REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000-000

A Novel Discrete Distribution Based on the Mixture of Poisson and Sum of Two Lindley Random Variables

Authors: Christophe Chesneau ✉
– Université de Caen Basse-Normandie, LMNO,
UFR de Sciences, F-14032, Caen, France. (christophe.chesneau@gmail.com)

Veena D'cruz
– Department of Statistics, Cochin University of Science and Technology,
Cochin 682 022, Kerala, India. (veenadicruz@gmail.com)

R. Maya
– Department of Statistics, Govt. College for Women,
Trivandrum 695 014, Kerala, India. (publicationsofmaya@gmail.com)

M.R. Irshad
– Department of Statistics, Cochin University of Science and Technology,
Cochin 682 022, Kerala, India. (irshadm24@gmail.com)

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- The distribution of the sum of two independent Lindley random variables having the same parameter defines the 2S-Lindley distribution recently introduced in the statistical literature. In this paper, we use it to create a new one-parameter discrete compound distribution called the Poisson 2S-Lindley distribution. More precisely, it is obtained by compounding the Poisson and 2S-Lindley distributions with the idea of combining their desirable functionalities. The mathematical and statistical properties of the proposed distribution are investigated and analyzed systematically. Subsequently, a statistical standpoint is adopted. Various estimation methods are considered, and an extensive simulation study is conducted to compare them. A count regression model as well as a first-order integer-valued autoregressive process based on the proposed model are constructed. In total, four real data sets in different fields are used to prove the empirical importance of the distribution.

Keywords:

- *2S-Lindley distribution; Compounding; Moments; Stress strength reliability ; Simulation; Count regression model; INAR(1) process*

AMS Subject Classification:

- 60E05, 62E15, 62F10.

1. INTRODUCTION

Numerous applied as well as theoretical fields, such as medicine, sports, insurance, etc., deal with a large number of count data sets. Hence, count data modelling is of paramount importance. The Poisson (P) and negative binomial (NB) distributions are the most commonly used for this purpose. When an over-dispersion problem occurs, the P distribution is not considered ideal due to its equi-dispersion nature. The NB distribution is then more useful. However, when data sets have large skewness and kurtosis, the NB distribution is not well-adapted, and thus we look for more flexible distributions.

Mixed P distributions were then introduced for modelling non-homogeneous populations (see [20] for more details). It consists of mixing the P distribution with well-known distributions. Mixed P distributions include the P Weibull distribution in [9], P transmuted Lindley distribution in [1], P transmuted exponential distribution in [10], P quasi Lindley distribution in [15], new P weighted exponential distribution in [3], P Bilal distribution in [4], and P xgamma distribution in [8], among others.

For other purposes, the Lindley distribution was originally proposed in [21] in the context of Bayesian statistics. Various extensions and generalizations of the Lindley distribution have been proposed in the literature over the last decade. The authors in [13] developed a continuous probability distribution with simple operations on two independent Lindley random variables. In particular, the distribution of the sum of two independent Lindley random variables with the same parameter is studied and named the 2S-Lindley (2S-L) distribution. They investigated the applications and structural properties of those models.

On the other hand, to model over-dispersed data, various count regression models based on many discrete distributions, including mixed P distributions with auxiliary variables, were introduced recently to predict the behavior of the count response variable similar to that of the P regression model. A count regression model based on the generalized Waring distribution was developed in [26]. A study on crash data was done in [23] based on the NB-Lindley regression model. The authors in [27] introduced the hyper-P regression model. The authors in [12] analyzed the P Weibull regression model with real-life data sets. The P-weighted exponential distribution with its count regression model was developed in [32]. The authors in [10] introduced a regression model for the over-dispersed dependent variable based on the P transmuted exponential distribution. Also, many authors compounded P with various generalized versions of the Lindley distribution and hence introduced count regression models based on it (see [24], [31], and [5]).

When the P distribution is used as the innovation distribution for the INAR(1) process, equi-dispersion (empirical mean equals empirical variance) turns out to be a disadvantage for modeling over-dispersed (empirical mean less than empirical variance) time series of counts, primarily in [25] and [2], independently. Hence, to solve this, researchers came up with various alternatives. The INAR(1) process with P-Lindley innovations (INAR(1)PL) in [22], INAR(1) process with new P weighted exponential innovations ((INAR(1)NPWE)) in [3], INAR(1) process with discrete three-parameter Lindley innovations in [14], INAR(1) process with P quasi xgamma innovations (INAR(1)PQX) in [7], INAR(1) process with Bell innovations (INAR(1)BL) in [17], INAR(1) with P transmuted exponential innovations (INAR(1)PTE) in [6], INAR(1) with discrete pseudo Lindley innovations (INAR(1)DPsL) in [18] are some of the recently developed over-dispersed INAR(1) processes.

