EJSIT

European Journal of Science, Innovation and Technology

ISSN: 2786-4936

www.ejsit-journal.com

Volume 3 | Number 6 | 2023

Comparative Evaluation of Zero-Inflated and Hurdle Models for Balanced and Unbalanced Data: Performance Assessment and Model Fit Analysis

Intesar N. El-Saeiti [1*], Gadir Alomair [2]

[1] Department of Statistics, University of Benghazi, Libya

[2] Department of Quantitative Methods, School of Business, King Faisal University, Al-Ahsa 31982, Saudi Arabia

ABSTRACT

Excessive zeros in count data pose challenges in statistical modeling, particularly in insurance applications. Zero-inflated (ZI) and hurdle models are commonly employed to address this issue by capturing both zero counts and regular counts. While these models share a similar objective, they differ in their treatment of zeros. Zero-inflated models consider zeros as a component of both zero and regular counts, while hurdle models treat zeros separately from non-zero observations. However, limited research exists on the comparative performance of these models, particularly in the presence of missing data. In this study, we assess the performance of four models: zero-inflated Poisson (ZIP), hurdle Poisson (HurP), zero-inflated negative binomial (ZINB), and hurdle negative binomial (HurNB) models, under balanced and unbalanced data conditions. Using an automobile insurance claims dataset, we employ Akaike's information criteria (AIC) and Bayesian information criteria (BIC) as model selection criteria. Our findings indicate that the ZIP model demonstrates the best fit for the claim frequency dataset, both in balanced and unbalanced data scenarios.

Keywords*:* Zero-inflated Poisson (ZIP), Hurdle Poisson (HurP), Zero-inflated Negative Binomial (ZINB), Hurdle Negative Binomial (HurNB), Balanced data, Unbalanced data

INTRODUCTION

Zero-inflated and hurdle models are commonly employed in statistical analysis to address excess zeros in count data. These models are particularly useful when dealing with datasets that exhibit zero-inflation, where the occurrence of zeros is greater than what would be expected under a standard count distribution. In many fields such as healthcare, economics, and ecology, researchers frequently encounter count data that display varying degrees of zero-inflation and imbalanced distributions. The performance evaluation of zero-inflated and hurdle models is of paramount importance in determining their suitability for analyzing balanced and unbalanced data. Balanced data refers to datasets where the number of observations in each category or group is roughly equal, while unbalanced data refers to datasets with unequal distribution among categories or groups. Assessing the performance of these models under both balanced and unbalanced data scenarios allows for a comprehensive understanding of their effectiveness and applicability in different contexts. Some authors are dealing and studding the Zero-inflated and hurdle models to fit count data with excessive zeros. These models have been compared in various studies, but the results are inconsistent. This paper aims to evaluate the performance of zero-inflated and hurdle Poisson models for overdispersion data through simulation studies and real data analysis (Aswi, Astuti, & Sudarmin, 2022; Feng, 2021; Nekesa, Odhiambo, & Chaba, 2019). The parameters of the logistic component of zero-inflated models, like the parameters of hurdle models, represent impacts on the probability of an observed zero, In contrast to hurdle models though, the parameters of the zero-inflated model's count component describe impacts on any count (i.e., zeros and positive counts) (Lalonde, 2014).

[*] Corresponding Author

The primary objective of this study is to conduct a comparative evaluation of zero-inflated and hurdle models for analyzing balanced and unbalanced data. Specifically, the performance of these models will be assessed in terms of their ability to accurately capture zero-inflation and model the count data distribution. Additionally, the impact of data balance, including cluster size balance, on the performance of these models will be examined. The simulation studies involve different sample sizes, means, and probabilities of zero, while the real data analysis focuses on HIV exposed infants in Kenya (Purnama, 2021). The results show that the zero-inflated Poisson (ZIP) model performs relatively the same or better than the hurdle Poisson model under different scenarios (Zhang, Pitt, & Wu, 2022). Model fit analysis will play a crucial role in the evaluation process. Goodness-of-fit measures, such as Akaike's information criteria (AIC) and Bayesian information criteria (BIC), will be employed to compare the fit of the zero-inflated and hurdle models to the observed data. These criteria will aid in selecting the most appropriate model for the given dataset and provide insights into the adequacy of the models in capturing the underlying count data structure. Additionally, the negative binomial model emerges as the best performing model when fitting data with both structured and non-structured zeros. These findings highlight the importance of considering the specific characteristics of the data and conducting model fit analysis when choosing between zero-inflated and hurdle models.

