doi.org/10.26398/IJAS.0034-001

ON ZERO INFLATED GENERALIZED POISSON-SUJATHA DISTRIBUTION AND ITS APPLICATIONS

Samuel Adewale Aderoju[1]

Department of Statistics and Mathematical Sciences, Kwara State University, Malete, P.M.B. 1530, Ilorin, Kwara State, Nigeria.

Isaac Adeola Adeniyi

Department of Mathematical Sciences, Federal University Lokoja, P.M.B. 1530, Lokoja, Kogi State, Nigeria.

Abstract: *A new model called Zero Inflated generalized Poisson-Sujatha distribution (ZIGPSD) is proposed in this paper. Some characteristics of the model were derived and the maximum likelihood method was used to obtain the estimators of the parameters numerically. Through simulation and application to two datasets, the goodness-of-fit of the ZIGPSD was examined in comparison with the zero inflated Poisson (ZIP), Poisson and the new generalized Poisson-Sujatha (NGPS) models. The results showed that the ZIGPSD provides better fit compared to the other existing models considered in some cases of count data with excess zeros.*

Keywords: *Zero inflated Poisson, Excess zeros, Goodness-of-fit, Moments, Maximum Likelihood*

1. INTRODUCTION

The Poisson distribution is generally considered as the standard model for modeling count data but overdispersion and excess zeros than expected from the Poisson distribution are common problems in modelling count data, which the model cannot handle. Johnson and Kotz (1969) were the first to define a mixture Poisson distribution that accounted for excess zeros in the data (Sirichantra & Bodhisuwan, 2017). However, in some cases, overdispersion is a function of excess zero count or none of it in the data. The count data with excess zeros are common in various fields like physical, natural, biological and social sciences as well as in engineering and agriculture. Regular discrete distribution may fail to fit such data either because of zero inflation or over- or underdispersion (Aryuyuen et al., 2014). There is increased interest in zero inflated distribution to account for extra zeros, which are

[1] Corresponding author: Samuel Adewale Aderoju: E-mail: samuel.aderoju@kwasu.edu.ng

common in count data

"The zero inflated distribution assumes that the observed data are the result of a two-part process that generates structural zeros and a process that generates random counts" said Aryuyuen et al., (2014). They further expressed that the distribution can be simply defined as follows:

$$P(X = x|\omega) = \omega\pi_0(x) + (1 - \omega)f(x|\lambda), \quad (1)$$

where ω is the extra proportion of zeros, *X* is the count variable, *f(x)* is the pmf of *X* with parameter λ, and π_0 *(x)* = 1 if *x* = 0; otherwise π_0 *(x)* = 0.

In this paper, we derived zero inflated generalized Poisson-Sujatha (ZIGPS) distribution by compounding zero inflated Poisson (ZIP) distribution with Tesfay and Shanker's proposed "another two-parameter Sujatha distribution (ATPSD)", Tesfay and Shanker (2019). Some characteristics of the ZIGPS distribution and its application are shown as well.

The rest of this paper is organized as follows: in section 2 the new zero inflated model is presented and some characteristics of the model are derived in section 3. The Maximum likelihood estimation (MLE) method to estimate the parameters of the model are discussed in section 4. Simulation studies are presented in section 5 while section 6 contains the application of the model to two real data sets. The conclusion is presented in section 7.

2. THE ZERO INFLATED GENERALIZED POISSON-SUJATHA DISTRIBUTION

The zero inflated Poisson distribution having probability mass function (pmf) as follows:

$$h(X = x) = \begin{cases} \omega + (1 - \omega)e^{-\lambda}, & if\ \ x = 0 \\ (1 - \omega)\dfrac{\lambda^x}{x!}e^{-\lambda}, & if\ \ x = 1, 2, \ldots \end{cases}, \quad (2)$$

where $0 \leq \omega \leq 1$ and $\lambda > 0$

Suppose the parameter λ in ZIP distribution is a random variable and it follows the ATPSD, such that $\lambda \sim$ *ATPS* (α, β) that with the probability density function (pdf):

$$g(\lambda) = \frac{\beta^3}{\beta^2 + \alpha\beta + 2\alpha}(1 + \alpha\lambda + \alpha\lambda^2)\, e^{-\beta\lambda}; \quad \lambda > 0, \beta > 0, \alpha \geq 0, \quad (3)$$

where β is a scale parameter and α a shape parameter (Tesfay and Shanker, 2019). Obviously, for $\alpha = 0$ and $\alpha = 1$ the ATPSD reduces to exponential distribution and Sujatha distribution respectively.

