Sains Malaysiana 47(8)(2018): 1931–1940
http://dx.doi.org/10.17576/jsm-2018-4708-35

The Extra Zeros in Traffic Accident Data: A Study on the Mixture of Discrete Distributions

(Lebihan Sifar dalam Data Kemalangan Jalan Raya: Satu Kajian bagi Taburan Diskret Campuran)

ZAMIRA HASANAH ZAMZURI*, MOHD SYAFIQ SAPUAN & KAMARULZAMAN IBRAHIM

ABSTRACT

The presence of extra zeros is commonly observed in traffic accident count data. Past research opt to the zero altered models and explain that the zeros are sourced from under reporting situation. However, there is also an argument against this statement since the zeros could be sourced from Poisson *trial process. Motivated by the argument, we explore the possibility of mixing several discrete distributions that can contribute to the presence of extra zeros. Four simulation studies were conducted based on two accident scenarios and two discrete distributions:* Poisson *and negative binomial; by considering six combinations of proportion values correspond to low, moderate and high mean values in the distribution. The results of the simulation studies concur with the claim as the presence of extra zeros is detected in most cases of mixed* Poisson *and mixed negative binomial data. Data sets that are dominated by* Poisson *(or negative binomial) with low mean show an apparent existence of extra zeros although the sample size is only 30. An illustration using a real data set concur the same findings. Hence, it is essential to consider the mixed discrete distributions as potential distributions when dealing with count data with extra zeros. This study contributes on creating awareness of the possible alternative distributions for count data with extra zeros especially in traffic accident applications.*

Keywords: Hurdle models; negative binomial; Poisson; proportion; simulation study; traffic accident; zero-inflated models

ABSTRAK

Kehadiran lebihan sifar sering dicerap dalam data bilangan kemalangan jalan raya. Kajian lepas cenderung kepada penggunaan model dengan ubah suaian sifar dan menjelaskan bahawa lebihan sifar ini berpunca daripada keadaan kemalangan tidak terlapor. Walau bagaimanapun, terdapat tentangan terhadap pernyataan ini dengan kehadiran lebihan sifar ini boleh berpunca daripada campuran beberapa taburan diskret yang mewakili taburan bagi masa atau lokasi berbeza. Maka, kajian ini bermatlamat untuk meneroka teori bahawa taburan disket tercampur boleh menyumbang kepada lebihan sifar dalam data bilangan. Empat kajian simulasi dijalankan berdasarkan dua senario kemalangan dan dua taburan diskret: Poisson *dan binomial negatif; dengan mengambil kira enam gabungan nilai perkadaran bagi nilai purata rendah, sederhana dan tinggi dalam taburan tersebut. Keputusan kajian bersetuju dengan teori tersebut dengan kehadiran lebihan sifar dapat dikenal pasti dalam kebanyakan kes data* Poisson *tercampur dan binomial negatif tercampur. Set data yang didominasi oleh* Poisson *(atau binomial negatif) dengan nilai purata rendah menunjukkan bilangan lebihan sifar yang ketara walaupun saiz sampel hanyalah 30. Oleh itu, adalah amat penting bagi pengkaji untuk mengambil kira taburan diskret tercampur ini apabila berhadapan data bilangan dengan lebihan sifar. Kajian ini menyumbang dalam mencetus kesedaran berkenaan potensi taburan alternatif untuk data bilangan terlebih sifar terutamanya dalam aplikasi kemalangan jalan raya.*

Kata kunci: Binomial negatif; kajian simulasi; kemalangan jalan raya; model lebihan sifar; model terpangkas; perkadaran; Poisson

INTRODUCTION

Understanding factors that impact on the occurrence of traffic accidents is essential in traffic accident modelling. Finding an appropriate distribution for the accident frequency is crucial, as this will determine the accuracy of the traffic accidents model. The oldest work on the development of traffic accident models can be traced back to Tanner (1953). In the early work of this area, traffic flow is considered as the most influential factor on the occurrence of traffic accidents (Breunning & Boone 1959). Hauer (1988) and Maycock and Hall (1984) produced groundbreaking work in this field, by associating the relationship between the accident rate and explanatory variables, using generalized linear models. The most basic model used is the Poisson regression model (Miao & Lum 1993; Miao et al. 1992). However, traffic accident data are typically over dispersed, hence attention shifted to the negative binomial regression model (Miao 2001, 1994; Vogt 1999; Zegeer et al. 2001). These are univariate models where the contributions of covariates are considered as fixed effects. Chin and Quddus (2003) and Kweon and Kockelman (2003) show that these univariate fixed effects

models are inadequate due to their inability to capture variation caused by unobserved covariates. A review on selected models applied in traffic accident analysis can be found in Zamzuri (2016).

In more recent studies, it is reported that accident counts display an excess presence of zeros than would be expected from either a Poisson or a negative binomial distribution (Chen et al. 2016; Dong et al. 2016; Kumara & Chin 2003; Qin et al. 2004). This raises the issue of finding a more suitable distribution. Zero-inflated distributions gained popularity in the traffic literature as they provide better fits compared to the two count distributions mentioned previously (Kim 2015; Li et al. 2008; Roshandeh 2016).

The fundamental property of zero-inflated distributions is that there are two processes that generate the zero counts in the distribution; namely the structural zeros process and the random zeros process. For example, in counting disease lesions on plants, a plant may have no lesions because of two reasons: It is resistant to the disease and no disease spores have landed on it. This is the distinction between structural zeros, which are unavoidable and random zeros, which occur by chance (Ridout et al. 1998). Martin et al. (2005) categorized different kinds of zeros that occur in ecological data. Another example is the study of species abundance by Welsh et al. (1996). In this study, the abundance data can be thought of as arising from two sources: Animal presence is not tenable since there are no source of foods on site (structural zeros) and zero occurrence by chance on site with foods (random zeros). The main reference for this paper is work by Warton (2005) that found the high frequency of zeros in abundance data is considered to arise from distribution where mean abundance is very low. Identifying the source of the extra zeros is considered necessary as it explains the nature of accident frequency. There are two different opinions about the source of excess zeros in accident data sets.

