Austrian Journal of Statistics
June 2018, Volume 47, 1–19.
http://www.ajs.or.at/
doi:10.17713/ajs.v47i3.590

On Zero-Modified Poisson-Sujatha Distribution to Model Overdispersed Count Data

Wesley Bertoli da Silva
Federal Technology University of Paraná

Angélica Maria T. Ribeiro
Federal Technology University of Paraná

Katiane S. Conceição
University of São Paulo

Marinho G. Andrade
University of São Paulo

Francisco Louzada
University of São Paulo

Abstract

In this paper we propose the zero-modified Poisson-Sujatha distribution as an alternative to model overdispersed count data exhibiting inflation or deflation of zeros. It will be shown that the zero modification can be incorporated by using the zero-truncated Poisson-Sujatha distribution. A simple reparametrization of the probability function will allow us to represent the zero-modified Poisson-Sujatha distribution as a hurdle model. This trick leads to the fact that proposed model can be fitted without any previously information about the zero modification present in a given dataset. The maximum likelihood theory will be used for parameter estimation and asymptotic inference concerns. A simulation study will be conducted in order to evaluate some frequentist properties of the developed methodology. The usefulness of the proposed model will be illustrated using real datasets of the biological sciences field and comparing it with other models available in the literature.

Keywords: zero-modified Poisson-Sujatha, overdispersion, inflation/deflation of zeros, hurdle models, maximum likelihood estimation.

1. Introduction

Most applications involving the analysis of count data are performed using the Poisson and Negative Binomial distributions. The latter is a well-known 2-parameter Poisson compound model that arises as alternative to fit overdispersed data since Poisson models are not applicable in this case. The literature concerning discrete models that accommodate different levels of dispersion is wide and provides several composed distributions as Poisson-Lindley (Sankaran 1970), Negative Binomial-Lindley (Zamani and Ismail 2010), Poisson-Exponential (Cancho, Louzada-Neto, and Barriga 2011), Poisson-Shanker (Shanker 2016a), among others.

A relevant drawback of such compound models is the fact that they do not fit well when a large amount of zeros is observed. To overcome this issue, several zero-inflated and hurdle approaches for standard Poisson model were proposed (Mullahy 1986; Lambert 1992; Zorn 1996). McDowell (2003) provides a insightful discussion about hurdle models. Such approaches were considered by several authors for applications and we will pointed out a few. Bohara and Krieg (1996) show that the modelling of migratory frequency data can be improved using zero-inflated Poisson models. Gurmu and Trivedi (1996) seek to deal with the excess of zeros on data from recreational trips. Ridout, Demétrio, and Hinde (1998) use several zero-inflated Poisson regression models for Apple shoot propagation data. In the social sciences, Bahn and Massenburg (2008) consider the hurdle Poisson model for the number of homicides in Chicago (Illinois - USA). Further applications were considered for quantitative studies on HIV-risk reduction (Heilbron and Gibson 1990; Hu, Pavlicova, and Nunes 2011) and for DNA sequencing data (Beuf, Schrijver, Thas, Criekinge, Irizarry, and Clement 2012). As zero-deflated data seldom arises in practice, there are very few literature addressing this case (Angers and Biswas 2003; Conceição, Andrade, and Louzada 2013) even if this situation is often referred in papers dealing with zero-inflated models.

Recently, the Poisson-Sujatha distribution was obtained by compounding the Poisson with a Sujatha distribution. The latter was introduced by Shanker (2016b) for modelling real lifetime in biological and engineering contexts. The author has shown that this model is a three component mixture of an Exponential distribution with scale parameter θ, a Gamma distribution having shape parameter 2 and scale parameter θ and a Gamma distribution having shape parameter 3 and scale parameter θ with mixing proportions given, respectively, by $\theta^2\upsilon^{-1}$, $\theta\upsilon^{-1}$ and $2\upsilon^{-1}$, being $\upsilon = \theta^2 + \theta + 2$. A comprehensive discussion about the statistical properties of the Sujatha distribution such as moments, hazard function, stochastic orderings, parameter estimation, among others is also presented on the mentioned paper.

The Poisson-Sujatha distribution was introduced and extensively studied by Shanker (2016c) which have discussed its various mathematical properties. Shanker and Fesshaye (2016a) consider the Poisson-Sujatha distribution to model overdispersed counts provided by ecological and genetic experiments. Shanker and Fesshaye (2016b) obtained the size-biased version of the Poisson-Sujatha distribution, presenting its properties and discussing its applications. The zero-truncated Poisson-Sujatha distribution, which will be of particular interest in this paper, was presented by Shanker and Fesshaye (2016c). Further, a detailed report on zero-truncated Poisson, Poisson-Lindley and Poisson-Sujatha distributions is provided by Shanker and Fesshaye (2016d).

Zero-modified models may arise when no information about the kind of zero modification in a given dataset is available. Dietz and Böhning (2000) proposed the zero-modified Poisson regression model for zero inflated/deflated samples and Conceição *et al.* (2013) consider a Bayesian approach for this model as an alternative to model Brazilian *leptospirosis* notification data. Once zero inflated/deflated models may also be useful to deal with data presenting overdispersion, this paper aims to introduce and present the usefulness of the zero modified version of the Poisson-Sujatha distribution, which is itself overdispersed. The proposed model is naturally more flexible than the original one since it takes into account inflation or deflation of zeros, being the first an issue often encountered when analysing count data. For our purpose, we consider a reparameterization of the zero-modified Poisson-Sujatha probability mass function, which will allow the likelihood function to be separable on the model parameters. The estimation procedure will be conducted under the frequentist point of view by the usual likelihood theory. A simulation study will be conducted in order to evaluate some frequentist properties of the maximum likelihood estimators. The usefulness of the proposed model will be illustrated by considering applications to real datasets from the biological sciences field. Standard model comparison will be also provided.

This paper is organized as follows. In Section 2, we briefly present the Poisson-Sujatha distribution, some of its mathematical properties and its zero-truncated version. In Section 3, we introduce the zero-modified Poisson-Sujatha distribution, demonstrating its flexibility to deal with zero inflated/deflated data. In Section 4, the zero-modified Poisson-Sujatha distribution is presented as a hurdle model. In Section 5, maximum likelihood estimation for the unknown parameters as well the asymptotic standard errors and confidence intervals are discussed. In Section 6, a simulation study is presented. In Section 7, the proposed model is considered for application to real datasets. Concluding remarks are addressed in Section 8.

2. Poisson-Sujatha distribution

A random variable ψ is said to have Sujatha (S) distribution if its probability density function (pdf) can be written as

$$g(\psi;\theta) = \frac{\theta^3}{\theta^2+\theta+2}\left(\psi^2+\psi+1\right)e^{-\theta\psi}, \qquad \psi > 0,$$

for $\theta > 0$.

The Poisson-Sujatha (PS) distribution is a probabilistic model that arises when the S distribution is chosen to describe the rate parameter (ψ) of the Poisson (P) distribution. In this case, a random variable X is said to have PS distribution if it follows the stochastic representation

$$\mathrm{X}|\psi \sim \mathrm{P}(\psi) \quad \text{and} \quad \psi \sim \mathrm{S}(\theta).$$

The unconditional distribution of the random variable X can be denoted by $\mathrm{PS}(\theta)$. Let $\mathcal{X}_z = \{z, z+1, \ldots\}$ the set of the integers greater or equal to z. We completed the definition by stating that a random variable X, defined on $\mathcal{X}_0$, will have PS distribution if its probability mass function (pmf) can be written as

$$f(x;\theta) = \frac{\theta^3}{\theta^2+\theta+2}\left[\frac{x^2+x(\theta+4)+\left(\theta^2+3\theta+4\right)}{(\theta+1)^{x+3}}\right], \qquad x \in \mathcal{X}_0, \tag{1}$$

for $\theta > 0$. Using the gamma integral, the above result can be easily proved by integrating $f(x|\psi)\,g(\psi;\theta)$ respect to ψ over $\mathbb{R}_+$, being $f(x|\psi)$ the conditional pmf of a P variable.

From the results provided by Shanker (2016c) we have that the r^{th} factorial moment about the origin of the PS distribution is given by

$$\mu_r' = \frac{r!\left[\theta^2+(r+1)\theta+(r+1)(r+2)\right]}{\theta^r\left(\theta^2+\theta+2\right)}, \tag{2}$$

which provides the moments about origin. Thus, the expected value and the variance are

$$\mu = \mu_1' = \frac{\theta^2+2\theta+6}{\theta\left(\theta^2+\theta+2\right)}, \tag{3}$$

and

$$\sigma^2 = \mu_2' - \left(\mu_1'\right)^2 = \frac{\theta^5+4\theta^4+14\theta^3+28\theta^2+24\theta+12}{\theta^2(\theta^2+\theta+2)^2}. \tag{4}$$

It is easily to see that the variance term can be written as

$$\sigma^2 = \mu\left[1+\frac{\theta^4+4\theta^3+18\theta^2+12\theta+12}{\theta\left(\theta^2+\theta+2\right)\left(\theta^2+2\theta+6\right)}\right] = \mu\tau, \tag{5}$$

being the ratio involving the parameter θ always positive. This implies that the PS distribution is overdispersed, i.e. whichever $\theta > 0$ we have that $\sigma^2 > \mu$. Further, the useful index of dispersion (τ) is clearly greater than 1, also implying overdispersion since $\tau = \sigma^2\mu^{-1}$. On the other hand, we have that $\tau \to 1$ $\left(\sigma^2 \to \mu\right)$ as $\theta \to \infty$, i.e. the PS distribution has the property of equidispersion for large values of θ.