Definition 1: Suppose a random variable *X* is said to be of zero inflated generalized Poisson-Sujatha distribution (ZIGPSD) if:

$$X|\lambda \sim ZIP(\lambda)$$

$$while \;\; \lambda|\alpha,\beta \sim ATPSD(\alpha,\beta)$$

for $\lambda > 0$, $\beta > 0$ and $\alpha \geq 0$. Hence, we denote the pmf of $X \sim ZIGPS(\alpha, \beta, \omega)$ is given by:

$$f(x) = \begin{cases} h(X=0)g(\lambda), & if \;\; x = 0 \\ (1-\omega)h(X \geq 1)g(\lambda) & if \;\; x = 1,2,\ldots \end{cases} \tag{4}$$

where, $0 < \omega < 1$ is the inflation parameter.

Theorem 1: If $X \sim ZIGPS(\alpha, \beta, \omega)$ be a Zero inflated Generalized Poisson-Sujatha distribution, then the pmf is:

$$f(x) = \begin{cases} \omega + (1-\omega)\dfrac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\left(\dfrac{\alpha(\beta+3)+(\beta^2+2\beta+1)}{(\beta+1)^3}\right), & if \;\; x = 0 \\ (1-\omega)\dfrac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\left(\dfrac{\alpha(x^2+\beta+3)+(\beta+4)\alpha x+(\beta^2+2\beta+1)}{(\beta+1)^{x+3}}\right), & if \;\; x = 1,2,\ldots \end{cases} \tag{5}$$

where $\beta > 0$, $\alpha \geq 0$ and $0 \leq \omega \leq 1$.

***Proof of part 1*:** [if X = 0]: Suppose $X|\lambda \sim ZIP(\lambda)$ *and* $\lambda|\beta, \alpha \sim ATPSD\ (\beta, \alpha)$ then the pmf of conditional random variable X is given as:

$$f_1(x) = \int_0^\infty h(X=0)g(\lambda)d\lambda\,,$$

where $\omega + (1-\omega)h(X=0)$ is ZIP when $x = 0$ and $g(\lambda)$ is ATPSD. Therefore,

$$f_1(x) = \int_0^\infty \omega + (1-\omega)e^{-\lambda}\frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}(1+\alpha\lambda+\alpha\lambda^2)\,e^{-\beta\lambda}d\lambda$$

$$= \omega + (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\int_0^\infty e^{-\lambda(\beta+1)}\,(1+\alpha\lambda+\alpha\lambda^2)\,d\lambda$$

$$= \omega + (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\left[\int_0^\infty e^{-\lambda(\beta+1)}\,d\lambda + \alpha\int_0^\infty \lambda e^{-\lambda(\beta+1)}\,d\lambda + \alpha\int_0^\infty \lambda^2 e^{-\lambda(\beta+1)}\,d\lambda\right]$$

$$= \omega + (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\left[\frac{0!}{(\beta+1)^1}+\frac{\alpha}{(\beta+1)^2}+\frac{2\alpha}{(\beta+1)^3}\right]$$

$$= \omega + (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\left[\frac{\alpha(\beta+3)+(\beta^2+2\beta+1)}{(\beta+1)^3}\right]$$

$$\therefore f_1(x) = \omega + (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\left[\frac{\alpha(\beta+3)+(\beta+1)^2}{(\beta+1)^3}\right]; \quad if\ x=0 \tag{6}$$

Proof of part 2 [if $X \geq 1$]: Suppose $X|\lambda \sim ZIP(\lambda)$ *and* $\lambda|\beta, \alpha \sim ATPSD(\beta, \alpha)$ then the pmf of conditional random variable X is given as:

$$f_2(x) = \int_0^\infty \omega + (1-\omega)h(X \geq 1)g(\lambda)d\lambda\,,$$

where ω + (1- ω) h ($X \geq 1$) is ZIP when $x > 0$ and (g)λ) is ATPSD. Therefore,

$$f_2(x) = \int_0^\infty (1-\omega)\frac{\lambda^x}{x!}e^{-\lambda}.\frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}(1+\alpha\lambda+\alpha\lambda^2)\,e^{-\beta\lambda}d\lambda$$

$$= (1-\omega)\frac{\beta^3}{x!\,(\beta^2+\alpha\beta+2\alpha)}\int_0^\infty \lambda^x e^{-\lambda(\beta+1)}\,(1+\alpha\lambda+\alpha\lambda^2)\,d\lambda$$

$$= (1-\omega)\frac{\beta^3}{x!\,(\beta^2+\alpha\beta+2\alpha)}\left[\int_0^\infty \lambda^x e^{-\lambda(\beta+1)}\,d\lambda + \alpha\int_0^\infty \lambda^{x+1}e^{-\lambda(\beta+1)}\,d\lambda + \alpha\int_0^\infty \lambda^{x+2}e^{-\lambda(\beta+1)}\,d\lambda\right]$$