In this paper, we construct a new one-parameter compound distribution based on the P distribution and the probability density function (pdf) of the 2S-L distribution as described above. The newly proposed distribution, named the P2S-L distribution, is a one-parameter distribution having simple and explicit forms for its probability mass function (pmf), cumulative distribution function (cdf), moments, probability generating function (pgf), moment generating function (mgf), stress-strength reliability function, etc. Under the P2S-L distribution, a count regression model and an INAR(1) process are proposed to model over-dispersed data sets. Four data sets are used to illustrate the capability of the P2S-L distribution.

The important feature of the P2S-L distribution is that it has a simpler explicit expression for moments as well as other statistical properties, and the P2S-L distribution is effective in modelling over-dispersed data sets. Furthermore, the 2S-L distribution was chosen over the Lindley and exponential distributions because the convolution operation defining its pdf makes it more flexible and suitable.

The rest of the paper is organized as follows: In Section 2, the proposed distribution is defined and some statistical properties are derived. Section 3 describes the estimation methods used to estimate the unknown parameter, and their finite sample performance through a simulation study is checked. Section 4 introduces a new count regression model, and Section 5 defines a new INAR(1)P2S-L process. Section 6 deals with four real-life data sets to prove the potential of the proposed distribution. Section 7 includes concluding observations.

2. THE POISSON 2S-LINDLEY DISTRIBUTION

Let us first define the 2S-L distribution proposed in [13]. Suppose that Y_1 and Y_2 are two independent random variables following the one-parameter Lindley distribution in [21] with the same parameter denoted by θ. Then the random variable Y defined as the sum of Y_1 and Y_2, i.e., $Y = Y_1 + Y_2$, possesses the following pdf:

$$f(y;\theta) = \frac{\theta^4}{(1+\theta)^2} y \left(\frac{y^2}{6} + y + 1\right) e^{-\theta y}, \quad y, \theta > 0.$$

It is understood that $f(y;\theta) = 0$ for $y \leq 0$. This pdf characterizes the 2S-L distribution. The corresponding cdf is given as

$$F(y;\theta) = 1 - \frac{1}{6(1+\theta)^2}\left[\theta^3 y\left(y^2+6y+6\right) + 3\theta^2\left(y^2+4y+2\right) + 6\theta(y+2) + 6\right] e^{-\theta y},$$
$$y, \theta > 0.$$

The 2S-L distribution is the basis of our proposed one-parameter discrete compound distribution, which is mathematically formulated in the definition below.

Definition 2.1. Suppose that X is a random variable having the Poisson 2S-Lindley (P2S-L) distribution such that

$$X|\lambda \sim \mathrm{P}(\lambda) \quad \text{and} \quad \lambda|\theta \sim \text{2S-L}(\theta),$$

where $\lambda > 0, \theta > 0$, and $\sim \mathcal{D}$ denotes "follows the $\mathcal{D}$ distribution". Then, the unconditional pmf of X is

$$P(x;\theta) = \frac{\theta^4(1+x)}{6(1+\theta)^{6+x}}\left[x^2+6(\theta+2)^2+x(11+6\theta)\right], \quad x=0,1,2,\ldots \tag{2.1}$$

The distribution defined with the pmf (2.1) is referred to as the P2S-L distribution with parameter θ.

Proof: Using the procedure of compounding and integral developments, we have

$$\begin{aligned}
P(x;\theta) &= \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\frac{\theta^4}{(1+\theta)^2}\lambda\left(\frac{\lambda^2}{6}+\lambda+1\right)e^{-\theta\lambda}\mathrm{d}\lambda \\
&= \frac{\theta^4}{x!(\theta+1)^2}\left[\frac{1}{6}\int_0^\infty e^{-\lambda(\theta+1)}\lambda^{x+3}\mathrm{d}\lambda + \int_0^\infty e^{-\lambda(\theta+1)}\lambda^{x+2}\mathrm{d}\lambda + \int_0^\infty e^{-\lambda(\theta+1)}\lambda^{x+1}\mathrm{d}\lambda\right] \\
&= \frac{\theta^4(1+x)}{6(1+\theta)^{6+x}}\left[x^2+6(\theta+2)^2+x(11+6\theta)\right].
\end{aligned}$$

The proof is completed. □

After some mathematical developments, the corresponding cdf of the P2S-L distribution is

$$\begin{aligned}
F(x;\theta) = \frac{1}{6(1+\theta)^{6+x}}\{6\left(-1+(1+\theta)^x\right) &+ \theta\left[-6(6+x)-3(5+x)(6+x)\theta\right. \\
&- (4+x)(5+x)(6+x)\theta^2 - 6(11+x(7+x))\theta^3 - 6(2+x)\theta^4 \\
&\left.+ 6(1+\theta)^x(2+\theta)\left(1+\theta+\theta^2\right)(3+\theta(3+\theta))\right]\}.
\end{aligned}$$

Proposition 1. The pmf given in (2.1) is log concave.