The under-reporting theory is the most common explanation used in the literature to explain the presence of extra zeros in accident count data (Kumara & Chin 2003; Qin et al. 2004; Shankar et al. 1997). This theory means that at the accident location, there were accidents that were not reported because they were minor, or non-fatality accidents. Since the accident was not reported, it was not recorded and contributes to the zero count in the data set. Accident count data in Malaysia also experiencing such scenario in which up to 1400% of slight injury in motorcycle accidents were not reported (Manan & Varhelyi 2012).

Most authors, for example, Miao (1994), Oh et al. (2006) and Shankar et al. (2003) define the dual states in a zero-inflated distribution as a structural zero accident state and a random zero accident state. The zeros from the structural zero accident state are attributable to unreported minor severity accidents, while zeros from the random zero accident state happen by chance and follow the count distribution. Parameter estimation of this distribution can be performed using maximum likelihood method. Zamzuri (2015) provides an alternative method to estimate the zero inflated distributions.

Lord et al. (2005) argued strongly against the use of zero-inflated models. The structural or true zeros state in zero-inflated distributions means that there is a totally safe situation, where an accident could not happen. They argue it is impossible to achieve such a scenario. They explained that fundamental to crash data is a process called Poisson trials. Poisson trials are Bernoulli trials with unequal probability of events. Each vehicle that enters the intersection has a different probability of experiencing accidents, resulting from a combination of the driver's behaviour, the road condition and other factors; which describe the variability in time and locations. Hence, it is more sensible to consider different discrete distributions for different time periods or locations; and the data set is structured from the combined distributions. Few studies have been conducted on a mixture of discrete distributions in various applications such as insurance claim (Ismail et al. 2004), sudden infant death syndrome (Dalrymple et al. 2003) and fall count data (Ullaha et al. 2010). However, the association between the mixture of discrete distributions with the presence of extra zeros is not extensively explored in the literature.

Motivated by the latter discussion, we propose that the excess zeros in accident data could be sourced from a mixture of several discrete distributions with different mean values. To check on this statement, we conduct simulation studies and record the combinations of several discrete distributions that resulting on the extra zeros in the data set. By fitting two generalized linear model with discrete distributions (Poisson and negative binomial) and four zero augmented models (zero-inflated Poisson, zero-inflated negative binomial, hurdle Poisson and hurdle negative binomial), we record the model that appear as the best fit for the simulated data sets. This illustrates the danger of the normal practice in traffic accident modelling in which the zero augmented models become the best choice when the data has extra zeros. It is crucial to see that although the data do have extra zeros, it is not necessarily distributed as the zero augmented distribution but actually comes from the mixture of discrete distributions. The aim of this paper was to offer an alternative explanation to the extra zeros in traffic accident data, which should be considered as one of the options other than the zero augmented models.

Methods

As discussed in the previous section, the presence of excess zeros in accident data is an unavoidable situation and two possible explanations have been discussed. In this paper, we consider the theory for excess zeros that has not been considered in the literature, in which the accident counts might be collected from two or more different scenarios. For example, the data may come from three traffic periods: on peak, off peak and between peaks. We consider two scenarios that explains the mixture distributions:

Scenario A

Let y_i be the number of accidents on *i*th day. Consider a data set of daily number of accidents in a year, that are categorized into three groups:

On peak: Days in which the traffic flows are seriously heavy, typically encountered during long weekend, festive season or the first and last day of school holidays. This part of the data is presented by distribution with high mean value. Off peak: Days in which the traffic flows are not heavy, presented by distribution with low mean value. Between peaks: days in which the traffic flows are in between (1) and (2). Typically encountered on normal working days. This part of the data is presented by distribution with moderate mean value.

Since the data in each category follows a certain Poisson distribution, hence this data set becomes a mixture of three Poisson distributions with different mean values. In a data set with size n, p_H% of the data is distributed as Poisson with high mean value (μ_H), p_M% of the data is distributed as Poisson with moderate mean value (μ_M) and p_L% of the data distributed as Poisson with low mean (μ_L). The model specification is as follows:

$$y_j \sim P_o(\mu_j)$$

$$\mu_j = \exp\left(X'_j\beta_j\right), j = H, M \,\&\, L$$

Similar scenario is considered for the negative binomial distribution.

Scenario B

Let y_i be the number of accidents at the *i*th intersection in a day, and let y_i is distributed as Poisson with mean μ. Since there are 24 h in a day, we categorize these hours into three groups, similarly as described in Scenario A. The difference here is the Poisson mean of the daily accident count is constructed by three categories of the hours; hence the incorporation is through the mean value as shown below,

$$y_j \sim P_o(\mu_j)$$

$$\mu i = p_H\mu_H + p_M\mu_M + p_L\mu_L$$

$$\mu_j = \exp\left(X'_j\beta_j\right), j = H, M \,\&\, L$$

where p_H is the proportion (weightage) for Poisson with high mean value; p_M is the proportion (weightage) for Poisson with moderate mean value; p_L is the proportion (weightage) for Poisson with low mean value; X_j is the vector of explanatory variables for the *j*th group; β_j is the vector of regression coefficients for the *j*th group.

The same scenario is also considered for the negative binomial distribution.