Again, using the relationship between the moments about mean and the moments about origin, some useful measures as the coefficient of variation (β), the coefficient of skewness (γ) and the coefficient of kurtosis (ζ) can be derived from equation (2). The expressions of such measures are

$$\beta = \frac{\sigma}{\mu} = \frac{\sqrt{\theta^5+4\theta^4+14\theta^3+28\theta^2+24\theta+12}}{\theta^2+2\theta+6},$$

$$\gamma = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\theta^8 + 7\theta^7 + 32\theta^6 + 110\theta^5 + 228\theta^4 + 300\theta^3 + 240\theta^2 + 144\theta + 48}{\left(\theta^5 + 4\theta^4 + 14\theta^3 + 28\theta^2 + 24\theta + 12\right)^{3/2}},$$

and

$$\begin{aligned}\zeta = \frac{\mu_4}{\mu_2^2} &= \frac{\theta^{11} + 15\theta^{10} + 99\theta^9 + 488\theta^8 + 1682\theta^7 + 4016\theta^6}{\left(\theta^5 + 4\theta^4 + 14\theta^3 + 28\theta^2 + 24\theta + 12\right)^2} + \\ &\quad \frac{7008\theta^5 + 9016\theta^4 + 8784\theta^3 + 6240\theta^2 + 2880\theta + 720}{\left(\theta^5 + 4\theta^4 + 14\theta^3 + 28\theta^2 + 24\theta + 12\right)^2}.\end{aligned}$$

The definition of the zero-truncated Poisson-Sujatha (ZTPS) distribution will be quite useful for the purpose of this paper. A random variable X is said to have ZTPS if its pmf can be written as

$$f_{\text{ZTPS}}\left(x;\theta\right) = \frac{\theta^3}{\theta^4 + 4\theta^3 + 10\theta^2 + 7\theta + 2}\left[\frac{x^2 + x\left(\theta + 4\right) + \left(\theta^2 + 3\theta + 4\right)}{\left(\theta + 1\right)^x}\right], \; x \in \mathcal{X}_1, \tag{6}$$

for $\theta > 0$. See Shanker and Fesshaye (2016c) for further details about the ZTPS distribution.

3. Zero-modified Poisson-Sujatha distribution

Let X be a random variable defined on $\mathcal{X}_0$. Thus, X is said to have zero-modified Poisson-Sujatha (ZMPS) distribution if its pmf can be written as

$$f_{\text{ZMPS}}\left(x;\theta,\pi\right) = \left(1 - \pi\right)\delta_x + \pi\, f\left(x;\theta\right), \qquad x \in \mathcal{X}_0, \tag{7}$$

for $\theta > 0$ and the parameter π is subject to the condition (called π-condition) given by

$$0 \leqslant \pi \leqslant \frac{1}{1 - f\left(0;\theta\right)}, \tag{8}$$

being $f\left(x;\theta\right)$ the pmf of a PS random variable. Further, δ_x is the indicator function, so that $\delta_x = 1$ if $x = 0$ and $\delta_x = 0$ otherwise. Note that (7) is not a mixture distribution typically fitted to zero-inflated data, since parameter π can assume values greater than 1. However, for all values of π between 0 and its upper boundary, the equation (7) corresponds to a properly pmf since $f_{\text{ZMPS}}\left(x;\theta,\pi\right)$ is positive for each x and sums to 1 on $\mathcal{X}_0$.

The expected value and the variance of X are

$$\mu_{\text{ZMPS}} = \pi\mu \quad \text{and} \quad \sigma^2_{\text{ZMPS}} = \pi\left[\sigma^2 + \left(1 - \pi\right)\mu^2\right], \tag{9}$$

where μ and σ^2 are given in equations (3) and (4). Under the ZMPS distribution, the index of dispersion and the coefficients of variation, skewness and kurtosis are given, respectively, by

$$\tau_{\text{ZMPS}} = \frac{\left(\theta^5 + 5\theta^4 + 18\theta^3 + 44\theta^2 + 48\theta + 48\right) - \pi\left(\theta^4 + 4\theta^3 + 16\theta^2 + 24\theta + 36\right)}{\theta\left(\theta^2 + \theta + 2\right)\left(\theta^2 + 2\theta + 6\right)},$$

$$\beta_{\text{ZMPS}} = \frac{\sqrt{\pi\left(\theta^5 + 5\theta^4 + 18\theta^3 + 44\theta^2 + 48\theta + 48\right) - \pi^2\left(\theta^4 + 4\theta^3 + 16\theta^2 + 24\theta + 36\right)}}{\pi\left(\theta^2 + 2\theta + 6\right)},$$

$$\gamma_{\text{ZMPS}} = \left[\frac{\theta^4+8\theta^3+30\theta^2+96\theta+120}{\theta^2+2\theta+6} - \frac{3\pi\left(\theta^3+4\theta^2+12\theta+24\right)}{\theta^2+\theta+2} + \frac{2\pi^2\left(\theta^2+2\theta+6\right)^2}{\left(\theta^2+\theta+2\right)^2}\right]\frac{\pi\left(\theta^2+2\theta+6\right)}{\theta^3\left(\theta^2+\theta+2\right)h^{3/2}}.$$

and

$$\zeta_{\text{ZMPS}} = \left[\frac{\theta^5+16\theta^4+84\theta^3+336\theta^2+840\theta+720}{\theta^2+2\theta+6} - \frac{4\pi\left(\theta^4+8\theta^3+30\theta^2+96\theta+120\right)}{\theta^2+\theta+2} + \frac{6\pi^2\left(\theta^3+4\theta^2+12\theta+24\right)\left(\theta^2+2\theta+6\right)}{\left(\theta^2+\theta+2\right)^2} - \frac{3\pi^3\left(\theta^2+2\theta+6\right)^3}{\left(\theta^2+\theta+2\right)^3}\right]\frac{\pi\left(\theta^2+2\theta+6\right)}{\theta^4\left(\theta^2+\theta+2\right)h^2}.$$

where

$$h = \frac{\pi\left(\theta^2+2\theta+6\right)}{\theta^2\left(\theta^2+\theta+2\right)}\left[\frac{\theta^3+4\theta^2+12\theta+24}{\left(\theta^2+2\theta+6\right)} - \frac{\pi\left(\theta^2+2\theta+6\right)}{\left(\theta^2+\theta+2\right)}\right].$$

The ZMPS distribution may be considered an interesting alternative to the usual zero-modified Poisson (ZMP) model since the basis distribution of the former can accommodate several levels of overdisperson, issue that the P distribution generally fails in deal with. Table 1 summarizes the nature and the behaviour of the presented measures, using selected values for the parameters θ and π.

Table 1: Theoretical descriptive measures for different values of θ and π.

θ	π	**Measures**					
		μ_{ZMPS}	σ^2_{ZMPS}	τ_{ZMPS}	β_{ZMPS}	γ_{ZMPS}	ζ_{ZMPS}
1.0	0.5	1.1250	3.8594	3.4306	1.7462	2.2984	9.4194
	1.2	2.7000	5.0100	1.8556	0.8290	1.4478	6.0346
3.0	0.5	0.2500	0.4256	1.7024	2.6095	3.4391	18.3409
	1.2	0.6000	0.8114	1.3523	1.5013	1.9409	8.1560
5.0	0.5	0.1281	0.1767	1.3794	3.2809	4.0635	24.0585
	1.2	0.3075	0.3689	1.1997	1.9753	2.3341	9.9550
7.0	0.5	0.0850	0.1066	1.2541	3.8424	4.5498	28.6871
	1.2	0.2039	0.2316	1.1359	2.3597	2.6613	11.5852
9.0	0.5	0.0634	0.0755	1.1909	4.3332	4.9771	32.9873
	1.2	0.1522	0.1677	1.1018	2.6908	2.9534	13.2027

It is clear that the coefficient of variation, the coefficient of skewness, and the coefficient of kurtosis are increasing as θ increases and π decreases. The higher values for the index of dispersion are obtained for small values of θ and π. On the other hand, combining small values of θ with higher as possible values of π will provide bigger values for the expected value and for the variance.