The findings of this study will contribute to the existing literature on zero-inflated and hurdle models, providing valuable insights into their performance and model fit for balanced and unbalanced data. Researchers and practitioners will benefit from a better understanding of the strengths and limitations of these models, allowing for more informed decisions when analyzing count data with excess zeros and varying data distributions.

STATISTICAL METHODOLOGY

In the context of excess-zero models, there are three key components that are interconnected. These components play a crucial role in the analysis and interpretation of the models. The first component is the random component, which involves specifying the assumed distribution. Typically, this distribution belongs to the exponential family distributions. The second component is the systematic component, which captures the relationship between the parameters and the predictors. It describes how the predictors influence the parameters of the model and ultimately affect the response variable. The third component is the link function, which establishes a connection between the mean of the response variable and the systematic component. In the case of the logistic component, the logit function is used, while for the Poisson or negative binomial components, the log function is employed. The link function helps in transforming the linear predictor to the appropriate scale of the response variable.

Together, these three components work in conjunction to form the excess-zero models, enabling the analysis of data with excess zeros (McCullagh & Nelder, 1989). Considering the random variable Y representing the frequency, the Zero-Inflated models incorporate a logistic regression model to predict "structured zeros" and count regression models to predict counts. These models are characterized by excess-zero distributions, with a probability π for the logistic part and a mean λ for the count part,

$$f_{ZI}(y;\pi,\lambda)=\begin{cases}\pi+(1-\pi)\Pr(\mathrm{K}=0) & y=0\\ \\ (1-\pi)\Pr(\mathrm{K}=\mathrm{y}) & y>0\end{cases} \tag{1}$$

And the distribution of the hurdle models that include a logistic regression model for prediction of a “structured zero” – the single source of zeros-, and zero truncated count regression models for prediction of counts can be written,

$$f_{Hur}(y;\pi,\lambda) = \begin{cases} \pi \; y = 0 \\ (1-\pi)\frac{\Pr(\mathrm{K=y})}{1-\Pr(\mathrm{K=0})} \; y > 0 \end{cases} \quad (2)$$

Where K is a random variable that may follow Poisson or negative binomial distributions. The systematic components and link functions for the excess- zero regression models are,

$$\mathrm{logit}(\pi) = \mathbf{X}_{\boldsymbol{l}}\boldsymbol{\beta}_{\boldsymbol{l}} \quad (3)$$

$$\ln(\lambda) = \mathbf{X}_{\boldsymbol{c}}\boldsymbol{\beta}_{\boldsymbol{c}} \quad (4)$$

Where $\mathbf{X}_{\boldsymbol{l}}$ and $\boldsymbol{\beta}_{\boldsymbol{l}}$ are the design matrix and the parameter vector corresponding to the logistic component, respectively, $\mathbf{X}_{\boldsymbol{c}}$ and $\boldsymbol{\beta}_{\boldsymbol{c}}$ are the design matrix and the parameter vector corresponding to the count component, respectively. Parameters for all the models are estimated using maximum likelihood estimation (MLE) method of estimation. Probability density functions of the distributions are summarized in Table 1 below.

Table 1. Probability density function of excess-zero models

Model	$\mathbf{Pr(K = y)}$
ZIP	$\frac{e^{-\lambda}\lambda^{y}}{y!}$
ZINB	$\binom{y+\kappa-1}{y}p^{\kappa}(1-p)^{y}$
HurP	$\frac{e^{-\lambda}\lambda^{y}}{y!}$
HurNB	$\binom{y+\kappa-1}{y}p^{\kappa}(1-p)^{y}$

The estimation of parameters is through maximization of log-likelihood functions. BFGS method is applied for iterative parameters estimation. Hessian matrix that contains the second derivatives of the log-likelihood function with respect to the parameters of the corresponding excess- zero models is used to find the standard errors of the estimated parameters. To compare the goodness-of-fit of excess-zero models, computing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for model selection (Burnham & Anderson, 2002), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are computed for each model,

$$AIC = -2\,log\,likelihood + 2k \quad (5)$$

$$BIC = -2\,log\,likelihood + k\,ln(n) \quad (6)$$

Where k = number of parameters and n = number of observations. AIC and BIC are model selection criteria based on a Bayesian measure of fit or adequacy and can be used to compare the fit of different models based on the loss of information. Smaller values of AIC and BIC are desirable as they indicate a better fit of the model to the data set.