THE SIMULATION STUDY

A simulation is the process of imitating the real process or subject on the study (Mahdavi & Mahdavi 2014). Based on the two accident scenarios described previously, we perform a simulation study to investigate further the existence of extra zeros in mixed Poisson data. The steps of the simulation study are as follows:

Generate y, the number of accidents based on two scenarios discussed in the previous section. Two explanatory variables considered which are minor (X_1) and major traffic flow (X_2); generated from the uniform distribution:

$$X_1 \sim Unif\,(100, 500)$$

$$X_2 \sim Unif\,(500, 1000)$$

Four different sample sizes are considered: 30, 100, 200 ans 500. Six possible combinations are considered for the proportion values (p_H, p_M, p_L): (0.1, 0.3, 0.6), (0.1, 0.6, 0.3), (0.3, 0.1, 0.6), (0.3,0.6,0.1), (0.6, 0.1, 0.3) and (0.6, 0.3, 0.1).

The regression coefficients $\boldsymbol{\beta}_j$:$\boldsymbol{\beta}_H$ = (–2, 0.5, 0.2), $\boldsymbol{\beta}_M$ = (–2, 0.5, 0.2), $\boldsymbol{\beta}_L$ = (–2, 0.5, 0.2).

The number of data sets generated for the simulation based on mixed Poisson data is 2 scenarios × 4 sample sizes × 6 combinations = 48 sets of data. Fit the regression models based on the Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), Hurdle Poisson (HP) and Hurdle negative binomial (HNB) to each data set generated in (1).

For each model fitted, the Akaike Information Criterion (AIC) and the log likelihood values (LL) are recorded. The best models (whether it is a zero based model or not) is determined by these criterion values. Steps (1) - (3) are repeated for 100 times. Compute the average value for AIC and LL values for each fitted model. Repeat the simulation for the negative binomial based data. Next, we present the details on the models based on the six distributions considered in the simulation study.

THE GENERALIZED LINEAR MODELS

As mentioned previously, the generalized linear models based on Poisson and negative binomial distributions are commonly used in traffic accident modelling. Here we briefly explain the specification of six models fitted to the data set along with model validation criterions considered in the study.

Poisson Model:

Let y_i be the numbers of accidents distributed as Poisson distribution with mean μ_i. The probability density function of y_i, can be expressed as:

$$P(Y_i = y_i) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}, y_i = 0, 1, 2, 3 \ldots$$

To specify the Poisson regression model, μ_i is expressed as a function of the explanatory variables through a log link function, as shown next:

$$\mu_i = \exp(X'B)$$

Negative Binomial Model:
The most common parametric model for over dispersion is negative binomial model, which is generalization of the Poisson regression model that allows differences between the variance and the mean. It assumes that the mean μ_i of Y_i is determined not only by X_i but also by a heterogeneous component of ε_i that is unrelated to X_i. It can be expressed as follows:

$$\widehat{\mu_i} = \exp(X'\beta + \varepsilon_i) = \exp(X'\beta)\exp(\varepsilon_i)$$

where $\exp(\varepsilon_i) \sim$ Gamma $(\alpha^{-1}, \alpha^{-1})$. As a result, the density function of Y_i is:

$$f\left(Y_i \middle| X_i\right) = \frac{\Gamma(Y_i + \alpha^{-1})}{\Gamma(Y_i + 1)\Gamma(\alpha^{-1})} \cdot \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)\alpha^{-1} \cdot \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)Y_i$$

The negative binomial distribution is derived from a gamma mixture of Poisson random variables with a conditional mean and variance of $E(y_i|x_i) = \mu_i = e^{x_i\beta}$ and $V\,ar\,(y_i|x_i) = \mu_i + \alpha\mu_i^2$, respectively. Parameter α caters for the over dispersion part in the data. When $\alpha = 0$, the model reduces to the Poisson regression model. Therefore, the negative binomial model has greater flexibility in modeling the relationship between the expected value and the variance of Y_i.

Hurdle Model:
In addition to overdispersion, many empirical count data sets exhibit more zero observations than would be allowed for by the Poisson model. One model class capable of capturing both properties is the Hurdle model. The Hurdle model combines a count data model $f_{count}(y; x, \beta)$ that is left truncated at y =1 and a zero Hurdle model $f_{count}(y; z, \gamma)$ right censored at y =1:

$$f_{count}(y; x, z, \beta, \gamma) = \begin{cases} f_{count}(0; z, \gamma) & if \quad y = 0 \\ \frac{(1 - f_{zero}(0; z, \gamma))\, f_{count}(y; x, \beta)}{(1 - f_{count}(0; x, \beta))}, & if \quad y > 0 \end{cases} \tag{7}$$

The model parameters β, γ, and potentially one or two additional dispersion parameters θ (if either f_{count} or f_{zero} or both are negative binomial densities) are estimated by maximum likelihood, where the specification of the likelihood has the advantage that the count and the Hurdle component can be maximized separately. The corresponding mean regression relationship is given by:

$$\log(\mu_i) = x\beta + \log(1 - f_{zero}(0; z, \gamma)) - \log(1 - f_{count}(0; x_i, \beta))$$

Zero-Inflated Model:
Zero-inflated models are another option that capable of dealing with excess zero counts. They are two component mixture models combining a point mass at zero with a count distribution such as Poisson and Negative Binomial. Zeros may come from both the point mass and from the count component. Let *f(y)* be the count distribution and be the proportion of structural zeros. The probability function of the zero-inflated distribution is given as:

$$P(y) = \begin{cases} n + (1 - n)\, f(0), & y = 0 \\ (1 - n) f(y), & y = 1, 2\ldots \end{cases}$$

Zero-inflated Poisson (ZIP) distribution is defined as:

$$P(y) = \begin{cases} n + (1 - n)e^{-\lambda}, & y = 0 \\ (1 - n)\frac{e^{-\lambda}\lambda^y}{y!}, & y = 1, 2\ldots \end{cases}$$

Zero-inflated negative binomial (ZINB) distribution is defined as:

$$P(y) = \begin{cases} n + (1 - n)\left(\frac{1}{1 + \theta\mu}\right)^{\frac{1}{\theta}}, & y = 0 \\ (1 - n)\binom{\frac{1}{\theta} + y - 1}{y}\left(\frac{1}{1 + \theta\mu}\right)^{\frac{1}{\theta}}\left(\frac{\theta\mu}{1 + \theta\mu}\right)^y, & y = 1, 2\ldots \end{cases}$$

In terms of finding the best-fitted model to the simulated data sets, we use two common criterions for model diagnostic and validation: The Akaike Information Criteria (AIC) and log likelihood values.