Proof: We need to prove that $\frac{P(x+1;\theta)}{P(x;\theta)}$ is a decreasing function in x. We have

$$\frac{P(x+1;\theta)}{P(x;\theta)} = \frac{(2+x)\left[x^2+6(2+\theta)(3+\theta)+x(13+6\theta)\right]}{(1+x)(1+\theta)\left[x^2+6(2+\theta)^2+x(11+6\theta)\right]}.$$

Hence, by considering this ratio as a continuous function with respect to x and taking its derivative, we obtain

$$\frac{\partial}{\partial x}\frac{P(x+1;\theta)}{P(x;\theta)} = -\frac{3\left[\frac{1}{\theta+1}+\frac{(4\theta+4x+10)+2(x+1)(2\theta(x+1)+3x+4)}{(6(\theta+2)^2+x^2+(6\theta+11)x)^2}\right]}{(x+1)^2},$$

which is clearly negative, implying that $\frac{P(x+1;\theta)}{P(x;\theta)}$ is a decreasing function in x for every possible value of θ. This ends the proof. □

Figure 1 shows the possible pmf shapes of the P2S-L distribution for various values of θ.

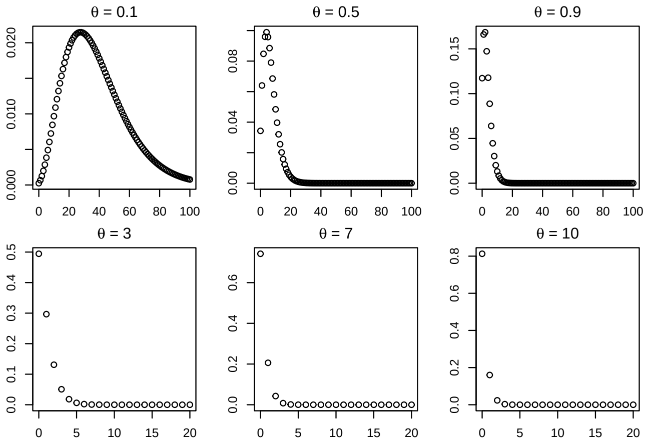

Figure 1: Some plots of the pmf of the P2S-L distribution for chosen values for θ.

The figure clearly indicates that the P2S-L distribution is right-skewed and, as θ goes larger, the mass will concentrate more on nearer to 0 values.

2.1. Moments, Skewness and Kurtosis

Some moment functions, skewness and kurtosis of the P2S-L distribution are now under investigation. If X represents a random variable having the P2S-L distribution, then the pgf of X is

$$G(s) = \mathrm{E}(s^X) = \sum_{x=0}^{\infty} s^x P(x;\theta) = \frac{\theta^4(2-s+\theta)^2}{(1+\theta)^2(1-s+\theta)^4}, \tag{2.2}$$

for $|s| < 1+\theta$. The mgf of X can be obtained by substituting s in (2.2) by e^t, and we obtain

$$M(t) = \mathrm{E}(e^{tX}) = \frac{\theta^4\left(2-e^t+\theta\right)^2}{(1+\theta)^2\left(1-e^t+\theta\right)^4},$$

for $t \leq \ln(1+\theta)$. The r^{th} factorial moment of X is described in the next result.

Proposition 2. The r^{th} factorial moment of X is derived as

$$\mu_{[r]} = \mathrm{E}(X(X-1)\ldots(X-r+1)) = \frac{\left[r^2+6(1+\theta)^2+r(5+6\theta)\right](r+1)!}{6\theta^r(1+\theta)^2}.$$

Proof: The r^{th} factorial moment of X is obtained by integrating. That is,

$$\mu_{[r]} = \int_0^\infty \frac{\theta^4}{(1+\theta)^2}\lambda^{r+1}\left(\frac{\lambda^2}{6}+\lambda+1\right)e^{-\theta\lambda}\mathrm{d}\lambda$$
$$= \frac{\theta^4}{(\theta+1)^2}\left(\frac{1}{6}\int_0^\infty \lambda^{r+3}e^{-\theta\lambda}\mathrm{d}\lambda + \int_0^\infty \lambda^{r+2}e^{-\theta\lambda}\mathrm{d}\lambda + \int_0^\infty \lambda^{r+1}e^{-\theta\lambda}\mathrm{d}\lambda\right)$$
$$= \frac{\left[r^2+6(1+\theta)^2+r(5+6\theta)\right](r+1)!}{6\theta^r(1+\theta)^2}.$$