The Data: The motor insurance dataset used in this study contains information about the frequency of insurance claims. In this study, claim count insurance data is analyzed. The modeling of the excess-zero distributions takes into account the extra zero proportion, which could be attributable to the influence of the deductible agreement and the no-claim discounts (NCD) system. The SAS Enterprise Miner database (SAS Institute Inc., 1998) was used to retrieve the data set on automobile insurance claim frequency. The distribution of claim counts is presented in Table 2. It is observed that the claim frequency variable has a significant number of zero counts. One possible explanation for this zero inflation in the dataset is the presence of a larger proportion of policyholders with low claim risk. The dataset includes various factors related to claims, policies, driving history, and personal information of policyholders. The claim profile file enables the calculation of the number of claims for each policyholder. The policy details, driving records, and personal particulars files provide information about potential risk variables that can impact claim experience. The insurance details file contains information such as the policy number, customer identification number, policy start date, home/working area, commute time, and details about the insured vehicle (value, type, usage, and color). The driving records file includes the policyholder's motor vehicle record points and whether their license has been revoked by government agencies in the previous seven years. These records contribute to assessing the policyholder's driving history. The personal particulars file includes demographic information about the policyholder, such as gender, age, date of birth, marital status, and number of children, annual income, job category, and education level. These details provide insights into the policyholder's personal characteristics. Overall, the dataset contains a comprehensive set of information encompassing claim profiles, policy details, driving records, and personal particulars. This information is crucial for analyzing and modeling the insurance claim frequency data.

Table 2. Number of claims in one year, ranging from 0 to 5

Number of claims in one year, ranging from 0 to 5	Frequency	Percent %
0	1706	60.7
1	351	12.5
2	408	14.5
3	268	9.5
4	74	2.6
5	5	0.2

RESULTS AND DISCUSSION

The study investigates the use of excess-zero regression models for analyzing insurance claim frequency data under balanced and unbalanced data conditions. The predictors used in the analysis were selected based on a previous study that utilized the same dataset. Thirteen variables were chosen to avoid multicollinearity, considering factors such as car usage, marital status, residence location, income, and gender of policyholders, which were found to be important in the Poisson regression model. The presence of zero inflation in the data is tested using a score statistic in the Poisson regression, and the p-value indicates a significant deviation from the Poisson distribution, suggesting the existence of an excess of zeros.

The results of fitting ZIP, ZINB, Hurp, and HurNB regression models to both balanced and unbalanced data are presented in Tables 3, 4, 5, and 6. Across all regression models, the variables related to car usage, annual income, gender, and area of residence show significance in the logistic component, indicating their importance in explaining claim frequency.

The parameter estimates differ between zero-inflated and hurdle models. Business vehicle usage is associated with a higher claim frequency rate in the logistic component of both hurdle models, regardless of the data balance. Policyholders residing in cities are more likely to file claims compared to those in suburbs. Negative estimates for income, gender, and marital status variables suggest that male policyholders, married individuals, and those with higher incomes tend to have a lower claim frequency rate. Under the logistic component of zero-inflated models, married policyholders, males, and those with higher incomes have a higher claim frequency rate. Negative coefficients for car usage and area of residence indicate that business vehicles and policyholders residing in cities are less likely to file claims. In terms of the count portion of both zero-inflated and hurdle models, the only negative estimate is for gender, indicating that male policyholders tend to have a lower claim frequency rate. Comparing the logistic hurdle models with previous studies, they exhibit better compatibility than the zero-inflated models. Based on log-likelihood value, AIC, and BIC, the ZIP model shows slightly better fit compared to the other models in both complete and missing data scenarios.