Results and Discussion

In this section, we discuss the main findings from the simulation study conducted. Through the obtained results, we wish to identify the combinations of proportion and the Poisson mean values in which the zero based model appear to be the best fit. When the zero altered model is the best fit to these combinations of mixed Poisson data, it illustrates the potential of alarming situation; in which a zero altered model is wrongly fitted to a data set generated based on a mixed Poisson distribution. Hence, it is essential to consider the possibility of the mixture of discrete distributions when dealing with extra zeros in count data. The first simulation result is the accident count data y_i that follows Poisson distribution; based on the specification as explained in Scenario A, in which the are three Poisson distributions in the data set with certain proportion values given as p_H, p_M and p_L. Table 1 displays the results for the first combination of the proportion values: (0.6, 0.3, 0.1). It is observed that when the sample size is 30, the best-fitted model is Poisson regression model since it has the lowest value of AIC= 103.05 and Log likelihood= -46.52 compared to the other models. When the sample size increases to 100 and 200, the best-fitted model is negative binomial regression. It shows that over dispersion exists in

the dataset when the sample size increases. For the largest sample size considered (n=500,) the best-fitted model is zero-inflated Poisson. Since 60% of the sample in this data set distributed as Poisson with high mean; it is expected that the existence of extra zeros is not apparent in small size data sets. As the sample size gets to 500, then only we can observe that the extra zeros do exist as the best model is shifted to the zero altered model, in this case the ZIP model.

TABLE 1. Results of Simulation A for data based on Poisson distribution with the combination of proportion 60% high, 30% moderate and 10% low

Size	Model	AIC	Loglik
n=30	PM*	103.0459	-46.5229
	NBM	105.0476	-48.5238
	HP	105.9387	-46.96936
	ZIP	105.589	-46.7945
	HNB	107.9389	-46.9694
	ZNB	107.581	-48.7899
n=100	PM	254.2411	-120.6206
	NBM*	253.2457	-120.0029
	HP	255.3513	-121.6756
	ZIP	255.631	-121.8155
	HNB	257.3515	-121.6757
	ZNB	257.6315	-121.8157
n=200	PM	589.846	-291.923
	NBM*	588.8621	-289.9311
	HP	592.261	-290.1305
	ZIP	592.5486	-290.2743
	HNB	594.2634	-290.1317
	ZNB	594.5489	-290.2745
n=500	PM	1440.452	-717.2258
	NBM	1442.48	-717.2402
	HP	1441.701	-714.8506
	ZIP*	1437.854	-712.9269
	HNB	1443.702	-714.8508
	ZNB	1439.855	-712.9273

Next, we look at the exact opposite combinations of the one found in Table 1, in which the combination is (0.1, 0.3, 0.6). This means that 60% of the data comes from Poisson distribution with low mean, that contributes many zeros in the data set. By observing Table 2, we can see a different pattern compared to the one in Table 1. In Table 2, the best fitted models is always the zero augmented model, means that the extra zeros do exist in the data sets even though for the smallest size, n=30. Recall that the true distribution for these data sets is actually a mixed Poisson distributions, this signify the potential danger in which researchers might fit a zero altered model to the data set with extra zeros that actually not distributed as one.

There are four more possible combinations of proportions of the Poisson mean values that are not displayed in this paper due to the space limitation. Figure 1 summarizes this result for simulation based on scenario A for Poisson generated data sets. Red shaded cells signify

TABLE 2. Results for simulation A for data based on Poisson distribution with the combination of proportions (10% high, 30% medium and 60% low)

Size	Model	AIC	Loglik
n=30	PM	65.85108	-29.9255
	NBM	67.8522	-29.9261
	HP	69.4536	-28.72682
	ZIP*	64.8993	-28.4497
	HNB	71.4538	-28.72688
	ZNB	71.1755	-28.5878
n=100	PM	200.2008	-97.1004
	NBM	202.205	-97.1025
	HP	200.7037	-94.3519
	ZIP*	195.5247	-90.7624
	HNB	202.7043	-94.3522
	ZNB	197.5215	-91.7607
n=200	PM	312.5165	-153.2583
	NBM	314.528	-153.264
	HP	306.9555	-147.4777
	ZIP*	303.0771	-145.5386
	HNB	308.9563	-147.4781
	ZNB	315.0746	-150.5373
n=500	PM	846.6711	-420.3355
	NBM	848.6917	-420.3458
	HP	813.4631	-400.7015
	ZIP*	808.2555	-400.0077
	HNB	815.4636	-400.7318
	ZNB	840.2559	-413.1279

that the zero augmented models are best fitted to the data for the corresponding combination of the proportion values. Based on this figure, the proportion combinations that consistently show the presence of extra zeros in the data set event though the sample size is small (n=30) is (0.1, 0.3, 0.6). As predicted, when most of the data (in this case, 60%) come from a Poisson distribution with low mean, the existence of extra zeros is apparent although the sample size is only 30. The second possible combination is (0.3, 0.1, 0.6) in which the zero augmented models appear to be the best fit here starting at n=100. As for other possible combinations, when the sample size gets larger, the presence of extra zeros is detected since the zero augmented models appear to be the best fit model for any combinations when n=500.