$$= (1-\omega)\frac{\beta^3}{x!\,(\beta^2+\alpha\beta+2\alpha)}\left[\frac{x!}{(\beta+1)^{x+1}}+\frac{\alpha(x+1)!}{(\beta+1)^{x+2}}+\frac{\alpha(x+2)!}{(\beta+1)^{x+3}}\right]$$

$$= (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\left[\frac{1}{(\beta+1)^{x+1}}+\frac{\alpha(x+1)}{(\beta+1)^{x+2}}+\frac{\alpha(x+2)(x+1)}{(\beta+1)^{x+3}}\right] \tag{7}$$

$$= (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\left[\frac{\alpha(x^2+\beta+3)+(\beta+4)\alpha x+(\beta+1)^2}{(\beta+1)^{x+3}}\right]$$

$$\therefore f_2(x) = (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\left(\frac{\alpha(x^2+\beta+3)+(\beta+4)\alpha x+(\beta+1)^2}{(\beta+1)^{x+3}}\right); \quad x = 1, 2, \ldots,$$

where $0 < \omega < 1$, $\beta > 0$ and $\alpha \geq 0$. Therefore, the full pmf of ZIGPSD is

$$f(x) = \begin{cases} f_1(x), & if\ x = 0 \\ f_2(x), & if\ x = 1, 2, \ldots \end{cases}$$

That is,

$$f(x) = \begin{cases} \omega + (1-\omega)\dfrac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\left(\dfrac{\alpha(\beta+3)+(\beta+1)^2}{(\beta+1)^3}\right), & if\ x = 0 \\ (1-\omega)\dfrac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\left(\dfrac{\alpha(x^2+\beta+3)+(\beta+4)\alpha x+(\beta+1)^2}{(\beta+1)^{x+3}}\right), & if\ x = 1,2,\ldots \end{cases}$$

Note that when $\omega = 0$ equation (5), reduces to new generalized Poisson-Sujatha (NGPS) distribution developed by Aderoju (2020).
Probability mass plots of ZIGPS distribution for particular values of ω, α and β are given in Figure 1 below.

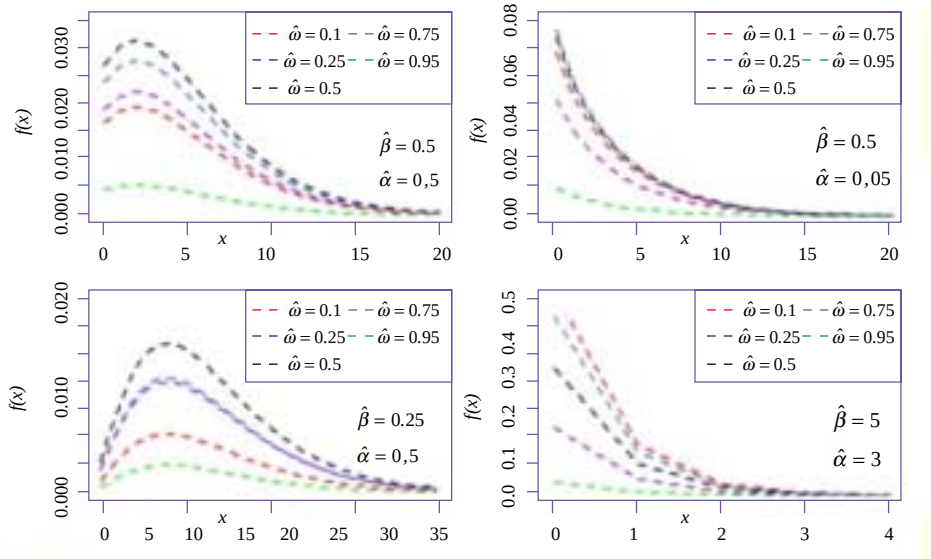

Fig. 1: The pmf of ZIGPS distribution with specified parameters

3. MATHEMATICAL PROPERTIES

The r[th] factorial moment $E[(X)_r]$ of the zero inflated generalized Poisson-Sujatha distribution with random variable X, if is given as:

$$\mu'_r = \frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\int_0^{\infty}(1-\omega)\sum_{x=1}^{\infty}\left[(x)_r\frac{\lambda^x}{x!}e^{-\lambda}\right](1+\alpha\lambda+\alpha\lambda^2)\,e^{-\beta\lambda}d\lambda$$

$$= \frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\int_0^{\infty}(1-\omega)\lambda^r\sum_{x=1}^{\infty}\left[(x)_r\frac{\lambda^{x-r}}{x!}e^{-\lambda}\right](1+\alpha\lambda+\alpha\lambda^2)\,e^{-\beta\lambda}d\lambda$$

Note that $(X)_r = X(X-1)(X-2)\cdots(X-r+1)$, therefore,

$$\mu'_r = \frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\int_0^{\infty}(1-\omega)\lambda^r\sum_{x=1}^{\infty}\left[\frac{\lambda^{x-r}}{(x-r)!}e^{-\lambda}\right](1+\alpha\lambda+\alpha\lambda^2)\,e^{-\beta\lambda}d\lambda$$