Hence the proof. □

Based on Proposition 2, the first four moments of X follow as

$$\mathrm{E}(X) = \frac{2(2+\theta)}{\theta(1+\theta)}, \quad \mathrm{E}(X^2) = \frac{2\left(\theta^3+6\theta^2+14\theta+10\right)}{\theta^2(\theta+1)^2},$$
$$\mathrm{E}(X^3) = \frac{2\left(\theta^4+12\theta^3+50\theta^2+90\theta+60\right)}{\theta^3(\theta+1)^2},$$
$$\mathrm{E}(X^4) = \frac{2\left(\theta^5+24\theta^4+158\theta^3+490\theta^2+720\theta+420\right)}{\theta^4(\theta+1)^2}$$

and the variance of X is

$$\mathrm{Var}(X) = \frac{2\left(\theta^3+4\theta^2+6\theta+2\right)}{\theta^2(\theta+1)^2}.$$

It is worth noting that when θ tends to 0, the above moment measures tend to ∞, and when θ tends to ∞, they tend to 0. The skewness and kurtosis of the P2S-L distribution can be calculated using the following formulas:

$$\mathrm{Skewness}(X) = \frac{\mathrm{E}(X^3)-3\mathrm{E}(X^2)\mathrm{E}(X)+2[\mathrm{E}(X)]^3}{[\mathrm{Var}(X)]^{3/2}} \tag{2.3}$$

and

$$\mathrm{Kurtosis}(X) = \frac{\mathrm{E}(X^4)-4\mathrm{E}(X^3)\mathrm{E}(X)+6\mathrm{E}(X^2)[\mathrm{E}(X)]^2-3[\mathrm{E}(X)]^4}{[\mathrm{Var}(X)]^2}, \tag{2.4}$$

respectively. Both equations (2.3) and (2.4) will result in explicit forms.

2.2. Dispersion Index and Coefficient of Variation

The expressions of the coefficient of variation and dispersion index of the P2S-L distribution are obtained in this section. The coefficient of variation (CV) is calculated as

$$\mathrm{CV}(X) = \frac{\sqrt{2\left(\theta^3+4\theta^2+6\theta+2\right)}}{2(\theta+2)}.$$

The dispersion index (DI) is given by

$$\mathrm{DI}(X) = 1+\frac{1}{\theta}+\frac{1}{\theta+1}-\frac{1}{\theta+2}.$$

Clearly, DI(X) is always greater than 1. As a result, the P2S-L distribution has over-dispersed properties.

Numerical values for the mean, variance, DI, skewness, and kurtosis for different sets of parameter values are reported in Table 1.

Table 1: Numerical values for some moment measures of the P2S-L distribution for various values of θ.

Measures	θ					
	0.1	0.5	1.9	2.0	5.0	8.0
Mean	38.1818	6.6667	1.4156	1.3333	0.4667	0.2778
Variance	436.5289	21.7778	2.2858	2.1111	0.5711	0.3156
DI	11.4329	3.2667	1.6147	1.5833	1.2238	1.1361
Skewness	1.0120	1.1297	1.4551	1.4731	1.9126	2.2638
Kurtosis	4.5261	4.8364	5.9078	5.9709	7.5983	9.0977

From Table 1, the P2S-L distribution would be appropriate for modelling over-dispersed, right-skewed and leptokurtic data sets.

2.3. Stress-Strength Analysis

Stress-strength reliability has wide applications in almost all fields of engineering and machine learning. Let X_{stress} and $X_{strength}$ be discrete random variables with positive integer values that model the stress and strength of a system, respectively. Then, the expected reliability can be calculated using the following formula:

$$R_{\text{Stress-Strength}} = \Pr\left(X_{\text{Stress}} < X_{\text{Strength}}\right) = \sum_{x=0}^{\infty} P_{X_{\text{Stress}}}(x)(1 - F_{X_{\text{Strength}}}(x)),$$

where $P_X(x)$ and $F_X(x)$ are the pmf and cdf of a random variable X, respectively. The expression for stress-strength reliability when $X_{Stress} \sim$ P2S-L with parameter θ_1 and $X_{Strength} \sim$ P2S-L with parameter θ_2, and X_{Stress} and $X_{Strength}$ are independent, is obtained in explicit form as follows:

$$\begin{aligned}
&R_{Stress-Strength} = \\
&\frac{\theta_2^4}{(1+\theta_1)^2(1+\theta_2)^2(\theta_2+\theta_1(1+\theta_2))^7} \times \theta_1^3\left\{35+\theta_1(5+2\theta_1)\left\{30+\theta_1\left[39+\theta_1\left[30+\theta_1(11+2\theta_1)\right]\right]\right\}\right\} \\
&\quad + \theta_1^2\left\{21+\theta_1\left\{217+\theta_1\left[711+\theta_1\left[1111+\theta_1\left(953+\theta_1\left(471+16\theta_1(8+\theta_1)\right)\right)\right]\right]\right\}\right\}\theta_2 \\
&\quad + \theta_1(1+\theta_1)(7+\theta_1)\left\{91+\theta_1\left\{434+\theta_1\left[918+\theta_1\left[987+\theta_1\left(573+\theta_1(177+25\theta_1)\right)\right]\right]\right\}\right\}\theta_2^2 \\
&\quad + (1+\theta_1)^2\left\{1+\theta_1\left\{21+\theta_1\left[126+\theta_1\left[350+\theta_1\left(477+\theta_1\left(331+\theta_1(119+19\theta_1)\right)\right)\right]\right]\right\}\right\}\theta_2^3 \\
&\quad + (1+\theta_1)^3\left\{2+\theta_1\left\{19+\theta_1\left[65+\theta_1\left(110+\theta_1\left(92+\theta_1(39+7\theta_1)\right)\right)\right]\right\}\right\}\theta_2^4 + (1+\theta_1)^9\theta_2^5.
\end{aligned}$$

Some numerical values for $R_{Stress-Strength}$ for different values of the parameters are given in Table 2.