Theorem 1. *The following statements holds.*

i) If $\pi = 0$ then $f_{ZMPS}(0;\theta,\pi) = 1$ and therefore, equation (7) relates to a degenerate distribution with all mass at zero;

ii) If $\pi = 1$ then $f_{ZMPS}(0;\theta,\pi) = f(0;\theta)$. Hence (7) is the usual PS distribution;

iii) *If* $\pi = [1 - f(0;\theta)]^{-1}$ *then* $f_{ZMPS}(0;\theta,\pi) = 0$*;*

iv) *If* $0 < \pi < 1$ *then* $f_{ZMPS}(0;\theta,\pi) > f(0;\theta)$ *and therefore, the ZMPS distribution has a proportion of zeros greater than the usual PS distribution;*

v) *If* $1 < \pi < [1 - f(0;\theta)]^{-1}$ *then* $f_{ZMPS}(0;\theta,\pi) < f(0;\theta)$ *and therefore, the ZMPS distribution has a proportion of zeros smaller than the usual PS distribution.*

Proof. Define the proportion of additional or missing zeros by

$$\begin{aligned} f_{\mathrm{ZMPS}}(0;\theta,\pi) - f(0;\theta) &= (1-\pi) + \pi f(0;\theta) - f(0;\theta) \\ &= (1-\pi)[1 - f(0;\theta)]. \end{aligned} \tag{10}$$

Follows from the previous expression that (i) and (ii) are obvious. As for (iii),

$$\begin{aligned} f_{\mathrm{ZMPS}}(0;\theta,\pi) - f(0;\theta) &= \left\{1 - [1 - f(0;\theta)]^{-1}\right\}[1 - f(0;\theta)] \\ &= f(0;\theta), \end{aligned}$$

hence $f_{\mathrm{ZMPS}}(0;\theta,\pi) = 0$. Statement (iv) follows from the fact that if $0 < \pi < 1$ then $0 < (1-\pi)[1 - f(0;\theta)] < 1$ since f is a probability measure. Therefore $f_{\mathrm{ZMPS}}(0;\theta,\pi) > f(0;\theta)$. For the latter, whichever $\pi > 1$, $(1-\pi) < 0$ and the result follows by the same argument for (iv). Hence, $f_{\mathrm{ZMPS}}(0;\theta,\pi) < f(0;\theta)$, which completes the proof. □

The inflation and deflation of zeros are characterized, respectively, by statements (iv) and (v) of the previous Theorem. We observe from (10) that the main role of the parameter π is to control the frequency of zeros. In such a way, very different values of π lead to completely different ZMPS distributions. For instance, fixing $\theta = 1.5$, if $\pi = 0.05$ then $\mathrm{P}(\mathrm{X} = 0) \approx 0.97$ and if $\pi = 0.95$ then $\mathrm{P}(\mathrm{X} = 0) \approx 0.43$.

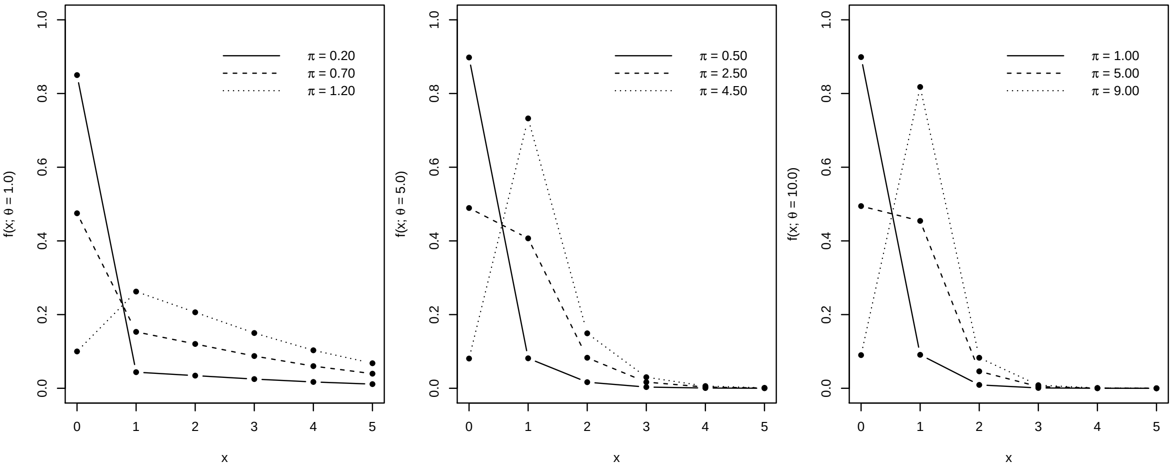


Figure 1: Behaviour of the ZMPS distribution for different values of θ and π.

Figure 1 depicts the pmf of the ZMPS distribution for $\theta = 1.0$ (implying $0 \leqslant \pi \leqslant 1.33$), for $\theta = 5.0$ (implying $0 \leqslant \pi \leqslant 4.90$) and for $\theta = 10.0$ (implying $0 \leqslant \pi \leqslant 9.90$).

4. Hurdle version of the ZMPS distribution

The class of hurdle models was introduced by Mullahy (1986). The relevant feature of such models is that the zero outcomes are treated separately from the positive ones. In the formulation, a binary probability model determines whether a zero or a non-zero outcome occurs and hence, an appropriated truncated discrete distribution is chosen to describe the positive outcomes (Saffar, Adnan, and Greene 2012).

Let us define the hurdle version of the ZMPS distribution. Firstly, the equation (7) can be expressed as

$$f_{\text{ZMPS}}\left(x;\theta,\pi\right)=\left\{1-\pi\left[1-f\left(0;\theta\right)\right]\right\}\delta_x+\pi\left(1-\delta_x\right)f\left(x;\theta\right), \qquad x\in\mathcal{X}_0, \tag{11}$$

Now, setting $\omega=\pi\left[1-f\left(0;\theta\right)\right]$, expression (11) becomes

$$f_{\text{ZMPS}}\left(x;\theta,\omega\right)=(1-\omega)\delta_x+\omega\, f_{\text{ZTPS}}\left(x;\theta\right), \qquad x\in\mathcal{X}_0, \tag{12}$$

being $f_{\text{ZTPS}}(x;\theta)$ the ZTPS distribution given by (6). Since $0\leqslant p\leqslant[1-f(0;\mu)]^{-1}$ then $0\leqslant p\,[1-f(0;\mu)]\leqslant 1$. Hence $0\leqslant\omega\leqslant 1$.

The pmf (12) can be seen as a hurdle version of the ZMPS distribution, where the probability of $\mathrm{X}=0$ is $(1-\omega)$ and the probability of $\mathrm{X}>0$, is $\omega\, f_{\text{ZTPS}}(x;\theta)$. Moreover, from equation (9) we have that the expected value and the variance of the ZMPS distribution, expressed in its hurdle version, depends on the probability of $\mathrm{X}=0$ under the PS distribution.

The ZMPS distribution expressed in a hurdle version contains the ZTPS distribution as one of its components, which differs from the traditional mixture representation of zero-inflated distributions. Indeed, this representation of the ZMPS distribution can be interpreted as a superposition of two processes, i.e. one that produces positive observations from a ZTPS distribution and another that produces only zero valued observations with probability $(1-\omega)$.

The hurdle version of the ZMPS distribution can be used to derive the maximum likelihood estimators (MLEs) for the parameters θ and ω. Furthermore, such approach allow us to use only the positive observations in a given dataset to estimate the parameter θ assuming that these observations comes from a ZTPS distribution, while the parameter ω can be estimated as the proportion of zeros in the sample. Hence, parameter π can be estimated subsequently using the equation

$$\pi=\frac{\omega}{1-f\left(0;\theta\right)}. \tag{13}$$

It is noteworthy that make inferences about the parameter π is essential to identify the kind of zero modification (inflation or deflation) is present in the analysed dataset.