Let $z = x-r$,

$$\therefore \mu'_r = \frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\int_0^{\infty}(1-\omega)\lambda^r\sum_{x=1}^{\infty}\left[\frac{\lambda^z}{z!}e^{-\lambda}\right](1+\alpha\lambda+\alpha\lambda^2)\,e^{-\beta\lambda}d\lambda$$

$$\sum_{x=1}^{\infty}\left[\frac{\lambda^z}{z!}e^{-\lambda}\right] = 1, \quad therefore,$$

$$\mu'_r = \frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}(1-\omega)\int_0^{\infty}\lambda^r(1+\alpha\lambda+\alpha\lambda^2)\,e^{-\beta\lambda}d\lambda$$

$$= (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}\left[\frac{r!}{\beta^{r+1}}+\frac{\alpha(r+1)!}{\beta^{r+2}}+\frac{\alpha(r+2)!}{\beta^{r+3}}\right]$$

$$= (1-\omega)\frac{\beta^3}{(\beta^2+\alpha\beta+2\alpha)}r!\left[\frac{1}{\beta^{r+1}}+\frac{\alpha(r+1)}{\beta^{r+2}}+\frac{\alpha(r+2)(r+1)}{\beta^{r+3}}\right]$$

$$\therefore \mu'_r = (1-\omega)r!\left[\frac{\beta^2+\alpha(r+1)(\beta+r+2)}{\beta^r(\beta^2+\alpha(\beta+2))}\right], \qquad r = 1,2,\ldots \tag{8}$$

To obtain the corresponding r^{th} factorial moments for the proposed distribution, the values of $r = 1, 2, 3, \ldots$ will be substituted into (8). Below are results of the first four moments of this distribution. When $r = 1, 2, 3, \ldots$ and 4:

$$\mu_1' = (1-\omega)\frac{\beta^2 + 2\alpha(\beta+3)}{\beta(\beta^2 + \alpha(\beta+2))}$$

$$\mu_2' = (1-\omega)\frac{2[\beta^2 + 3\alpha(\beta+4)]}{\beta^2(\beta^2 + \alpha(\beta+2))}$$

$$\mu_3' = (1-\omega)\frac{6[\beta^2 + 4\alpha(\beta+5)]}{\beta^3(\beta^2 + \alpha\beta + 2\alpha)}$$

$$\mu_4' = (1-\omega)\frac{24[\beta^2 + 5\alpha(\beta+6)]}{\beta^4(\beta^2 + \alpha\beta + 2\alpha)}$$

Hence, the variance (σ^2) of the distribution is obtained as:

$$\sigma^2 = \mu_2' - (\mu_1')^2 + \mu_1'$$

$$= (1-\omega)\frac{(\beta^2 + \alpha(\beta+2))\left[2(\beta^2 + 3\alpha(\beta+4)) + \beta(\beta^2 + 2\alpha(\beta+3))\right] - (1-\omega)(\beta^2 + 2\alpha(\beta+3))^2}{\beta^2(\beta^2 + \alpha(2+\beta))^2}$$

The coefficient of variation is given as

$$CV = \frac{\sqrt{(1-\omega)\left[(\beta^2 + \alpha(\beta+2))\left[2(\beta^2 + 3\alpha(\beta+4)) + \beta(\beta^2 + 2\alpha(\beta+3))\right] - (1-\omega)(\beta^2 + 2\alpha(\beta+3))^2\right]}}{(1-\omega)(\beta^2 + 2\alpha(\beta+3))}$$

The probability generating function (PGF) of the distribution can be expressed as follows:

$$\eta_X(t) = \sum_{x=0}^{\infty} t^x f(x)$$

$$= \omega + (1-\omega)\frac{\beta^3}{\beta^2 + \alpha\beta + 2\alpha}\left[\frac{(\beta+1-t)^2 + \alpha(\beta+1-t) + 2\alpha}{(\beta+1-t)^3}\right] \qquad (9)$$

This PGF can also be used to obtain moments. For example, the first momentis given as $\eta_X'(t=1)$, where

$$\eta'_X(t) = \frac{d\eta_X(t)}{dt} = (1-\omega)\left(\frac{1-2t+2\beta+t^2-2t\beta+\beta^2+2\alpha(4-t+\beta)}{(\beta+1-t)^4(\beta^2+\alpha\beta+2\alpha)}\right).$$

Hence, $$\mu = \eta'_X(t=1) = \frac{d\eta_X(t)}{dt}\bigg|_{t=1} = (1-\omega)\frac{\beta^2+2\alpha(\beta+3)}{\beta(\beta^2+\alpha(\beta+2))}$$

as obtained earlier.

4. MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS

The likelihood function of the ZIGPSD is:

$$\begin{aligned} L(\omega,\alpha,\beta) = \prod_{i=1}^{n}\Bigg[\omega + (1-\omega)\frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\left(\frac{\alpha(\beta+3)+(\beta^2+2\beta+1)}{(\beta+1)^3}\right)\Bigg] \\ + \Bigg[(1 \\ - \omega)\frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\left(\frac{\alpha\left(x_i^2+\beta+3\right)+(\beta+4)\alpha x_i+(\beta^2+2\beta+1)}{(\beta+1)^{x_i+3}}\right)\Bigg] \end{aligned} \tag{10}$$

The log-likelihood function of the *ZIGPS*(α, β, ω) can be expressed as follows:

$$\begin{aligned} \mathcal{L} = \sum_{i=1}^{n} log\Bigg[\omega + (1-\omega)\frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\left(\frac{\alpha(\beta+3)+(\beta^2+2\beta+1)}{(\beta+1)^3}\right)\Bigg] \\ + \sum_{i=1}^{n} log\Bigg[(1 \\ - \omega)\frac{\beta^3}{\beta^2+\alpha\beta+2\alpha}\left(\frac{\alpha\left(x_i^2+\beta+3\right)+(\beta+4)\alpha x_i+(\beta^2+2\beta+1)}{(\beta+1)^{x_i+3}}\right)\Bigg] \end{aligned} \tag{11}$$

The estimates of the parameters in the nonlinear equation (11) can be obtained by numerical optimization using “optim” or “nlm” functions in the R software (R Core Team, 2021).

5. SIMULATION STUDY

In this section, a simulation study to examine the goodness-of-fit performance of the ZIGPS is presented. A maximum likelihood estimation scheme was implemented within the R environment to carryout model fitting for the ZIGPS model. The ZIGPS is compared with the Poisson, ZIP and NGPS distributions. The assessment

is done under various settings of sample sizes and proportion of excess zeros. All simulations were carried out in the R environment.

5.1 SIMULATION SETTING

The simulation settings for the three cases considered are defined as follows.

Case 1: The data are from a Poisson distribution with parameter λ. We considered cases of $\lambda = 5, 50$ at sample sizes of n = 20, 50, 200.

Case 2: The data are from a ZIP distribution with parameters $\lambda = 10$ and $\omega = 0.1$, 0.2, 0.5, 0,8. We considered sample sizes of n = 50, 200.

Case 3: The data are from a ZIGPS distribution with parameters $\alpha = 0.03$, $\beta = 0.8$ and $\omega = 0.1, 0.2, 0.5, 0.8$. We considered sample sizes of n = 50, 200.

5.2 SIMULATION RESULTS

The performance of the methods were evaluated over 100 replications of each case discussed above. The evaluation criteria are: Loglikelihood (Loglik), Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The AIC and BIC are defined as $AIC = -2L + 2p$ and $BIC = -2L + p\,(\log n)$, where L is the loglikelihood, p is the number of parameters to be estimated for the model and n is the number of observations. Tables 1-3 summarizes the means of Loglik, AIC and BIC over 100 replications. It should be noted that higher Loglik and lower AIC and BIC indicate better fit (Hastie, et al., 2001; Adeniyi et al, 2018).

Table 1 presents the results for case 1 where the data are Poisson distributed. The results indicate that at $\lambda = 5$, the performance of Poisson and ZIP are similar across all the sample sizes considered. The performance of the ZIGPS is the least as expected since the data do not contain excess zeros.

The results for case 2, where the data follow the ZIP distribution, are presented in Table 2. Under this setting, the ZIP produced the best performance followed by the ZIGPS across the various levels of excess zeros and sample sizes. The least performance, as expected, was produced by the Poisson distribution as the data were zero-inflated.

Table 3 presents the results for case 3, where the generated data are from the ZIGPS distribution.

Tab. 1: Means of for Loglik, AIC and BIC over 100 replications for Poisson distributed data (CASE 1).

n	λ	Model	Loglik	AIC	BIC
20	***5***	**Poisson**	-43.95	89.90	90.89
		ZIP	-43.81	91.62	93.61
		NGPS	-48.07	100.14	102.13
		ZIGPS	-48.20	102.40	105.39
	50	**Poisson**	-67.28	136.56	137.55
		ZIP	-67.28	138.56	140.55
		NGPS	-87.42	178.84	180.83
		ZIGPS	-92.94	191.89	194.88
50	***5***	**Poisson**	-109.27	220.54	222.45
		ZIP	-109.16	222.32	226.14
		NGPS	-120.31	244.62	248.45
		ZIGPS	-120.36	246.71	252.45
	50	**Poisson**	-167.94	337.87	339.78
		ZIP	-167.94	339.87	343.70
		NGPS	-218.30	440.60	444.43
		ZIGPS	-232.77	471.54	477.28
200	***5***	**Poisson**	-441.32	884.65	887.95
		ZIP	-441.05	886.11	892.70
		NGPS	-482.28	968.57	975.16
		ZIGPS	-482.46	970.91	980.81
	50	**Poisson**	-675.08	1352.16	1355.46
		ZIP	-675.08	1354.16	1360.75
		NGPS	-873.39	1750.77	1757.37
		ZIGPS	-969.52	1945.04	1954.94