Table 3. Results of fitting ZIP, ZINB, HurP and HurNB regression models for balanced data

Parameters	ZIP		ZINB		Hurp		HurNB	
	Count	logit	Count	logit	Count	logit	Count	logit
Intercept	0.4399 (0.1123)*	1.3723 (0.1690)*	0.4399 (0.1124)*	1.3723 (0.1690)	0.4557 (0.1129)*	-1.7128 (0.1429)*	0.4556 (0.1129)*	-1.7128 (0.1429)*
Usage	0.0264 (0.0576)	-0.5776 (0.1267)*	0.0264 (0.0576)	-0.5776 (0.1267)*	0.0332 (0.0574)	0.4775 (0.0903)*	0.0331 (0.0574)	0.4775 (0.0903)*
Income	0.0016 (0.0062)	0.0643 (0.0122)*	0.0016 (0.0062)	0.0644 (0.0122)*	0.0002 (0.0062)	-0.0507 (0.0093)*	0.0002 (0.0062)	-0.0507 (0.0093)*
Gender	0.0151 (0.0561)	0.2915 (0.1156)*	0.0151 (0.0561)	0.2915 (0.1156)*	0.0115 (0.0562)	-0.2227 (0.0855)*	0.0116 (0.0562)	-0.2227 (0.0855)*
Married	-0.0153 (0.0551)	0.2070 (0.1149)	-0.0153 (0.0551)	0.2070 (0.1149)	-0.0411 (0.0545)	-0.1521 (0.0835)	-0.0410 (0.0545)	-0.1521 (0.0835)
Area	0.0718 (0.1100)	-2.1383 (0.1576)*	0.0719 (0.1100)	-2.1384 (0.1576)*	0.0796 (0.1095)	1.9244 (0.1317)*	0.0796 (0.1095)	1.9244 (0.1317)*
Dispersion Parameter	-	-	15.1449	-	-	-	13.1524	-
Log-likelihood	-3190		-3190		-3190		-3190	
AIC	6403.247		6405.247		6404.581		6406.582	
BIC	6474.546		6482.488		6475.881		6483.824	

Table 4. Results of fitting ZIP, ZINB, HurP and HurNB regression models for unbalanced data (10%)

Parameters	ZIP		ZINB		Hurp		HurNB	
	Count	logit	Count	logit	Count	logit	Count	logit
Intercept	0.4584 (0.1143)*	1.3468 (0.1757)*	0.4584 (0.1143)*	1.3467 (0.1757)*	0.4925 (0.1200)*	-1.7380 (0.1524)*	0.4925 (0.1200)*	-1.7380 (0.1524)*
Usage	0.0646 (0.0601)	-0.5595 (0.1310)*	0.0646 (0.0601)	-0.5595 (0.1310)*	0.0314 (0.0618)	0.4226 (0.0955)*	0.0314 (0.0618)	0.4226 (0.0955)*
Income	0.0021 (0.0066)	0.0614 (0.0129)*	0.0021 (0.0066)	0.0614 (0.0129)*	0.0010 (0.0067)	-0.0542 (0.0099)*	0.0010 (0.0067)	-0.0542 (0.0099)*
Gender	0.0004 (0.0589)	0.2589 (0.1207)*	0.0004 (0.0589)	0.2589 (0.1207)*	-0.0078 (0.0602)	-0.1938 (0.0903)*	-0.0078 (0.0602)	-0.1938 (0.0903)*
Married	-0.0259 (0.0580)	0.1877 (0.1199)	-0.0259 (0.0580)	0.1877 (0.1199)	-0.0762 (0.0580)	-0.1114 (0.0882)	-0.0762 (0.0580)	-0.1114 (0.0882)
Area	0.0421 (0.1125)	-2.0598 (0.1628)*	0.0421 (0.1125)	-2.0600 (0.1628)*	0.0565 (0.1165)	1.9345 (0.1404)*	0.0565 (0.1165)	1.9345 (0.1404)*

Dispersion Parameter	-	-	15.6636	-	-	-	12.3686	-
Log-likelihood	-2877		-2877		-2854		-2854	
AIC	5834.571		5836.572		5835.305		5837.307	
BIC	5904.603		5912.439		5905.336		5913.174	

Table 5. Results of fitting ZIP, ZINB, HurP and HurNB regression models for unbalanced data (15%)