Figure 2 summarizes the findings for simulation A for the negative binomial distribution. The results are not strikingly different from Poisson based data. The two possible combinations of the proportions values are the sane with Figure 1, that 60% of the data come from a Poisson distribution with low mean. A slight difference is observed in which all combinations start to tend to the zero based models when the sample size is 200.

The next two figures (Figures 3 & 4) summarize results from Scenario B simulation study for both Poisson and negative binomial distributions. Recall that in Scenario B, the proportion values represent the hours in a day and are induced through the mean value of the count distribution.

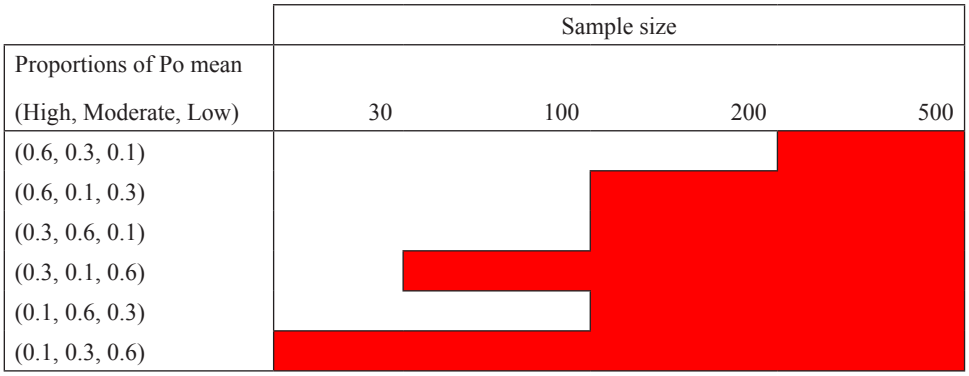

FIGURE 1. Summary of the results for simulation A (data based on Poisson distribution)

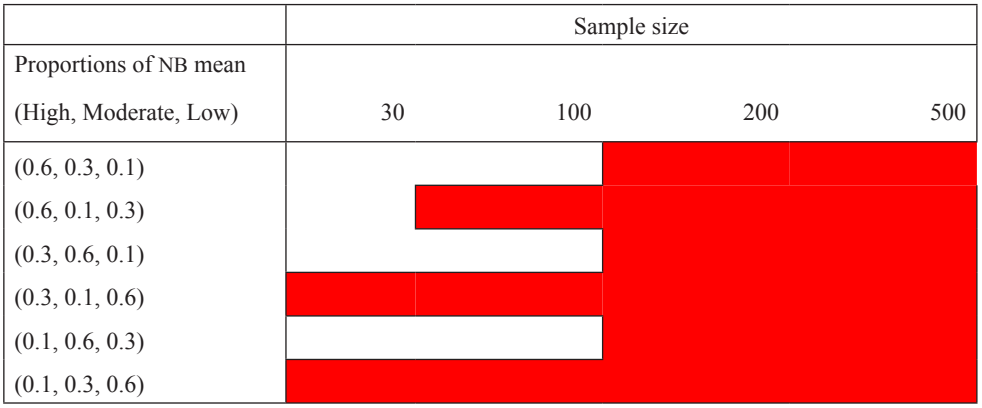

FIGURE 2. Summary of the results for simulation A (data based on the negative binomial distribution)

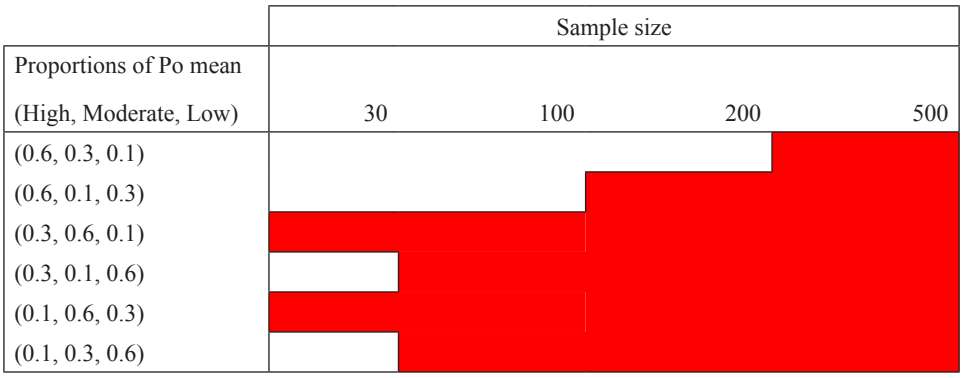

FIGURE 3. Summary of the results for simulation B (data based on Poisson distribution)

By inspecting Figure 3, the combinations that contribute to the presence of extra zeros in the data set are the ones with p_M=0.6. The moderate mean value represents the mean of the accident counts occurred in between peaks hour. This is realistic and consistent with most real-life data, since most hours in a day are in the between peaks periods. Meanwhile for the negative binomial based data, the combinations identified are the ones with p_H=0.1, which means that the negative binomial distribution with high mean value has the lowest weightage here.

ILLUSTRATION

For an illustration purpose, we consider fitting the models in the simulation study to a real data sourced from Bruin (2006). The data set consists of 250 observations on visitors to a fishing park with the following variables recorded: fishing status, camping status, number of child in the group and the number of fish caught. As can be seen in Figure 5, the data has a large zero count, hence zero-augmented models may be suitable options here.