5. Maximum likelihood estimation

Let $\mathbf{X}=(\mathrm{X}_{\mathbf{1}},\ldots,\mathrm{X}_{\mathbf{n}})$ a random sample of size n from the ZMPS distribution and $\boldsymbol{x}=(x_1,\ldots,x_n)$ its observed values. Thus, the likelihood function for the parameters θ and ω is given by

$$\mathrm{L}_n\left(\theta,\omega;\boldsymbol{x}\right)=\prod_{j=1}^{n}\left(1-\omega\right)^{\delta_{x_j}}\left[\omega\, f_{\text{ZTPS}}\left(x_j;\theta\right)\right]^{1-\delta_{x_j}}. \tag{14}$$

One can note that the likelihood values of hurdle models are computed separately for each pmf. The MLEs of θ and ω can be obtained by direct maximization of the log-likelihood function

$$\begin{aligned}\ell_n\left(\theta,\omega;\boldsymbol{x}\right) &= \sum_{j=1}^{n}\left(1-\delta_{x_j}\right)\left\{\log\left[f_{\text{ZTPS}}\left(x_j;\theta\right)\right]+\log\left(\omega\right)\right\}+\sum_{j=1}^{n}\delta_{x_j}\log\left(1-\omega\right)\\ &= \sum_{j=1}^{n}\left(1-\delta_{x_j}\right)\times\end{aligned}$$

$$
\begin{aligned}
& \log\left[\frac{\theta^3}{\theta^4+4\theta^3+10\theta^2+7\theta+2}\left(\frac{x_j^2+x_j\left(\theta+4\right)+\left(\theta^2+3\theta+4\right)}{\left(\theta+1\right)^{x_j}}\right)\right]+ \\
& \sum_{j=1}^{n}\left[\log\left(\omega\right)-\delta_{x_j}\log\left(\frac{\omega}{1-\omega}\right)\right] \\
= & \sum_{j=1}^{m}\log\left[\frac{\theta^3}{\theta^4+4\theta^3+10\theta^2+7\theta+2}\left(\frac{x_j^2+x_j\left(\theta+4\right)+\left(\theta^2+3\theta+4\right)}{\left(\theta+1\right)^{x_j}}\right)\right]+ \\
& \sum_{j=1}^{n}\log\left(\omega\right)-\sum_{j=1}^{q}\log\left(\frac{\omega}{1-\omega}\right) \\
= & \, 3m\log\left(\theta\right)-m\log\left(\theta^4+4\theta^3+10\theta^2+7\theta+2\right)-m\overline{x}_m\log\left(\theta+1\right)+ \\
& n\log\left(\omega\right)-q\log\left(\frac{\omega}{1-\omega}\right)+\sum_{j=1}^{m}\log\left[x_j^2+x_j\left(\theta+4\right)+\left(\theta^2+3\theta+4\right)\right], \qquad (15)
\end{aligned}
$$

where m denotes the number of positive outcomes and q the number of zero ones. Indeed $m+q=n$. Moreover, $\overline{x}_m$ is the sample mean obtained from the set of positive values.

From (15) it is straightforward to see that the parameters θ and ω are orthogonal and that all terms in the log-likelihood function depending on θ take into account only the positive values of the sample vector $\boldsymbol{x}$. Denoting by $\boldsymbol{x}^{(m)}$ the vector of positive values from $\boldsymbol{x}$, the log-likelihood function for θ based on the assumption that $x_j^{(m)}$, $j=1,\ldots,m$, are generated from a ZTPS distribution is given by

$$
\begin{aligned}
\ell_n\left(\theta,\omega;\boldsymbol{x}^{(m)}\right) = & \, 3m\log\left(\theta\right)-m\log\left(\theta^4+4\theta^3+10\theta^2+7\theta+2\right)-m\overline{x}_m\log\left(\theta+1\right)+ \\
& \sum_{j=1}^{m}\log\left[x_j^2+x_j\left(\theta+4\right)+\left(\theta^2+3\theta+4\right)\right]. \qquad (16)
\end{aligned}
$$

Indeed $\ell_n\left(\theta,\omega;\boldsymbol{x}^{(m)}\right)=\ell_m\left(\theta;\boldsymbol{x}\right)$, since each x_j present in the log-likelihood of θ are generated by a zero-truncated distribution. Therefore, evaluate θ under ZMPS distribution is equivalent to assuming that the positive values of $\boldsymbol{x}$ comes entirely from a ZTPS distribution. On the other hand, denoting by $\boldsymbol{x}^{(q)}$ the vector of zero outcomes from $\boldsymbol{x}$, the log-likelihood function for ω is given by

$$
\ell_n\left(\theta,\omega;\boldsymbol{x}^{(q)}\right)=\ell_n\left(\omega;\boldsymbol{x}\right)=n\log\left(\omega\right)-q\log\left(\frac{\omega}{1-\omega}\right). \qquad (17)
$$

Now, the corresponding score vector is given by

$$
\mathrm{U}\equiv\mathrm{U}\left(\theta,\omega;\boldsymbol{x}\right)=\left[u_\theta,u_\omega\right]^\intercal, \qquad (18)
$$

where

$$
\begin{aligned}
u_\theta=\frac{\partial\,\ell_m\left(\theta;\boldsymbol{x}\right)}{\partial\theta} = & \, \frac{3m}{\theta}-m\left[\frac{4\theta^3+12\theta^2+20\theta+7}{\theta^4+4\theta^3+10\theta^2+7\theta+2}\right]-\frac{m\overline{x}_m}{\theta+1}+ \\
= & \sum_{j=1}^{m}\frac{x_j+2\theta+3}{x_j^2+x_j\left(\theta+4\right)+\left(\theta^2+3\theta+4\right)}, \qquad (19)
\end{aligned}
$$

and

$$
u_\omega=\frac{\partial\,\ell_n\left(\omega;\boldsymbol{x}\right)}{\partial\omega}=\frac{\left(n-q\right)}{\omega}-\frac{q}{\left(1-\omega\right)}. \qquad (20)
$$

The observed information, i.e. the Hessian matrix is given by

$$
\mathrm{K}\equiv\mathrm{K}\left(\theta,\omega;\boldsymbol{x}\right)=-\left[\begin{array}{cc} k_{\theta\theta} & k_{\theta\omega} \\ k_{\omega\theta} & k_{\omega\omega}\end{array}\right],
$$

where

$$k_{\theta\theta} = \frac{\partial^2 \ell_m(\theta; \boldsymbol{x})}{\partial \theta^2} = -\frac{3m}{\theta^2} + m\left[\frac{4\theta^6 + 24\theta^5 + 68\theta^4 + 132\theta^3 + 176\theta^2 + 92\theta + 9}{\left(\theta^4 + 4\theta^3 + 10\theta^2 + 7\theta + 2\right)^2}\right] + \frac{m\overline{x}_m}{(\theta+1)^2} + \sum_{j=1}^{m} \frac{x_j^2 - 2x_j\theta + 2x_j - 2\theta^2 - 6\theta - 1}{\left[x_j^2 + x_j(\theta+4) + (\theta^2 + 3\theta + 4)\right]^2}, \qquad (21)$$

and

$$k_{\omega\omega} = \frac{\partial^2 \ell_n(\omega; \boldsymbol{x})}{\partial \omega^2} = -\frac{(n-q)}{\omega^2} - \frac{q}{(1-\omega)^2}. \qquad (22)$$

By orthogonality of θ and ω, the crossed partial derivatives are null, then $k_{\theta\omega} = k_{\omega\theta} = 0$. Hence the set up of the information matrix is complete. As usual, the curvature of the log-likelihood function can be evaluated locally at $\widehat{\theta}$ and $\widehat{\omega}$.

Proposition 1. *Let $\widehat{\omega}$ the MLE of parameter ω. The following statements holds.*

i) $\widehat{\omega} = mn^{-1}$;

ii) $\widehat{\omega}$ *is an unbiased estimator for* ω;

iii) The lower bound for the variance of $\widehat{\omega}$ *is that for binary probability models.*

Proof. Item (i) is straightforward. Take $n - q = m$ and isolate ω in the equation $u_\omega = 0$. Now,

$$\begin{aligned}
\mathrm{E}_{\mathrm{X}}(\widehat{\omega}) &= \mathrm{E}_{\mathrm{X}}\left(\frac{m}{n}\right) \\
&= \frac{1}{n}\mathrm{E}_{\mathrm{X}}\left\{\sum_{j=1}^{n}\left(1 - \delta_{\mathrm{X}_j}\right)\right\} \\
&= \frac{1}{n}\left\{n - \sum_{j=1}^{n}\mathrm{E}_{\mathrm{X}}\left(\delta_{\mathrm{X}_j}\right)\right\} \\
&= \frac{1}{n}\left\{n - \sum_{j=1}^{n} f_{\mathrm{ZMPS}}(0; \theta, \omega)\right\} \\
&= \frac{1}{n}\left\{n - n(1-\omega)\right\} = \omega,
\end{aligned}$$

and (ii) holds. For (iii), firstly note that the set $\{x : f_{\mathrm{ZMPS}}(0; \theta, \omega) > 0\}$ does not depend on θ nor ω. Moreover, it is clear from (20) that for all $x \geqslant 0$, u_ω exists and is finite whenever $\omega \neq 0$. For the moment, such conditions are sufficient and allow us to make use of the Cramér-Rao bound for the variance of an unbiased MLE, which is the case of $\widehat{\omega}$. Then,

$$\mathrm{Var}(\widehat{\omega}) \geqslant \mathrm{J}^{-1}(\omega),$$

where $\mathrm{J} = -\mathrm{E}_{\mathrm{X}}(\mathrm{K})$ is the expected information, i.e. the Fisher information matrix. From (22) we have that

$$\begin{aligned}
-\mathrm{E}_{\mathrm{X}}(k_{\omega\omega}) &= \mathrm{E}_{\mathrm{X}}\left\{\frac{m}{\omega^2} + \frac{(n-m)}{(1-\omega)^2}\right\} \\
&= \frac{1}{\omega^2}\mathrm{E}_{\mathrm{X}}(m) + \frac{n}{(1-\omega)^2} - \frac{1}{(1-\omega)^2}\mathrm{E}_{\mathrm{X}}(m) \\
&= \frac{n}{\omega} + \frac{n}{(1-\omega)^2} - \frac{n\,\omega}{(1-\omega)^2} \\
&= n\left\{\frac{1}{\omega} + \frac{1}{(1-\omega)}\right\} \\
&= \frac{n}{\omega(1-\omega)},
\end{aligned}$$

and hence,

$$\operatorname{Var}\left(\widehat{\omega}\right) \geqslant \frac{\omega\left(1-\omega\right)}{n}, \tag{23}$$

which completes the proof. It is straightforward to show that the variance of $\widehat{\omega}$ is exactly $n^{-1}\omega\left(1-\omega\right)$ and therefore, coincides with the Cramér-Rao lower bound. □

There is no closed form for the MLE of θ, see (19). However, using (17) the parameter θ can be estimated using standard numeric optimization algorithms such the Newton-Raphson, the Bisection and the Regula-Falsi methods. By the usual maximum likelihood theory, an asymptotic approximation for the variance of $\widehat{\theta}$ can be obtained from $k_{\theta\theta}^{-1}$, which evaluated at $\widehat{\theta}$ provides a consistent estimator for such a measure. On the other hand, the variance of $\widehat{\omega}$ can be estimated by its lower bound provided by the previous Proposition which, in fact, corresponds to the exact variance.