Tab. 2: Means of for Loglik, AIC and BIC over 100 replications for Zero-inflated Poisson (ZIP) distributed data (CASE 2).

n	ω	Model	Loglik	AIC	BIC
50	*0.1*	**Poisson**	-160.49	322.98	324.89
		ZIP	-130.54	265.08	268.9
		NGPS	-151.08	306.17	309.99
		ZIGPS	-147.57	301.13	306.87
	0.2	**Poisson**	-188.11	378.22	380.13
		ZIP	-127.21	258.41	262.24
		NGPS	-150.74	305.49	309.31
		ZIGPS	-142.16	290.32	296.06
	0.5	**Poisson**	-234.73	471.46	473.37
		ZIP	-98.03	200.05	203.88
		NGPS	-133.59	271.18	275.01
		ZIGPS	-108.30	222.61	228.34
	0.8	**Poisson**	-182.87	367.74	369.65
		ZIP	-50.23	104.45	108.28
		NGPS	-92.65	189.30	193.13
		ZIGPS	-57.74	121.48	127.22
200	*0.1*	**Poisson**	-649.59	1301.18	1304.47
		ZIP	-524.88	1053.77	1060.36
		NGPS	-605.85	1215.69	1222.29
		ZIGPS	-591.02	1188.03	1197.93
	0.2	**Poisson**	-766.26	1534.52	1537.82
		ZIP	-510.97	1025.93	1032.53
		NGPS	-604.16	1212.32	1218.92
		ZIGPS	-568.17	1142.34	1152.24
	0.5	**Poisson**	-944.54	1891.07	1894.37
		ZIP	-395.70	795.41	802.00
		NGPS	-535.13	1074.26	1080.86
		ZIGPS	-433.96	873.92	883.81
	0.8	**Poisson**	-745.72	1493.45	1496.74
		ZIP	-202.37	408.73	415.33
		NGPS	-376.10	756.20	762.80
		ZIGPS	-235.75	477.5	487.39

Tab. 3: Means of for Loglik, AIC and BIC over 100 replications for Zero Inflated generalized Poisson-Sujatha distribution (ZIGPSD) distributed data (CASE 3).

n	ω	Model	Loglik	AIC	BIC
50	*0.1*	**Poisson**	-91.60	185.20	187.11
		ZIP	-82.07	168.14	171.96
		NGPS	-78.90	161.80	165.62
		ZIGPS	-78.36	162.71	168.45
	0.2	**Poisson**	-90.08	182.16	184.08
		ZIP	-77.23	158.46	162.28
		NGPS	-75.02	154.05	157.87
		ZIGPS	-73.59	153.19	158.92
	0.5	**Poisson**	-73.42	148.83	150.74
		ZIP	-56.27	116.55	120.37
		NGPS	-59.03	122.07	125.89
		ZIGPS	-54.61	115.23	120.96
	0.8	**Poisson**	-43.87	89.74	91.66
		ZIP	-28.90	61.81	65.63
		NGPS	-35.08	74.16	77.98
		ZIGPS	-28.35	62.71	68.44
200	*0.1*	**Poisson**	-373.29	748.57	751.87
		ZIP	-332.34	668.67	675.27
		NGPS	-320.03	644.06	650.66
		ZIGPS	-318.57	643.14	653.04
	0.2	**Poisson**	-358.67	719.34	722.64
		ZIP	-310.75	625.5	632.09
		NGPS	-300.47	604.94	611.53
		ZIGPS	-296.97	599.94	609.84
	0.5	**Poisson**	-294.83	591.66	594.95
		ZIP	-229.11	462.22	468.81
		NGPS	-236.72	477.44	484.04
		ZIGPS	-220.71	447.42	457.32
	0.8	**Poisson**	-171.65	345.3	348.6
		ZIP	-114.04	232.09	238.69
		NGPS	-138.09	280.18	286.78
		ZIGPS	-110.83	227.66	237.55

6. APPLICATIONS TO REAL LIFE DATASETS

To examine the goodness-of-fit of the ZIGPSD in modelling zero inflated as well as overdispersed count data sets, we use two real data sets. The first data referred to the counts of cysts from 111 steroid-treated kidneys (McElduff, et al., 2010; Kumar & Ramachandran, 2019) while the second data were the number of

Mammalian Cytogenetic dosimetry Lesions in Rabbit Lymphoblast induced by *lymphoblast streptonigrin* (NSC-45383), exposure- 60 μg/kg provided by Shanker and Fesshaye (2016). The model evaluation in this section was based on the chi-squared (χ^2) goodness of fit test to compare between observed (O_i) and expected (E_i) values of data as well as the AIC and BIC:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \sim \quad X^2_{n-p}$$

where p is the number of parameters and is the number of classes after the data have been grouped into a frequency distribution.