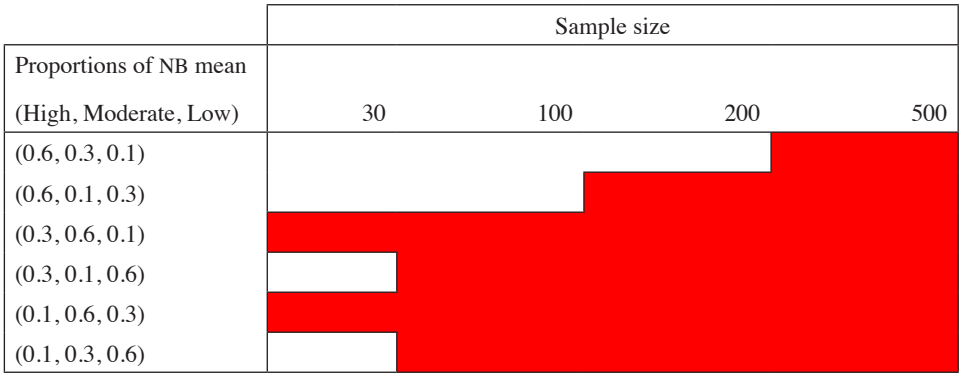

FIGURE 4. Summary of the results for simulation A (data based on the negative binomial distribution)

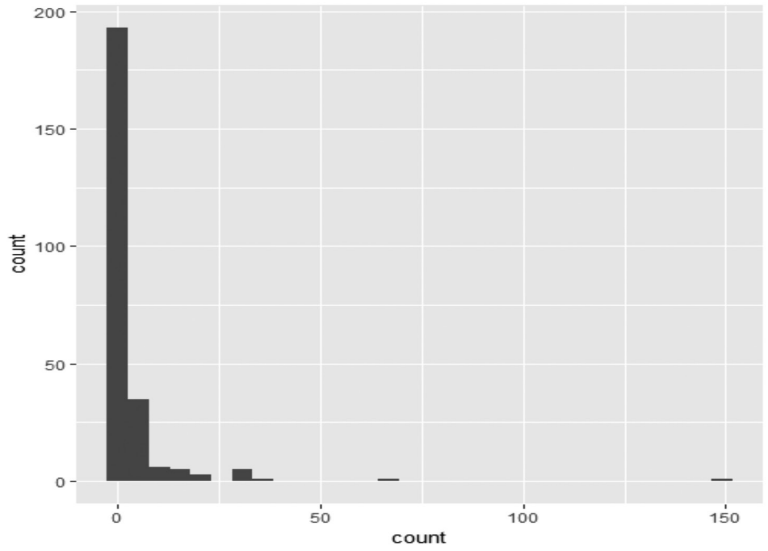

FIGURE 5. Histogram of the fish data

We also want to consider the mixture of discrete distributions as discussed extensively in the simulation study. We consider zero-inflated Poisson and zero inflated negative binomial for zero adjusted models. Note that the hurdle model is not considered here as it is unrealistic to claim that the zeros only come from one source. To consider the mixture distributions, we partitioned the data into four sets based on fishing status and camping status variables, as given in Table 3. We can see that these sets are analogous to the different periods of time (on peak, off peak and between peak) as discussed in traffic accident scenario.

Figure 6 exhibits four histograms correspond to the four sets of data as given in Table 3. We can see that Sets 1 and 2 have higher count of zeros, around 50 - 60; compared to Sets 3 and 4 which the frequency of zero count is around 20 - 30. This shows that there are two distinct groups based on the fishing status, in which the visitors who do not fish contribute to the extra zeros in Sets 1 and 2. If we compare the dispersion between these four data sets, Set 2 is the most overdispersed, with the highest number of fish caught, around 150. To consider the mixture of discrete distribution, which is four different Poisson distributions and four different negative binomial distributions. We fit both Poisson and negative binomial distributions individually to these four data sets.

TABLE 3. The four partitions of the fish data

Set	Fishing status (1=Yes, 0=No)	Camping status (1=Yes, 0=No)	Sample size
1	0	0	72
2	0	1	31
3	1	0	104
4	1	1	43