As aforementioned, the parameter π is a non-linear function depending on θ and ω. By the invariance principle, the MLE of π can be obtained as

$$\begin{aligned}\widehat{\pi} = s\left(\theta,\omega\right) \quad &= \quad \frac{\widehat{\omega}}{1-f\left(0;\widehat{\theta}\right)} \\ &= \quad \frac{m\left(\widehat{\theta}^2+\widehat{\theta}+2\right)\left(\widehat{\theta}+1\right)^3}{n\left(\widehat{\theta}^4+4\widehat{\theta}^3+10\widehat{\theta}^2+7\widehat{\theta}+2\right)}.\end{aligned}$$

Now, the variance of $\widehat{\pi}$ can be estimated using the delta-method. Since $\widehat{\theta}$ and $\widehat{\omega}$ are orthogonal, $\operatorname{Cov}\left(\widehat{\theta},\widehat{\omega}\right)=0$. Hence,

$$\begin{aligned}\widehat{\operatorname{Var}}\left(\widehat{\pi}\right) \quad &\approx \quad \widehat{\operatorname{Var}}\left(\widehat{\theta}\right)\left[\frac{\partial}{\partial\,\theta}\,s\left(\theta,\omega\right)\right]^2+\widehat{\operatorname{Var}}\left(\widehat{\omega}\right)\left[\frac{\partial}{\partial\,\omega}\,s\left(\theta,\omega\right)\right]^2 \\ &\approx \quad \frac{m^2}{n^2}\,\widehat{\operatorname{Var}}\left(\widehat{\theta}\right)\frac{\widehat{\theta}^4\left(\widehat{\theta}+1\right)^4\left(\widehat{\theta}^4+6\widehat{\theta}^3+25\widehat{\theta}^2+32\widehat{\theta}+24\right)^2}{\left(\widehat{\theta}^4+4\widehat{\theta}^3+10\widehat{\theta}^2+7\widehat{\theta}+2\right)^4}+ \\ &\quad\quad \frac{mq}{n^3}\,\frac{\left(\widehat{\theta}^2+\widehat{\theta}+2\right)^2\left(\widehat{\theta}+1\right)^6}{\left(\widehat{\theta}^4+4\widehat{\theta}^3+10\widehat{\theta}^2+7\widehat{\theta}+2\right)^2},\end{aligned}$$

being the variance of $\widehat{\theta}$ estimated numerically. The terms inside the brackets are evaluated at the MLEs of θ and ω. Now, to obtain intervallic estimates, we can use large sample approximations for the $100\left(1-\alpha\right)\%$ two sided confidence intervals (CIs) for the parameters θ, ω and π that are given, respectively, by

$$\widehat{\theta}\pm z_{\alpha/2}\,\widehat{\text{SE}}\left(\widehat{\theta}\right), \qquad \widehat{\omega}\pm z_{\alpha/2}\,\widehat{\text{SE}}\left(\widehat{\omega}\right) \qquad \text{and} \qquad \widehat{\pi}\pm z_{\alpha/2}\,\widehat{\text{SE}}\left(\widehat{\pi}\right),$$

being z_α the upper α^{th} percentile of the standard Normal distribution. The standard errors (SEs) are estimated as the squared root of the variance of the MLE of each model parameters.

In the following two sections, we presented the results obtained in the simulation study and the application of the proposed model to real datasets. To attain the numerical results, all computations were performed under the R environment (R Development Core Team 2007).

6. Simulation study

In this section, we seek to evaluate the frequentist properties of the proposed methodology by performing a simulation study. The simulation process consists in generating $N = 10,000$

pseudo-random samples of sizes $n = 50, 100, 150$ and 200 of a variable X having ZMPS distribution in its hurdle version. Our procedure is based on the Monte Carlo simulation method to estimate the average bias and the mean squared error of the MLEs of the parameters θ and ω as well the coverage probability of the asymptotic CIs derived from such estimates. The results obtained for the parameter π are also presented. Assuming $\phi = \theta, \omega$ or π, the measures computed using the generated samples are given by

$$\mathrm{B}\left(\widehat{\phi}\right) = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{\phi}_j - \phi\right), \quad \mathrm{MSE}\left(\widehat{\phi}\right) = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{\phi}_j - \phi\right)^2 \quad \text{and} \quad \mathrm{CP}\left(\phi\right) = \frac{1}{N}\sum_{j=1}^{N}\delta_{\mathrm{A}_j},$$

where $\mathrm{A}_j = \left\{\widehat{\phi}_j - z_{\alpha/2}\,\widehat{\mathrm{se}}\left(\widehat{\phi}\right) < \phi < \widehat{\phi}_j + z_{\alpha/2}\,\widehat{\mathrm{se}}\left(\widehat{\phi}\right)\right\}$ and therefore, δ_{A_j} assumes 1 whenever the CI obtained from the j^{th} simulation contains the true value ϕ.

The following algorithm can be used to generate a single random variable from a ZMPS distribution. The process to generate a random sample consists to run the algorithm as often as necessary, say n times. The sequential-search is a black-box type of algorithm and works with any computable probability vector. The main advantage of such procedure is its ease of implementation. More informations on this algorithm can be found at Hörmann, Leydold, and Derflinger (2013).

Algorithm 1 Sequential-Search

```
procedure SeqSea($\theta$, $\omega$)
    Generate $u \sim \mathrm{U}(0,1)$
    Set $x \leftarrow 0$
    Set $p \leftarrow (1-\omega)$
    while $u > p$ do
        Set $x \leftarrow x + 1$
        Set $p \leftarrow p + \omega\, f_{\mathrm{ZTPS}}(x;\theta)$
    end while
    return $x$
end procedure
```

Under ZMPS distribution, the expected number of iterations (NI), i.e. the expected number of comparisons in the while condition is given by