The observed mean and variance are 1.486 and 7.07027 respectively, which indicate presence of overdispersion. Moreover, the expected zero counts in the data are 25, that is, $111 \times e^{-1.486} = 25.104$ while there are 65 zeros in the data, which indicate presence of excess zero. With these evidences, it is clear that the Poisson model is not appropriate for this data set.

Going by the rule proposed by Lawal (1980), that expected value can be as small as $\frac{r}{d^{3/2}}$ (where r is the number of expected values less than 3 and $\boldsymbol{d}$ is the degree of freedom under such model) without violating the X^2 assumption. Hence, the minimum expected values required from the ZIGPSD for this dataset 1 will be $\frac{7}{9^{3/2}} = 0.2592$; only 7 expected values are less than 3 and the degrees of freedom *df* = 9 Hence, since there is no expected value less than 0.259, there is no need to collapse cells. However, the minimum expected values from Poisson, ZIP and NGPS distributions are 0.2192, 0.1897 and 0.2214 respectively. Therefore, cells less than the values were collapsed as shown in Table 1 and the degree of freedom adjusted appropriately.

Based on the computed values of X^2 statistic given in Table 4, it is obvious that the ZIGPSD fits the data well (p-value = 0.7675) while ZIP (pvalue < 0.0001) and NGPS (pvalue < 0.0001) distributions failed to fit the data at 5% significant level. It is worthy of note that Kumar & Ramachandran (2019) fitted zero inflated Hermite (ZIH), zero inflated generalized Poisson (ZIGP), Negative Binomial (NB) and zero inflated Negative Binomial (ZINB) distributions among others to this same data set. Their results showed that even ZINB distribution failed woefully at fitting the data, only ZIH distribution slightly fit at 5% level of significance (having pvalue = 0.0914).

Tab. 4: Observed and expected frequencies from fitted models to the two data set on counts of cysts from 111 steroid-treated kidneys

X (count)	Observed Freq.	Poisson	ZIP	NGPS	ZIGPS
0	65	25.1	65.0	45.3	65.0
1	14	37.3	5.1	26.4	13.1
2	10	27.7	8.9	15.6	9.2
3	6	13.8	10.3	9.3	6.6
4	4	$7.1 = \begin{cases} 5.1 \\ 1.5 \\ 0.4 \\ 0.1 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{cases}$	8.9	5.6	4.7
5	2		6.2	3.4	3.4
6	2		3.6	2.1	2.5
7	2		1.8	1.3	1.8
8	1		0.8	0.8	1.3
9	1		$0.4 = \begin{cases} 0.3 \\ 0.1 \\ 0.0 \\ 0.0 \end{cases}$	0.5	1.0
10	1			$0.7 = \begin{cases} 0.3 \\ 0.2 \\ 0.2 \end{cases}$	0.7
11	2				0.5
12	1				1.2
Total	111	111	111	111	111
MLE		$\hat{\lambda} = 1.4865$	$\hat{\lambda} = 3.4760$ $\hat{\omega} = 0.5724$	$\hat{\alpha} = 0.0311$ $\hat{\beta} = 0.8074$	$\hat{\alpha} = 0.0263$ $\hat{\beta} = 0.5280$ $\hat{\omega} = 0.4162$
Loglik		-263.2722	-191.866	-184.8609	-172.8206
X^2 ***(P-value)***		104.8565 (<0.0001)	76.6887 (<0.0001)	35.0950 <0.0001	5.720 (**0.7675**)
df		3	**7**	7	9
AIC		528.5444	387.7321	373.7219	**351.6413**
BIC		526.5444	383.7321	374.6917	**359.7699**

Tab. 5: Observed and expected frequencies from fitted models to the data set on number of Mammalian Cytogenetic dosimetry Lesions in Rabbit Lymphoblast induced by streponigrin (NSC-45383), *exposure - 60 µg/kg.*