Table 2: Numerical values for $R_{Stress-Strength}$ for different values of θ_1 and θ_2.

$\theta_1 \rightarrow$ $\theta_2 \downarrow$	0.1	0.2	0.5	0.9	1	1.5	2	2.3
0.1	0.507498	0.196622	0.031747	0.008971	0.007219	0.003305	0.002029	0.001639
0.2	0.817939	0.514575	0.154903	0.058235	0.048808	0.025438	0.016749	0.013911
0.5	0.974228	0.871305	0.534687	0.310976	0.27908	0.183003	0.13734	0.120468
0.9	0.993732	0.958349	0.765952	0.560421	0.52376	0.397247	0.326012	0.297247
1	0.995124	0.966264	0.796984	0.602639	0.56673	0.439911	0.366428	0.336299
1.5	0.998087	0.984944	0.886158	0.742172	0.712211	0.597391	0.523796	0.49193
2	0.998971	0.991334	0.925507	0.815831	0.79141	0.693082	0.626112	0.596133
2.3	0.999227	0.993309	0.939357	0.844368	0.822612	0.733147	0.670611	0.642212

From Table 2, it is clear that the stress-strength reliability increases as $\theta_2 \rightarrow \infty$ and it decreases as $\theta_1 \rightarrow \infty$.

2.4. Generating Random Values from the P2S-L Distribution

Random values from the P2S-L distribution can be generated using the steps given below. For $i = 1, 2, 3, \ldots, n$, consider

1. $\lambda_i \sim \text{Lindley}(\theta)$ and $\lambda_i^* \sim \text{Lindley}(\theta)$, independently,
2. an observation x_i of $X_i \sim \text{P}(\lambda_i + \lambda_i^*)$.

The obtained generated values are $x_1, x_2, \ldots, x_n$.

3. ESTIMATION METHODS

To accurately predict the behavior of a given distribution, the unknown parameter(s) must be precisely estimated. Here, four classical estimation methods, such as the method of moments (MM), maximum likelihood (ML), least square (LS) and weighted least square (WLS) estimation methods, are used for this purpose. We assume that n observations of a random variable X with the P2S-L distribution having the unknown parameter θ are $x_1, x_2, \ldots, x_n$.

3.1. Method of Moments

In the first method of estimation, the parameter of the P2S-L distribution is estimated with the MM. The theoretical and empirical moments are equated and solved to obtain the same. Based on the first sample moment given as $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, the MM estimate

(MME) of θ, say $\hat{\theta}_{MME}$, satisfies $\mathrm{E}(X) = \overline{x}_n$, which implies that

$$\hat{\theta}_{MME} = \frac{2 - \overline{x}_n + \sqrt{4 + 12\overline{x}_n + \overline{x}_n^2}}{2\overline{x}_n},$$

provided that $\overline{x}_n > 0$. A biased property of the MME is provided in the next proposition.

Proposition 3. The estimate $\hat{\theta}_{MME}$ is positively biased.

Proof: We have $\hat{\theta}_{MME} = g(\overline{x}_n)$, where $g(t) = \frac{2-t+\sqrt{4+12t+t^2}}{2t}$, with $t > 0$. Now, since

$$g''(t) = \frac{8\left[2 + \sqrt{4 + t(12+t)}\right] + 2t\left\{12\left[3 + \sqrt{4 + t(12+t)}\right] + t\left[30 + 3t + \sqrt{4 + t(12+t)}\right]\right\}}{t^3\left[4 + t(12+t)\right]^{3/2}} > 0,$$

$g(t)$ is strictly convex. The Jensen inequality states that $\mathrm{E}\left(g\left(\bar{X}_n\right)\right) > g\left(\mathrm{E}\left(\bar{X}_n\right)\right)$ when considering the random version of $\overline{x}_n$, denoted by $\bar{X}_n$. This implies that the random version of $\hat{\theta}_{MME}$, say $\tilde{\theta}_{MME}$, satisfies the inequality $\mathrm{E}(\tilde{\theta}_{MME}) > \theta$. Therefore, $\hat{\theta}_{MME}$ is positively biased. This ends the proof. □