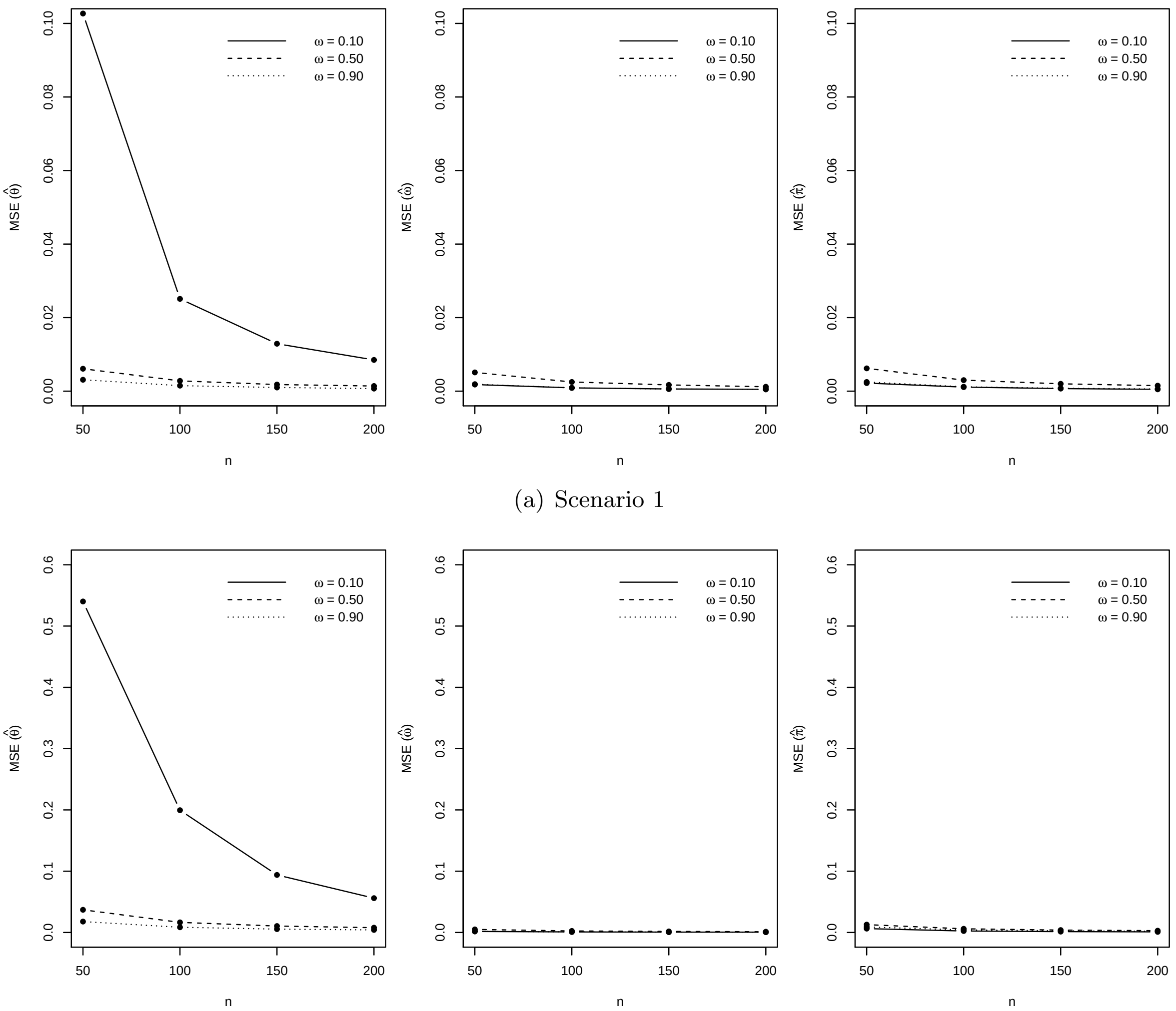
$$\begin{aligned} \mu_{\mathrm{NI}} = \mu_{\mathrm{ZMPS}} + 1 &= \pi\mu + 1 \\ &= \frac{\omega\left(\theta+1\right)^3\left(\theta^2+2\theta+6\right)}{\theta\left(\theta^4+4\theta^3+10\theta^2+7\theta+2\right)} + 1. \end{aligned}$$

To run the simulation, we have established four scenarios in which a single value of θ was chosen for each one. The selected values were $\theta = 0.5, 1.0, 2.0$ and 3.0. Moreover, we consider $\omega = 0.1, 0.5$ and 0.9 varying in each scenario. On the other hand, since parameter π depends on the values of θ and ω, its values vary within each scenario and are closer of ω for small θ and when ω approaches to zero. Further, in order to evaluate if the coverage probability of the asymptotic CIs are around the nominal level of 95%, we fix $\alpha = 0.05$ to compute such measures in the simulation process.

In Table 2, the bias and the coverage probability of the MLEs are presented for each parameter involving the ZMPS model. Figures 2 and 3 depicts the mean squared error of such estimates. The results shows that both bias and mean squared error tends to zero when the sample size increases. It is noteworthy that the MLE of ω is negative biased in some cases. The mean squared error of $\widehat{\omega}$ remains quite small even for small n, which also occurs for $\widehat{\pi}$ the smaller the value of θ. The coverage probabilities were found between 93% and 97% in most cases, indicating that the coverage of the 95% asymptotic CIs is relatively accurate. On the other hand, one can note that for small n, the coverage probability of the CIs obtained for ω decre-

Table 2: Estimated bias of the MLEs and coverage probability of the CIs.

Parameter	**Value**	$n = 50$		$n = 100$		$n = 150$		$n = 200$	
		Bias	**CP**	**Bias**	**CP**	**Bias**	**CP**	**Bias**	**CP**
Scenario 1									
θ	0.50	0.0897	95.47	0.0371	95.29	0.0237	95.47	0.0173	95.39
ω	0.10	0.0014	89.02	0.0003	93.20	-0.0001	92.74	0.0001	92.56
π	0.11	0.0057	93.91	0.0021	95.47	0.0011	95.45	0.0008	95.48
θ	0.50	0.0104	94.97	0.0054	94.94	0.0040	94.97	0.0035	94.87
ω	0.50	-0.0008	93.29	0.0001	94.20	-0.0006	94.35	-0.0004	94.43
π	0.54	0.0020	93.64	0.0015	97.08	0.0003	97.25	0.0004	97.45
θ	0.50	0.0060	94.85	0.0032	94.87	0.0028	94.90	0.0023	94.72
ω	0.90	-0.0001	86.92	-0.0002	93.09	-0.0001	92.34	-0.0002	92.67
π	0.98	0.0027	99.75	0.0013	99.66	0.0010	99.72	0.0007	99.74
Scenario 2									
θ	1.00	0.2274	94.78	0.1156	95.34	0.0701	95.62	0.0490	95.68
ω	0.10	0.0025	89.96	0.0004	93.25	0.0001	92.80	-0.0001	92.48
π	0.13	0.0207	96.62	0.0088	97.20	0.0051	97.71	0.0035	97.85
θ	1.00	0.0299	95.21	0.0146	94.92	0.0101	95.24	0.0080	95.01
ω	0.50	-0.0006	93.34	0.0001	94.23	-0.0004	94.31	-0.0003	94.35
π	0.67	0.0112	99.09	0.0059	99.53	0.0034	99.56	0.0027	99.65
θ	1.00	0.0160	95.05	0.0075	95.01	0.0060	94.92	0.0047	94.82
ω	0.90	-0.0002	86.97	-0.0001	93.10	-0.0001	92.36	-0.0002	92.68
π	1.20	0.0113	99.97	0.0053	99.97	0.0041	99.97	0.0030	99.99
Scenario 3									
θ	2.00	0.4278	93.33	0.4225	94.36	0.2783	95.06	0.1806	95.21
ω	0.10	0.0053	91.71	0.0006	93.50	0.0001	92.87	0.0001	92.55
π	0.21	0.0597	96.46	0.0394	98.12	0.0252	98.98	0.0162	99.25
θ	2.00	0.1179	95.01	0.0551	95.12	0.0372	95.28	0.0289	95.26
ω	0.50	-0.0007	93.29	0.0001	94.22	-0.0004	94.33	-0.0003	94.35
π	1.04	0.0550	99.67	0.0265	99.91	0.0169	99.94	0.0132	99.95
θ	2.00	0.0621	95.68	0.0287	95.20	0.0215	94.24	0.0169	95.06
ω	0.90	-0.0002	86.98	-0.0001	93.10	-0.0001	92.36	-0.0002	92.68
π	1.87	0.0536	99.99	0.0247	99.99	0.0183	99.99	0.0140	99.99
Scenario 4									
θ	3.00	0.3821	91.24	0.7463	93.52	0.6301	94.31	0.4740	94.66
ω	0.10	0.0078	92.69	0.0013	93.99	0.0003	92.97	0.0001	92.70
π	0.30	0.0832	96.70	0.0796	98.51	0.0620	99.22	0.0453	99.46
θ	3.00	0.3197	94.38	0.1425	95.06	0.0914	95.67	0.0690	95.21
ω	0.50	-0.0007	93.25	0.0001	94.22	-0.0004	94.32	-0.0003	94.38
π	1.48	0.1552	99.86	0.0707	99.97	0.0443	99.99	0.0335	99.99
θ	3.00	0.1436	95.05	0.0629	94.99	0.0451	95.14	0.0333	94.94
ω	0.90	-0.0002	86.93	-0.0002	93.07	-0.0001	92.38	-0.0002	92.69
π	2.67	0.1289	99.99	0.0567	99.99	0.0405	99.99	0.0297	99.99