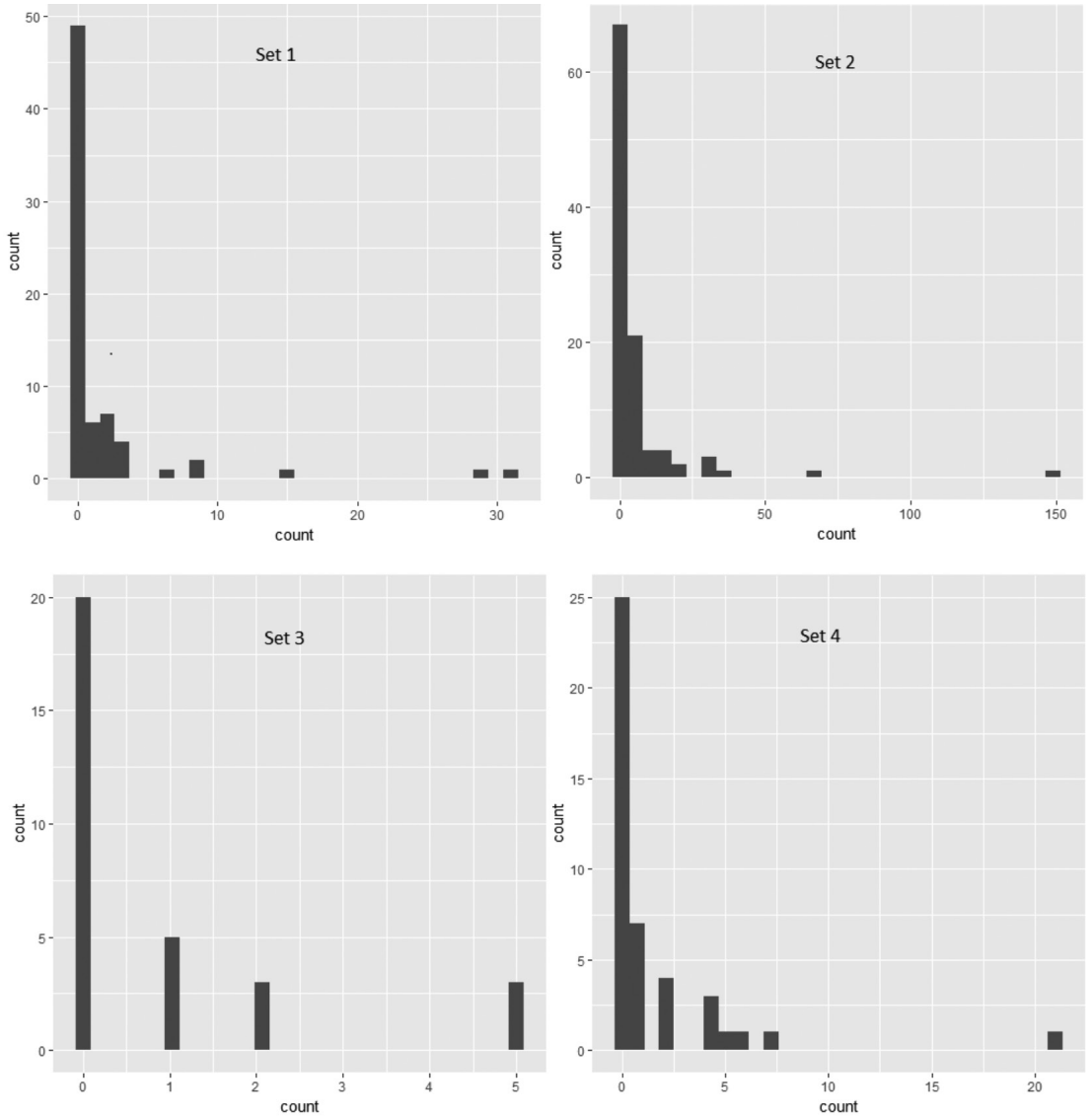

FIGURE 6. Histogram of the four sets of fish data

Table 4 displays the AIC values for the fitted models. Since the data is overdispersed, the mixture of four Poisson model recorded the highest value of AIC. When we fit the ZIP model, the AIC value drops by 708.2, indicates that the ZIP is a better fit to the data compared to the mixture four Poisson. Although ZIP does cater the extra zeros part, the data is still overdispersed; hence when we consider the mixture of four negative binomial and ZINB models, it is apparent that the value of AIC dropped significantly. Furthermore, between the two negative binomial based models, the mixture of four negative binomial distributions fits better to the data set compared to the ZINB model. This concludes the importance on considering the mixture of several discrete distributions to count data with extra zeros.

CONCLUSION

This paper aimed to explain the existence of extra zeros commonly observed in accident count data through the mixed discrete distributions. Most research in traffic

TABLE 4. Comparison on the AIC values for the fitted models

Fitted model	AIC
Mixture of four Poisson distributions	2966.239
Zero-inflated Poisson	2258.046
Mixture of four negative binomial distributions	918.735
Zero-inflated negative binomial	934.879

accident literature is too focusing on the zero-augmented models, in which could not explain the true zero state realistically. Hence, it is very essential to explore the possibility of other potential models that can explain the presence of extra zeros in the count data in a more realistic manner.

To achieve this aim, a simulation study is conducted based on two possible scenarios for Poisson and negative binomial distributions. The scenarios consider three Poisson distributions with different mean values. Each of the distribution's contribution is reflected through the proportion values either as a percentage in the data sample size or weightage in the distribution mean value. By fitting Poisson, negative binomial and four zero augmented models (ZIP, ZINB, HP and HNB), we identified the combinations of proportion and mean values that resulting in existence of extra zeros in the data set. Consistent results are obtained in which the higher contribution from Poisson with low mean values (or negative binomial), the more apparent extra zeros count observed. The findings from this study verify the claim to that the mixture of discrete distributions can have extra zeros in the data sets. A reminder is launched to the researchers especially those in accident modelling analysis, to consider the mixed discrete distributions as possible alternative distribution when dealing with the extra zeros. We include an illustration using a real data set to describe the capability of mixed discrete distribution dealing on count data with extra zeros.

We conclude this paper with several suggestions for directions in future research. Since the mixed discrete distribution has been identified as an alternative distribution for extra zeros count data, the next step is to develop the generalized linear model based on these distributions. Bayesian framework can be considered in the model development since there is more parameters need to be estimated. With the advent in technology, aligned with popularity of data science; a number of unsupervised learning algorithms can also be used for the estimation procedure. The next suggestion is to enhance the results of the simulation study conducted. In this study, the proportion and the regression coefficients are fixed at certain values. A more thorough result can be obtained if we let these parameter values to be random in the next simulation study.

ACKNOWLEDGEMENTS

We would like to thank Universiti Kebangsaan Malaysia and the Ministry of Higher Education (MOHE) for sponsoring this study through grant FRGS/1/2015/ST06/UKM/02/1.

REFERENCES

Breunning, S.M. & Bone, A.J. 1959. *Interchange Accident Exposure Highway Research Board Bulletin 240*, Washington D.C: National Research Council. pp: 44-52.

Bruin, J. 2006. Newtest: Command to compute new test. UCLA: Statistical Consulting Group. https://stats.idre.ucla.edu/stata/ado/analysis/.