3.2. Maximum Likelihood Estimation

The second method of estimation consists of finding the ML estimate (MLE) of θ. To this end, the likelihood function is

$$\mathrm{L} = \prod_{i=1}^{n} \frac{\theta^4(1+x_i)}{6(1+\theta)^{6+x_i}}\left[x_i^2 + 6(\theta+2)^2 + x_i(11+6\theta)\right]$$

and the log of the likelihood function is

$$\log \mathrm{L} = 4n\log\theta - 6n\log(1+\theta) - n\log(6) + \sum_{i=1}^{n}\log(1+x_i) + \sum_{i=1}^{n}\log\left[x_i^2 + 6(\theta+2)^2 + x_i(11+6\theta)\right] - \log(\theta+1)\sum_{i=1}^{n}x_i.$$

Now, the MLE of θ, say $\hat{\theta}_{MLE}$, satisfies

$$\hat{\theta}_{MLE} = \mathrm{argmax}_{\theta>0}\,\mathrm{L} \quad \text{or} \quad \hat{\theta}_{MLE} = \mathrm{argmax}_{\theta>0}\log\mathrm{L},$$

where $\mathrm{argmax}_{\{c\}}\,G(c)$ are the points c of some function G at which the function values are maximized. It can be obtained by solving the following non-linear equation:

$$\frac{\partial \log \mathrm{L}}{\partial\theta} = 0 \implies \frac{4n}{\theta} - \frac{6n}{\theta+1} - \frac{1}{1+\theta}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\frac{6(4+x_i+2\theta)}{x_i^2 + 6(2+\theta)^2 + x_i(11+6\theta)} = 0.$$

(3.1)

We obtain $\hat{\theta}_{MLE}$ by numerically solving (3.1) using a mathematical software like MATHEMATICA, MATHCAD and R. In our study, we employ the optim function of the R software to obtain $\hat{\theta}_{MLE}$. The variance and standard error (SE) of $\hat{\theta}_{MLE}$ are calculated using the inverse of the observed scalar information matrix. The fdHess function of the R software is used for the same purpose. For $\alpha \in (0,1)$, the asymptotic $100(1-\alpha)\%$ confidence interval (CI) for the parameter θ is

$$\left[\hat{\theta}_{MLE} \pm z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_{MLE})}\right],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

3.3. Least Square and Weighted Least Square Estimation

Let the ordered values of $x_1, x_2, \ldots, x_n$ be $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and set

$$\mathrm{LS} = \sum_{i=1}^{n}\left[F\left(x_{(i)};\theta\right) - \frac{i}{n+1}\right]^2$$

and

$$\mathrm{WLS} = \sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}\left[F\left(x_{(i)};\theta\right) - \frac{i}{n+1}\right]^2.$$

We obtain the LS estimate (LSE) and WLS estimate (WLSE) of θ by

$$\hat{\theta}_{LSE} = \mathrm{argmin}_{\theta>0}\,\mathrm{LS} \quad \text{and} \quad \hat{\theta}_{WLSE} = \mathrm{argmin}_{\theta>0}\,\mathrm{WLS},$$

respectively, where $\mathrm{argmin}_{\{c\}}\, G(c)$ are the points c of some function G at which the function values are minimized. The LS and WLS estimates are optimized using the optim function of the R software.

3.4. Simulation Study

An extensive simulation study is used in this section to compare the above-mentioned methods for estimating the unknown parameter θ. Some arbitrarily chosen values for θ, i.e., $\theta = 0.2, 0.5, 0.8, 1.3, 1.7, 2.2$, are considered for different sample sizes (n = 25, 50, 100, 200, 400 and 800). Also, $N = 1000$ replications are used for the same. In addition, measures such as bias and mean square error (MSE) are calculated with the following formulas:

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}|\hat{\theta}_i - \theta|, \quad \mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2,$$

where the index i refers to the i^{th} experiment. Simulation results, including the bias and MSEs of the MME, MLE, LSE, and WLSE, are reported in Table 3.

Table 3: Simulation results for the P2S-L distribution.