(a) Scenario 1

(b) Scenario 2

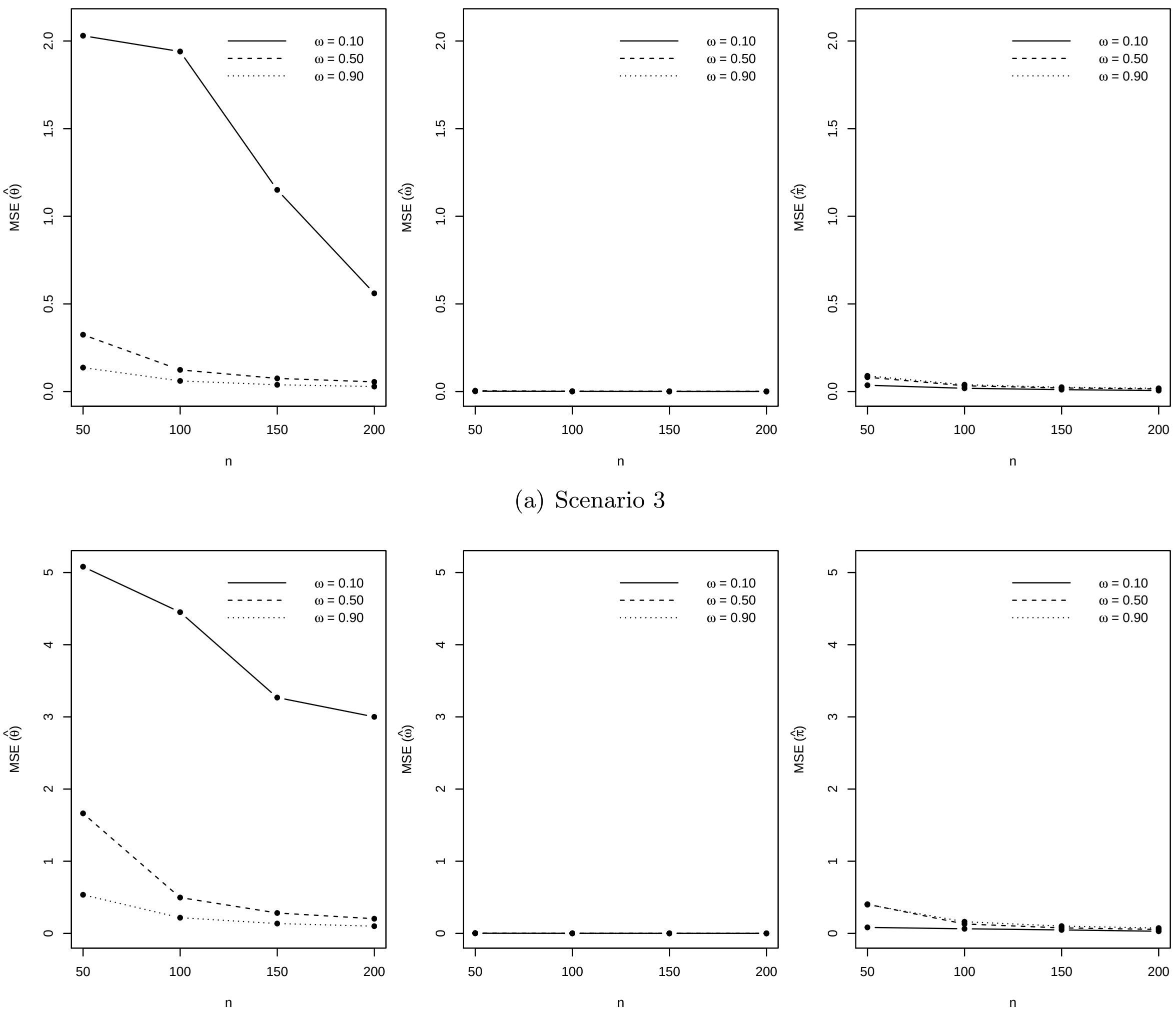
Figure 2: Estimated mean squared error of the MLEs (scenarios 1 and 2).

ases at ≈ 89% when its true value is close to boundaries of the parametric space. The last relevant result is that the coverageprobability of the CIs obtained for π increases as the values of θ and ω increases, attaining values greater than 99% in the last cases of the third and fourth scenarios.

7. Application to real data

In this section, the ZMPS distribution is considered as an attempt to adequately model four datasets from biological science field. The goodness of fit of the proposed model is compared with those accessed by the P, the PS and the ZMP distributions. The first dataset is due to Beall (1940). The response is the number of *Pyrausta nubilalis* observed in small unit areas of a field in 1937. The second one refers to the number of chromatid aberrations observed on chemically induced chromosome aberrations in cultures of human leukocytes (Loeschcke and Köhler 1976). The third and fourth datasets relates to the number of mammalian cytogenetic dosimetry lesions induced in rabbits by *lymphoblast streptonigrin* at the exposure of 60 mg/kg and 90 mg/kg, respectively. A broader description of the last three datasets is provided by Shanker and Fesshaye (2016a).

Table 3 presents some descriptive statistics. The last column shows the observed proportion of zeros (PZ) in each dataset. One can note that more than 50% of the observations are zero-valued in all samples. Also, the initial analysis highlights the presence of overdispersion (see the index of dispersion), justifying the choice of the ZMPS model to describe such data.



(a) Scenario 3

(b) Scenario 4

Figure 3: Estimated mean squared error of the MLEs (scenarios 3 and 4).

In Table 4 we present the frequency distribution of each sample. The expected frequencies was obtained using the estimated probabilities considering the MLEs of the parameters θ and π. Frequencies in bold relate to those one closer to the real values. The results show that the expected frequencies provided by the ZMPS model are the closest in most cases.

Table 3: Variables and descriptive statistics for each dataset.

Label	**Variable**	n	**Mean**	**Variance**	**ID** (%)	**PZ**
Dataset 1	Number of Pyrausta nublilalis	56	0.7500	1.3182	175.76	0.5893
Dataset 2	Number of Chromatid aberrations	400	0.5475	1.1256	205.58	0.6700
Dataset 3	Number of mammalian cytogenetic lesions (E-60)	601	0.4742	0.7398	156.00	0.6872
Dataset 4	Number of mammalian cytogenetic lesions (E-90)	300	0.8533	1.3697	160.51	0.5167

The MLEs, the SEs and the 95% asymptotic CIs for the parameter θ of each fitted model are presented in Table 5. The model selection was performed using the Akaike information

Table 4: Observed and expected frequencies from each fitted model.

Counts	Observed	Expected Frequency			
	Frequency	P	PS	ZMP	ZMPS
		Dataset 1			
0	33	26.45	31.47	**33.00**	**33.00**
1	12	19.84	14.17	10.83	**12.30**
2	6	7.44	6.13	7.34	**5.91**
3	3	1.86	2.55	3.32	**2.71**
4	1	0.35	**1.03**	1.12	1.20
5	1	0.05	0.40	0.30	**0.52**
		Dataset 2			
0	268	231.36	257.61	**268.00**	**268.00**
1	87	126.67	**92.98**	71.81	79.05
2	26	34.68	32.71	40.04	**32.38**
3	9	6.33	**11.19**	14.88	12.80
4	4	0.87	3.73	**4.15**	4.90
5	2	0.09	1.22	0.93	**1.83**
6	1	0.01	0.39	0.17	**0.67**
7	3	0.00	0.12	0.03	**0.24**
		Dataset 3			
0	413	374.05	406.13	**413.00**	**413.00**
1	124	177.38	132.99	115.99	**123.26**
2	42	**42.06**	42.67	52.14	43.02
3	15	6.65	13.37	15.62	**14.61**
4	5	0.79	4.10	3.51	**4.83**
5	0	**0.07**	1.23	0.63	1.56
6	2	0.01	0.36	0.09	**0.50**
		Dataset 4			
0	155	127.80	157.53	**155.00**	**155.00**
1	83	109.05	77.56	71.93	**80.64**
2	33	46.53	**36.42**	45.66	36.81
3	14	**13.24**	16.37	19.32	16.11
4	11	2.82	**7.10**	6.13	6.81
5	3	0.48	**2.99**	1.56	2.79
6	1	0.07	1.23	0.33	**1.12**

criterion with correction for finite samples (AICc) and the Bayesian information criterion (BIC). The goodness of fit was evaluated by the χ^2 statistic. It is noteworthy that the smaller AICc's are provided by the ZMPS model. On the other hand, in some cases the PS distribution presents similar fit when compared with its zero-modified version, as can be seen by that obtained from Dataset 4. In fact, there exist evidences that the proposed model adheres better to the considered datasets and hence, can be considered as a suitable option to model zero inflated/deflated count data in the presence of overdispersion.

The summary for the parameters ω and π can be found at Table 6. Under the ZMPS model, the inference about the parameter π allow us to identify that may exists an evidence that the first three datasets are zero-inflated while the last one may be classified as zero-deflated. By the 95% CIs obtained for the parameter π, we can also conclude that the PS distribution

Table 5: Summary for the parameter θ and comparison criteria.

Dataset	**Model**	$\widehat{\theta}$	SE $\left(\widehat{\theta}\right)$	95% **CI**		**AIC$_\mathbf{c}$**	**BIC**	χ^2
				Lower	**Upper**			
1	P	0.7500	0.1157	0.5232	0.9769	145.24	147.19	24.08
	PS	2.2415	0.3167	1.6208	2.8622	136.03	137.98	1.38
	ZMP	1.3551	0.4196	0.5327	2.1776	62.31	66.14	2.00
	ZMPS	1.9779	0.4516	1.0929	2.8630	**61.88**	**65.71**	**0.53**
2	P	0.5475	0.0370	0.4750	0.6200	881.04	885.02	13473.41
	PS	2.8291	0.1707	2.4945	3.1637	809.33	813.32	71.88
	ZMP	1.1151	0.2388	0.6471	1.5831	322.99	330.94	338.30
	ZMPS	2.3863	0.2476	1.9009	2.8717	**300.23**	**308.18**	**35.41**
3	P	0.4742	0.0281	0.4191	0.5292	1167.36	1171.75	726.48
	PS	3.1258	0.1640	2.8043	3.4472	1115.68	1120.07	9.76
	ZMP	0.8990	0.2731	0.3637	1.4344	376.34	385.12	42.19
	ZMPS	2.8538	0.2722	2.3204	3.3874	**369.61**	**378.38**	**6.16**
4	P	0.8533	0.0533	0.7488	0.9579	802.94	806.63	65.49
	PS	2.0342	0.1183	1.8023	2.2658	767.89	771.58	**3.27**
	ZMP	1.2694	0.1853	0.9063	1.6325	362.08	369.45	13.25
	ZMPS	2.1041	0.1966	1.7187	2.4894	**354.15**	**361.51**	3.35

Table 6: Summary for the parameters ω and π under ZMP and ZMPS models.