Parameters	**ZIP**		**ZINB**		**Hurp**		**HurNB**	
	Count	**logit**	**Count**	**logit**	**Count**	**logit**	**Count**	**logit**
Intercept	0.4118 (0.1200)*	1.2796 (0.1820)*	0.4119 (0.1200)*	1.2797 (0.1820)*	0.4285 (0.1205)*	-1.6323 (0.1512)*	0.4283 (0.1206)*	-1.6323 (0.1512)*
Usage	0.0258 (0.0625)	-0.5750 (0.1368)*	0.0258 (0.0625)	-0.5750 (0.1368)*	0.0267 (0.0628)	0.4791 (0.0972)*	0.0268 (0.0628)	0.4791 (0.0972)*
Income	0.0032 (0.0068)	0.0658 (0.0134)*	0.0032 (0.0068)	0.0658 (0.0134)*	0.0020 (0.0069)	-0.0506 (0.0102)*	0.0020 (0.0069)	-0.0506 (0.0102)*
Gender	-0.0116 (0.0613)	0.2946 (0.1255)*	-0.0116 (0.0613)	0.2945 (0.1255)*	-0.0130 (0.0617)	-0.2451 (0.0924)*	-0.0130 (0.0617)	-0.2451 (0.0924)*
Married	-0.0227 (0.0599)	0.1326 (0.1241)	-0.0228 (0.0599)	0.1325 (0.1241)	-0.0443 (0.0596)	-0.1007 (0.0901)	-0.0444 (0.0596)	-0.1007 (0.0901)
Area	0.0888 (0.1166)	-2.0155 (0.1680)*	0.0887 (0.1166)	-2.0157 (0.1680)*	0.0924 (0.1161)	1.8128 (0.1379)*	0.0926 (0.1161)	1.8128 (0.1379)*
Dispersion Parameter	-	-	15.1673	-	-	-	12.5011	-
Log-likelihood	-2708		-2708		-2708		-2708	
AIC	5424.2		5426.2		5426.079		5428.08	
BIC	5493.548		5501.327		5495.427		5503.208	

Table 6. Results of fitting ZIP, ZINB, HurP and HurNB regression models for unbalanced data (20%)

Parameters	**ZIP**		**ZINB**		**Hurp**		**HurNB**	
	Count	**logit**	**Count**	**logit**	**Count**	**logit**	**Count**	**logit**
Intercept	0.4717 (0.1279)*	1.4561 (0.1923)*	0.4717 (0.1279)*	1.4561 (0.1923)*	0.4832 (0.1287)*	-1.7704 (0.1636)*	0.4832 (0.1287)*	-1.7704 (0.1636)*
Usage	0.0062 (0.0631)	-0.6594 (0.1456)*	0.0062 (0.0631)	-0.6594 (0.1456)*	0.0107 (0.0635)	0.5231 (0.1020)*	0.0107 (0.0635)	0.5231 (0.1020)*
Income	0.0041 (0.0068)	0.0766 (0.0137)*	0.0041 (0.0068)	0.0766 (0.0137)*	0.0027 (0.0069)	-0.0587 (0.0104)*	0.0027 (0.0069)	-0.0587 (0.0104)*
Gender	-0.0274 (0.0616)	0.2716 (0.1304)*	-0.0274 (0.0616)	0.2716 (0.1304)*	-0.0199 (0.0621)	-0.2445 (0.0963)*	-0.0200 (0.0621)	-0.2445 (0.0963)*
Married	-0.0209 (0.0605)	0.2426 (0.1308)	-0.0209 (0.0605)	0.2426 (0.1308)	-0.0480 (0.0602)	-0.1795 (0.0941)	-0.0480 (0.0602)	-0.1795 (0.0941)

Area	0.0592 (0.1266)	-2.3463 (0.1827) *	0.0592 (0.1266)	-2.3463 (0.1827) *	0.0682 (0.1260)	2.0910 (0.1520) *	0.0682 (0.1260)	2.0910 (0.1520)*
Dispersion Parameter	-	-	13.0049	-	-	-	12.4056	-
Log-likelihood	-2557		-2557		-2557		-2557	
AIC	5178.125		5180.125		5179.764		5181.765	
BIC	5246.749		5254.468		5248.388		5256.108	