Chen, F., Suren, C. & Ma, X. 2016. Crash frequency modeling using real-time environmental and traffic data and unbalanced panel data models. *Int. J. Environ. Res. Public Health* 13(6). doi: 10.3390/ijerph13060609.

Chin, H.C.C. & Quddus, M.A. 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis and Prevention* 35(2): 253-259.

Dalrymple, M.L., Hudson, I.L. & Hudson, R.P. 2003. Finite mixture, zero-inflated *poisson* and hurdle models with application to SIDS. *Computational Statistics & Data Analysis* 41(3): 491-504.

Dong, C., Clarke, D.B., Yan, X., Khattak, A. & Huang, B. 2016. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. *Accident Analysis and Prevention* 70: 320-329.

Hauer, E., Ng, J.C.N. & Lovell, J. 1988. Estimation of safety at signalized intersections. *Transportation Research Record* 1185: 48-61.

Ismail, N., Mohd Ali, K.M. & Chiew, A.C. 2004. A model for insurance claim count with single and finite mixture distribution. *Sains Malaysiana* 33(2): 173-194.

Kim, D.H., Ramjan, M.N. & Mak, K. 2016. Prediction of vehicle crashes by drivers' characteristics and past traffic violations in Korea using a zero-inflated negative binomial model. *Traffic Injury Prevention* 17(1): 86-90.

Kumara, S.S.P. & Chin, H.C. 2003. Modelling accident occurrence at signalized tee intersections with special emphasis on excess zeros. *Traffic Injury Prevention* 4(1): 53-57.

Kweon, Y.J. & Kockelman, K.M. 2003. Overall injury risk to different drivers: Combining exposure, frequency, and severity models. *Accident Analysis & Prevention* 35(4): 441-450.

Li, Z., Knight, S., Cook, L.J., Holubkov, R. & Olson, L.M. 2008. Modeling motor vehicle crashes for street racers using zero-inflated models. *Accident Analysis and Prevention* 40(2): 835-839.

Lord, D., Washington, S.P. & Ivan, J.N. 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention* 37(1): 35-46 .

Mahdavi, M. & Mahdavi, M. 2014. Stochastic lead time demand estimation via monte carlo simulation technique in supply chain planning. *Sains Malaysiana* 43(4): 629-636.

Manan, M. & Varhelyi, A. 2012. Motorcycle fatalities in Malaysia. *IATSS Research* 36: 30-39.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., LowChoy, S.J., Tyre, A.J. & Possingham, H.P. 2005. Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters* 8(11): 1235-1246.

Maycock, G. & Hall, R.D. 1984. Accidents at 4-arm roundabouts. Laboratory Report LR1120, Transport Research Laboratory, Crowthorne, Berks, UK (Unpublished).

Miaou, S.P. 2001. Estimating Roadside Encroachment Rates with the Combined Strengths of Accident and Encroachment-Based Approaches (FHWARD-01-124). Oak Ridge, TN: Oak Ridge National Laboratory (Unpublished).

Miao, S.P. 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention* 26: 471-482.

Miaou, S.P. & Lum, H. 1993. Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis & Prevention* 25(6): 689-709.

Miaou, S.P., Hu, P.S., Wright, T., Rathi, A.K. & Davis, S.C. 1992. Relationship between truck accidents and highway geometric design: A Poisson regression approach. *Transportation Research Record* 1376: 10-18.

Oh, J., Washington, S.P. & Nam, D. 2006. Accident prediction model for railway- highway interfaces. *Accident Analysis & Prevention* 38: 346-356.

Roshandeh, A.M., Agbelie, B. & Lee, Y. 2016. Statistical modelling of total crash frequency at highway intersections. *Journal of Traffic and Transportation Engineering* 3(2): 166-171.

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

Ridout, M., Clarice, G.B. & Hinde, J. 1998. Models for count data with many zeros. *International Biometric Conference*, Cape Town.

Shankar, V., Milton, J. & Mannering, F.L. 1997. Modelling accident frequency as zero-altered probability processes: An empirical enquiry. *Accident Analysis & Prevention* 29: 829-837.

Shankar, V.N., Gudmundur, F.U., Ram, M.P. & MaryLou, B.N. 2003. Modelling crashes involving pedestrians and motorized traffics. *Safety Science* 41: 627-640.

Tanner, J.C. 1953. Accidents at rural three way junctions. *Journal of the Institution of Highway Engineers* 2(11): 56-67.

Ullaha, S., Caroline, F. & Fincha, L.D. 2010. Statistical modelling for falls count data. *Accident Analysis & Prevention* 42(2): 384-392.

Warton, D.I. 2005. Many zeros does not mean zero inflation: Comparing the goodness- of-fit of parametric models to multivariate abundance data. *Environmetrics* 16(3): 275- 289.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F. & Lindenmayer, D.B. 1996. Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling* 88(13): 297-308.

Zamzuri, Z.H. 2016 Selected models for correlated traffic accident count data. Advances in industrial and applied mathematics. *Proceedings of 23rd Malaysian National Symposium of Mathematical Sciences, SKSM 2015*. American Institute of Physics Inc. p. 1750.

Zamzuri, Z. 2015. An alternative method for fitting a zero inflated negative binomial distribution. *Global Journal of Pure and Applied Mathematics* 11(4): 2461-2467.

Zegeer, C.V., Stewart, J.R., Huang, H.H. & Lagerwey, P.A. 2001. Safety effects of marked vs. unmarked crosswalks at uncontrolled locations: Analysis of pedestrian crashes in 30 cities (with discussion and closure). *Transportation Research Record* 1773: 56-68.

Pusat Pengajian Sains Matematik
Fakulti Sains & Teknologi
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: zamira@ukm.edu.my

Received: 29 March 2018
Accepted: 2 April 2018