Dataset	**Model**	$\widehat{\omega}$	SE $(\widehat{\omega})$	95% **CI**		$\widehat{\pi}$	SE $(\widehat{\pi})$	95% **CI**	
				Lower	**Upper**			**Lower**	**Upper**
1	ZMP	0.411	0.066	0.282	0.540	0.554	0.161	0.237	0.870
	ZMPS					0.846	0.206	0.442	1.249
2	ZMP	0.330	0.024	0.284	0.376	0.491	0.061	0.371	0.611
	ZMPS					0.795	0.092	0.616	0.975
3	ZMP	0.313	0.019	0.276	0.350	0.528	0.055	0.421	0.634
	ZMPS					0.886	0.095	0.700	1.072
4	ZMP	0.483	0.029	0.427	0.540	0.672	0.076	0.523	0.821
	ZMPS					1.046	0.102	0.846	1.247

may be a reasonable choice to model the last two datasets, since the provided CIs contains the value 1. However, in the cases where exist evidence of zero modification, zero-modified models remain preferred.

8. Concluding remarks

In this paper, the ZMPS distribution was introduced as an alternative to model overdispersed count data having inflation or deflation of zeros. We discuss some of its mathematical properties as the expected value, the variance and the coefficients of variation, skewness and kurtosis. Moreover, using the hurdle version of the proposed model we derive the log-likelihood function, the score function, the information matrix and present some properties concerning the MLEs of the model parameters. Also, we performed a simulation study where the bias and the mean squared error of the MLEs as well the coverage probability of the asymptotic CIs were computed and indicated the suitability of the considered methodology. The usefulness of the proposed distribution was evaluated by fitting it to four datasets obtained from biological science field with characteristics of overdispersion and zero modification. The model selection was performed using the AICc and the BIC criteria. The goodness of fit was accessed by the χ^2 statistic. The provided results demonstrate the superiority of the proposed model over the P, the PS and the ZMP distributions, confirming its applicability to model overdispersed and zero-modified count data.

Acknowledgements

We would like to thank the associate editor and the anonymous reviewers for their insightful suggestions to improve the manuscript. In addition, the first two authors would like to thank the Federal Technology University of Paraná for the financial support during this research.

References

Angers JF, Biswas A (2003). "A Bayesian Analysis of Zero–Inflated Generalized Poisson Model." *Computational Statistics & Data Analysis*, **42**(1), 37–46.

Bahn GD, Massenburg R (2008). "Deal with Excess Zeros in the Discrete Dependent Variable, the Number of Homicide in Chicago Census Tract." In *Joint Statistical Meetings of the American Statistical Association*, pp. 3905–12.

Beall G (1940). "The Fit and Significance of Contagious Distributions When Applied to Observations on Larval Insects." *Ecology*, **21**(4), 460–474.

Beuf KD, Schrijver JD, Thas O, Criekinge WV, Irizarry RA, Clement L (2012). "Improved Base–Calling and Quality Scores for 454 Sequencing Based on a Hurdle Poisson Model." *BMC bioinformatics*, **13**(1), 303.

Bohara AK, Krieg RG (1996). "A Zero–Inflated Poisson Model of Migration Frequency." *International Regional Science Review*, **19**(3), 211–222.

Cancho VG, Louzada-Neto F, Barriga GDC (2011). "The Poisson–Exponential Lifetime Distribution." *Computational Statistics & Data Analysis*, **55**(1), 677–686.

Conceição KS, Andrade MG, Louzada F (2013). "Zero–Modified Poisson Model: Bayesian Approach, Influence Diagnostics, and an Application to a Brazilian Leptospirosis Notification Data." *Biometrical Journal*, **55**(5), 661–678.

Dietz E, Böhning D (2000). "On Estimation of the Poisson Parameter in Zero–Modified Poisson Models." *Computational Statistics & Data Analysis*, **34**(4), 441–459.

Gurmu S, Trivedi PK (1996). "Excess Zeros in Count Models for Recreational Trips." *Journal of Business & Economic Statistics*, **14**(4), 469–477.

Heilbron DC, Gibson DR (1990). "Shared Needle Use and Health Beliefs Concerning AIDS: Regression Modeling of Zero–Heavy Count Data. Poster Session." In *Proceedings of the Sixth International Conference on AIDS, San Francisco, CA*.

Hörmann W, Leydold J, Derflinger G (2013). *Automatic Nonuniform Random Variate Generation*. Springer Science & Business Media.

Hu MC, Pavlicova M, Nunes EV (2011). "Zero–Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial." *The American journal of drug and alcohol abuse*, **37**(5), 367–375.

Lambert D (1992). "Zero–Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics*, **34**(1), 1–14.

Loeschcke V, Köhler W (1976). "Deterministic and Stochastic Models of the Negative Binomial Distribution and the Analysis of Chromosomal Aberrations in Human Leukocytes." *Biometrische Zeitschrift*, **18**(6), 427–451.

McDowell A (2003). "From the Help Desk: Hurdle Models." *The Stata Journal*, **3**(2), 178–184.

Mullahy J (1986). "Specification and Testing of Some Modified Count Data Models." *Journal of Econometrics*, **33**(3), 341–365.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ridout M, Demétrio CGB, Hinde J (1998). "Models for Count Data with Many Zeros." In *Proceedings of the XIXth international biometric conference*, volume 19, pp. 179–192.

Saffar SE, Adnan R, Greene W (2012). "Parameter Estimation on Hurdle Poisson Regression Model with Censored Data." *Jurnal Teknologi*, **57**(1), 189–198.

Sankaran M (1970). "The Discrete Poisson–Lindley Distribution." *Biometrics*, **26**(1), 145–149.

Shanker R (2016a). "The Discrete Poisson–Shanker Distribution." *Jacobs Journal of Biostatistics*, **1**(1), 005.

Shanker R (2016b). "Sujatha Distribution and Its Applications." *Statistics in Transition–New series*, **17**(3), 1–20.

Shanker R (2016c). "The Discrete Poisson–Sujatha Distribution." *International Journal of Probability and Statistics*, **5**(1), 1–9.

Shanker R, Fesshaye H (2016a). "On Poisson–Sujatha Distribution and Its Applications to Model Count Data from Biological Sciences." *Biometrics & Biostatistics International Journal*, **3**(4), 1–7.

Shanker R, Fesshaye H (2016b). "Size–Biased Poisson–Sujatha Distribution with Applications." *American Journal of Mathematics and Statistics*, **6**(4), 145–154.

Shanker R, Fesshaye H (2016c). "Zero–Truncated Poisson–Sujatha Distribution with Applications." *Journal of Ethiopian Statistical Association*, **24**, 55–63.

Shanker R, Fesshaye H (2016d). "On Zero–Truncation of Poisson, Poisson–Lindley and Poisson–Sujatha Distributions and Their Applications." *Biometrics & Biostatistics International Journal*, **3**(5), 1–13.

Zamani H, Ismail N (2010). "Negative Binomial–Lindley Distribution and Its Application." *Journal of Mathematics and Statistics*, **6**(1), 4–9.

Zorn CJW (1996). "Evaluating Zero–Inflated and Hurdle Poisson Specifications." *Midwest Political Science Association*, **18**(20), 1–16.

Affiliation:

Wesley Bertoli da Silva
Department of Mathematics
Federal Technology University of Paraná
Curitiba, PR, Brazil
E-mail: wbsilva@utfpr.edu.br

Angélica Maria T. Ribeiro
Department of Mathematics
Federal Technology University of Paraná
Curitiba, PR, Brazil
E-mail: angelicaribeiro@utfpr.edu.br

Katiane S. Conceição
Institute of Mathematical Sciences and Computation
Department of Statistics
University of São Paulo
São Carlos, SP, Brazil
E-mail: katiane@icmc.usp.br

Marinho G. Andrade
Institute of Mathematical Sciences and Computation
Department of Statistics
University of São Paulo
São Carlos, SP, Brazil
E-mail: marinho@icmc.usp.br

Francisco Louzada
Institute of Mathematical Sciences and Computation
Department of Statistics
University of São Paulo
São Carlos, SP, Brazil
E-mail: louzada@icmc.usp.br

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 47 *Submitted:* 2016-08-08
June 2018 *Accepted:* 2016-10-23