From the results, several conclusions can be drawn. The ZIP regression model exhibits lower AIC and BIC values, indicating a superior fit to the motor insurance data in both complete and unbalanced datasets. In comparison to the ZINB and HurNB models, the Hurp model also demonstrates a satisfactory fit across all scenarios. This suggests that utilizing Poisson as a count distribution for this dataset is more preferable than using NB, as it avoids additional variance in the count component of the models. The parameter estimates for all four models display consistent coefficients that are close to the estimates obtained from the full data. Additionally, the standard errors for the four models are moderately low for both complete and missing datasets, indicating that the models possess strong statistical power and robustness. Across all models and data designs, the standard errors for the logistic components are greater than those for the count components. Furthermore, even with a relatively high percentage of missing data (20%), the results remain consistent. Generally, all models indicate the importance of car usage, residential location, income, and gender of policyholders in the logistic element, while marital status does not exhibit significance. The significance of these variables remains consistent across all models and levels of missing values, demonstrating consistent conclusions in the presence of unbalanced data. In summary, the ZIP model slightly outperforms the other excess zero models in terms of fitting the claim count data for both complete and missing data variations. Despite a higher fraction of missing values, all models produce stable and consistent results with relatively small standard errors. When choosing an appropriate model from the excess zero models for this investigation, the ZIP model is the preferred choice, followed by the Hurp model. The results of logistic hurdle models are more compatible with prior studies than zero-inflated models (Yip & Yau, 2005). Researchers should consider their goals regarding the structure of the zeros and the assumption about inflation of the variance in the count sections when deciding between a ZIP and a HurP model. Zero-inflated and hurdle models are based on different assumptions about the distribution of the outcome variable, resulting in different interpretations. Hurdle models expect zero counts to be produced only as specific zeros. Under the hurdle approach, the count distributions only yield positive results and are therefore truncated at zero.

CONCLUSION

The models indicated the importance of car usage, residential location, income, and gender of policyholders in the logistic element, while marital status does not exhibit significance. The significance of these variables remains consistent across all models and levels of missing values, demonstrating consistent conclusions in the presence of unbalanced data. In summary, the ZIP model slightly outperforms the other excess zero models in terms of fitting the claim count data for both complete and missing data variations. Despite a higher fraction of missing values, all models produce stable and consistent results with relatively small standard errors. When choosing an appropriate model from the excess zero models for this investigation, the ZIP model is the preferred choice, followed by the Hurp model. Researchers should consider their goals regarding the structure of the zeros and the assumption about inflation of the variance in the count sections when deciding between a ZIP and a HurP model.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Aswi, A., Astuti, S. A., & Sudarmin, S. (2022). Evaluating the Performance of Zero-Inflated and Hurdle Poisson Models for Modeling Overdispersion in Count Data. *Inferensi: Jurnal Statistika*, *5*(1), 17-22. doi: 10.12962/j27213862.v5i1.124

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media.

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications*, *8*(1), 8. doi: 10.1186/S40488-021-00121-4

Lalonde, T. L. (2014). *Modeling Correlated Counts with Excess Zeros and Time-Dependent Covariates: A Comparison of ZIP and Hurdle Mixed Models*. Joint Statistical Meetings.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. CRC Press.

Nekesa, F., Odhiambo, C., & Chaba, L. (2019). Comparative assessment of zero-inflated models with application to HIV exposed infants data. *Open Journal of Statistics*, *9*(6), 664-685. doi: 10.4236/OJS.2019.96043

Purnama, D. I. (2021). Comparison of Zero Inflated Poisson (ZIP) Regression, Zero Inflated Negative Binomial Regression (ZINB) and Binomial Negative Hurdle Regression (HNB) to Model Daily Cigarette Consumption Data for Adult Population in Indonesia. *Jurnal Matematika, Statistika dan Komputasi*, *17*(3), 357-369. doi: 10.20956/J.V17I3.12278

SAS Institute Inc. (1998). *Solving Business Problems Using SAS Enterprise Miner Software*. SAS Institute White Paper, SAS Institute Inc., Cary, NC.

Yip, K. C., & Yau, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics, 36*(2), 153-163.

Zhang, P., Pitt, D., & Wu, X. (2022). A comparative analysis of several multivariate zero-inflated and zero-modified models with applications in insurance. *arXiv preprint arXiv:2212.00985*.