X (count)	Obs. Freq	Poisson	ZIP	NGPS	ZIGPS
0	413	374.0	413.0	408.6	413.0
1	124	177.4	116.0	130.1	124.1
2	42	42.1	52.1	42.0	42.1
3	15	6.6	15.6	13.7	14.4
4	5	$0.9 = \begin{cases} 0.8 \\ 0.1 \\ 0.0 \end{cases}$	3.5	4.5	4.9
5	0		0.6	1.5	1.7
6	2		0.2	0.6	0.8
Total	601	601	601	601	601
MLE		$\hat{\lambda} = 0.4742$	$\hat{\lambda} = 0.8990$ $\hat{\omega} = 0.4725$	$\hat{\alpha} = 0.1472$ $\hat{\beta} = 2.4154$	$\hat{\alpha} = 0.2514$ $\hat{\beta} = 2.4160$ $\hat{\omega} = 0.0702$
Loglik		-582.6775	-559.5806	-556.4111	**-556.1823**
X^2 ***(P-value)***		72.1766 (<0.0001)	19.9756 (0.0005)	0.5171 (0.2598)	3.5273 (**0.3172**)
df		3	4	4	3
AIC		1167.355	1123.161	1116.822	**1112.365**
BIC		1169.247	1119.161	1116.406	**1112.365**

The observed mean and variance are 0.4742 and 0.7397 respectively, which indicate presence of overdispersion. Moreover, the expected zero counts in the data are 374, that is, $601 \times e^{-0.4742} = 374$ while there are 413 zeros in the data, which indicate excess zeros. With these evidences, it is clear that Poisson model is not appropriate for this second data set.

Similarly, based on the values of X^2 statistic in Table 5, the results shown that ZIGPSD (with pvalue = 0.3172) and NGPSD (pvalue = 0.2598) fits the data well while Poisson (pvalue < 0.0001) and ZIP (pvalue = 0.0005) distributions failed to fit the data at 5% level of significance.

Moreover, the minimum expected values required from the ZIGPSD for this dataset will be $\frac{2}{3^{3/2}} = 0.3849;$ only two expected values are less than 3 and the *df* = (7-3-1) = 3. So, there is no need to collapse cells except for Poisson model - where we collapse cells X = 4, 5 and 6.

7. CONCLUSION

In this paper we proposed the zero inflated Generalized Poisson-Sujatha (ZIGPS) distribution and derived some of its mathematical characteristics. Maximum Likelihood Estimates of the parameters through direct maximization of the log-likelihood function is proposed and implemented numerically using the R software. A simulation study was used to examine the goodness-of-fit performance of the ZIGPS distribution. Results from simulation study revealed that the ZIGPS - distribution is a good alternative for modelling count data with excess zeros.) = Application of the model was made to two real data sets. It has also been shown that the proposed model fits the two data sets well at 5% significant level. Note that the model performs extremely well in the two data sets, which makes it a model to reckon with in modeling zero inflated count data that is highly overdispersed. R codes used in this work are available on request from the authors.

ACKNOWLEDGMENTS

The authors greatly appreciate the anonymous reviewers for their helpful comments and suggestions towards the revision of this paper.

REFERENCES

Adeniyi, I.A., Yahya, W.B., & Ezenweke, C.P. (2018). A Note on Pharmacokinetics Modelling of Theophylline Concentration Data on Patients with Respiratory Diseases. *Turkiye Klinikleri Journal of Biostatistics*, *10*(1), 27-45. doi:10.5336/biostatic.2017-58451.

Aderoju, S.A (2020). A New Generalized Poisson-Sujatha Distribution and its Applications. Applied Mathematical Sciences, 14(5), 229–234. https://doi.org/10.12988/ams.2020.914185

Aryuyuen, S., Bodhisuwan, W. and Supapakorn T., (2014), Zero Inflated negative binomial-generalized exponential distribution and its applications; J. Sci. Technol. 36 (4), 483-491.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Johnson, N. and Kotz, S. (1969). *Discrete distributions, Distributions in statistics*. Houghton Mifflin,

Kumar, C.S and Ramachandran, R. (2019): On some aspects of a zero-inflated overdispersed model and its applications, *Journal of Applied Statistics, DOI: 10.1080/02664763.2019.1645098*

Lawal, H. B. (1980), Tables of percentage points of Pearson's goodness-of-fit statistic for use with small expectations, *Applied Statistics*, **29** (1980), 292–298. https://doi.org/10.2307/2346904

McElduff, F., Cortina-Borja, M., Chan, S.K. and Wade, A. (2010). When t-tests or Wilcoxon-Mann-Whitney tests won't do, Adv. Physiol. Educ. 34 , pp. 128-133

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Sirichantra C.'and Bodhisuwan W. (2017).Parameter Estimation of the Zero Inflated Negative Binomial Beta Exponential Distribution.

AIP Conference Proceedings 1905, 050041. https://doi.org/10.1063/1.5012260

Tesfay, M and Shanker R (2019). Another Two-Parameter Sujatha Distribution with Properties and Applications. J*ournal of Mathematical Sciences and Modelling, 2* (1), 1-13. DOI: http://dx.doi.org/10.33187/jmsm.416628

Winkelmann, D. R. (2000). *Econometric Analysis of Count Data;* Springer Berlin Heidelberg.