$\theta = 0.2$								
n	MME		MLE		LSE		WLSE	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
25	0.002781	0.000533	0.002770	0.000533	0.005383	0.000513	0.005375	0.000487
50	0.001035	0.000257	0.001018	0.000257	0.005271	0.000272	0.005362	0.000257
100	0.000993	0.000126	0.000991	0.000126	0.005270	0.000146	0.005299	0.000145
200	0.000832	0.000061	0.000833	0.000061	0.005268	0.000085	0.005206	0.000081
400	0.000315	0.000028	0.000314	0.000028	0.005161	0.000056	0.005172	0.000055
800	0.000109	0.000015	0.000111	0.000015	0.004723	0.000042	0.004943	0.000042
$\theta = 0.5$								
n	MME		MLE		LSE		WLSE	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
25	0.008855	0.003876	0.008701	0.003872	0.034054	0.004318	0.036604	0.004153
50	0.003064	0.001945	0.002984	0.001943	0.034036	0.002693	0.036014	0.002635
100	0.002575	0.001028	0.002517	0.001028	0.033953	0.001911	0.035450	0.001949
200	0.001375	0.000482	0.001366	0.000481	0.033330	0.001505	0.035239	0.001593
400	0.000750	0.000225	0.000731	0.000226	0.033178	0.001346	0.034293	0.001471
800	0.000674	0.000123	0.000667	0.000123	0.031546	0.001264	0.032465	0.001431
$\theta = 0.8$								
n	MME		MLE		LSE		WLSE	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
25	0.012386	0.012897	0.012017	0.012838	0.089365	0.014970	0.110187	0.015583
50	0.008235	0.006071	0.007991	0.006056	0.089177	0.011887	0.105940	0.012316
100	0.002922	0.003083	0.002862	0.003090	0.089012	0.009989	0.102034	0.012215
200	0.002850	0.001628	0.002788	0.001627	0.087833	0.008724	0.096044	0.011598
400	0.001502	0.000753	0.001477	0.000754	0.086976	0.008421	0.088295	0.011250
800	0.000509	0.000358	0.000511	0.000358	0.078654	0.008254	0.088295	0.010699
$\theta = 1.3$								
n	MME		MLE		LSE		WLSE	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
25	0.029639	0.048833	0.034871	0.052769	0.245340	0.072775	0.338869	0.115103
50	0.013136	0.022640	0.016509	0.022893	0.243812	0.067841	0.324150	0.105609
100	0.006243	0.011242	0.009289	0.010663	0.243389	0.064562	0.310919	0.097875
200	0.005737	0.005420	0.006728	0.005411	0.241816	0.060674	0.297188	0.091021
400	0.005690	0.002699	0.002866	0.002688	0.240467	0.060552	0.281971	0.085582
800	0.000949	0.001443	0.001221	0.001348	0.230707	0.059784	0.2664455	0.084874
$\theta = 1.7$								
n	MME		MLE		LSE		WLSE	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
25	0.067738	0.109771	0.053360	0.103309	0.417770	0.188135	0.590104	0.348501
50	0.024998	0.044997	0.034529	0.047393	0.417183	0.182087	0.569252	0.324724
100	0.007982	0.023980	0.017156	0.022596	0.416915	0.181277	0.546202	0.299772
200	0.006952	0.011087	0.007388	0.010343	0.414872	0.175299	0.524531	0.278230
400	0.002851	0.004889	0.001480	0.005598	0.411906	0.175281	0.498656	0.256103
800	0.000955	0.002690	0.000496	0.002784	0.399852	0.174821	0.465326	0.233700
$\theta = 2.2$								
n	MME		MLE		LS		WLS	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
25	0.116734	0.290218	0.115604	0.290072	0.688772	0.483882	0.954364	0.911138
50	0.037157	0.095885	0.036454	0.095643	0.686642	0.482239	0.925549	0.857396
100	0.012432	0.045527	0.012031	0.045469	0.686362	0.480097	0.894067	0.801097
200	0.009737	0.024009	0.009611	0.024004	0.686263	0.475233	0.863627	0.749363
400	0.005487	0.009871	0.005440	0.009866	0.684225	0.472904	0.826796	0.692542
800	0.004243	0.005719	0.004180	0.005725	0.666975	0.472544	0.780233	0.629468

According to Table 3, as n increases, the bias and MSE tend to 0. As a result, it is clear that MME and MLE perform almost identically when estimating the unknown parameter θ, and thus both methods are appropriate for small and large sample sizes.

4. P2S-L REGRESSION MODEL

In Section 2.1, it is clearly mentioned that the P2S-L model is capable of modelling over-dispersed data sets. Over-dispersed data sets are of paramount importance since they are most often observed in real life. Hence, we specialize in over-dispersion. This section introduces a count regression model based on the P2S-L distribution for modeling over-dispersed data sets.

4.1. Model Construction

Let Y be the random variable which represents the counted number of occurrences of an event. We assume that it has the P2S-L distribution and satisfies $\mathrm{E}(Y) = \mu$, implying the following reparametrization:

$$\theta = \frac{2 - \mu + \sqrt{4 + 12\mu + \mu^2}}{2\mu}.$$

Assume we have n observations of Y, which is also the response variable, with the i^{th} observation being a realization of a random variable Y_i for $i = 1, 2, \ldots, n$, and the mean of Y_i is linked to the random covariate $X^T = (X_1, X_2, \ldots, X_k)$ using the log-link function given by

$$\mu_i = \mathrm{E}(Y_i) = e^{x_i^T \gamma}, \tag{4.1}$$

where $x_i^T = (x_{i1}, x_{i2}, \ldots, x_{ik})$ is the covariate vector and $\gamma = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_k)$ is the unknown vector of regression coefficients. Based on (4.1), a linear form for the pmf of $Y_i|X_i^T = x_i^T$ which follows the P2S-L distribution with parameter μ_i is obtained as

