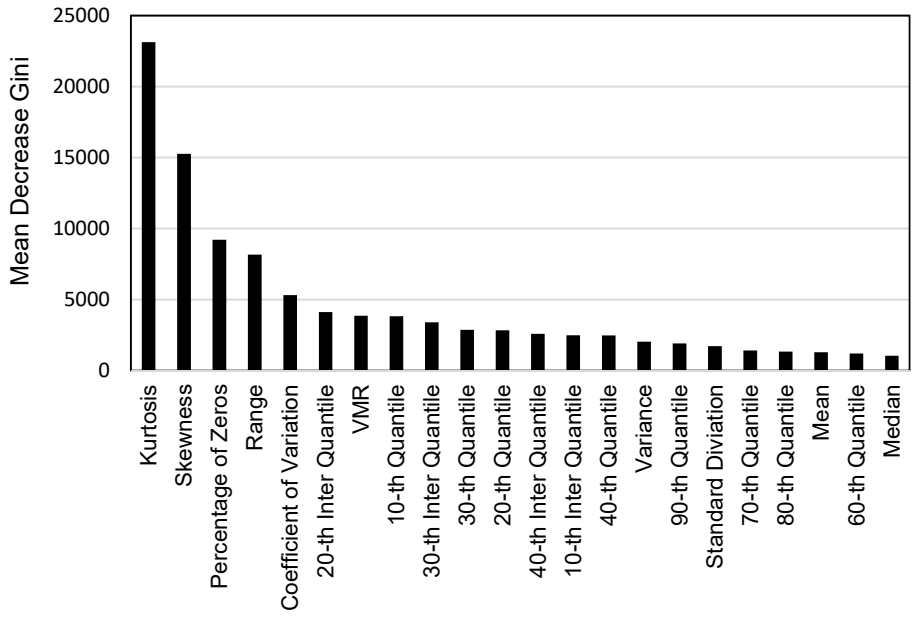

FIGURE 3: Importance of the Summary Statistics to Select a Distribution between the PG and PLN Based on the Mean Decrease Gini, Given the Results of the Random Forest Classifier.