$$\begin{aligned} P(y_i; e^{x_i^T\gamma}) = {} & \frac{\left[\frac{2 - e^{x_i^T\gamma} + \sqrt{4 + 12e^{x_i^T\gamma} + (e^{x_i^T\gamma})^2}}{2e^{x_i^T\gamma}}\right]^4 (1 + y_i)}{6\left\{1 + \left[\frac{2 - e^{x_i^T\gamma} + \sqrt{4 + 12e^{x_i^T\gamma} + (e^{x_i^T\gamma})^2}}{2e^{x_i^T\gamma}}\right]\right\}^{6+y_i}} \times \\ & \left\{ y_i^2 + 6\left[\left(\frac{2 - e^{x_i^T\gamma} + \sqrt{4 + 12e^{x_i^T\gamma} + (e^{x_i^T\gamma})^2}}{2e^{x_i^T\gamma}}\right) + 2\right]^2 \right. \\ & \left. + \, y_i \left[11 + 6\left(\frac{2 - e^{x_i^T\gamma} + \sqrt{4 + 12e^{x_i^T\gamma} + (e^{x_i^T\gamma})^2}}{2e^{x_i^T\gamma}}\right)\right]\right\}, \end{aligned}$$

where $y_i = 0, 1, 2, \ldots$

4.2. Estimation of the Model

The ML method is used to estimate the regression coefficients in γ. The log of the likelihood function, say U, of the P2S-L count regression model is given by

$$\begin{aligned} \log \mathrm{U} = {} & 4\sum_{i=1}^{n}\left[\log\left(2 - e^{x_i^T\gamma} + \sqrt{4 + 12e^{x_i^T\gamma} + (e^{x_i^T\gamma})^2}\right) - \log(2e^{x_i^T\gamma})\right] + \sum_{i=1}^{n}\log(1 + y_i) \\ & - n\log 6 - \sum_{i=1}^{n}(6 + y_i)\log\left[1 + \left(\frac{2 - e^{x_i^T\gamma} + \sqrt{4 + 12e^{x_i^T\gamma} + (e^{x_i^T\gamma})^2}}{2e^{x_i^T\gamma}}\right)\right] \\ & + \sum_{i=1}^{n}\log\left\{ y_i^2 + 6\left[\left(\frac{2 - e^{x_i^T\gamma} + \sqrt{4 + 12e^{x_i^T\gamma} + (e^{x_i^T\gamma})^2}}{2e^{x_i^T\gamma}}\right) + 2\right]^2 \right. \\ & \left. + \, y_i\left[11 + 6\left(\frac{2 - e^{x_i^T\gamma} + \sqrt{4 + 12e^{x_i^T\gamma} + (e^{x_i^T\gamma})^2}}{2e^{x_i^T\gamma}}\right)\right]\right\}. \end{aligned} \tag{4.2}$$

Now an estimation of the unknown vector γ is obtained by maximizing (4.2):

$$\hat{\gamma} = \text{argmax}_{\gamma} \log \text{U}.$$

In this regard, we use the optim function in the R software. Also, the SE of this estimate is calculated using the fdHess function in the R software.

4.3. Simulation of the P2S-L Regression Model

In this part, the ML method employed to estimate the unknown regression parameters is analyzed via a simulation study. By taking $k = 2$, the parametric combinations ($\gamma_0 = 0.5,\ \gamma_1 = 0.1,\ \gamma_2 = 0.4$) and ($\gamma_0 = 0.2,\ \gamma_1 = 0.7,\ \gamma_2 = 0.5$) are used to generate $N = 1000$ samples of sizes $n = 50$, 250, 500 and 1000 from the model, $\log(\mu_i) = \gamma_0+\gamma_1 x_{i1}+\gamma_2 x_{i2}$. We assume that x_{i1} and x_{i2} are generated from the $\text{U}(0,1)$ distribution. Here, measures such as estimates, bias, and MSEs are used to prove the asymptotic property of the MLEs. Table 4 reports the simulation results.

Table 4: Simulation results for the P2S-L regression model.

n	Parameters	$\gamma_0 = 0.5,\ \gamma_1 = 0.1,\ \gamma_2 = 0.4$			$\gamma_0 = 0.2,\ \gamma_1 = 0.7,\ \gamma_2 = 0.5$		
		Estimates	Bias	MSE	Estimates	Bias	MSE
50	γ_0	0.459330	0.040670	0.069193	0.158021	0.041979	0.078119
	γ_1	0.115366	0.015366	0.077800	0.716583	0.016583	0.073192
	γ_2	0.403656	0.003656	0.079126	0.502873	0.002873	0.079101
250	γ_0	0.492112	0.007888	0.011192	0.190608	0.009392	0.013913
	γ_1	0.097662	0.002338	0.014621	0.702769	0.002769	0.014656
	γ_2	0.402847	0.002847	0.014029	0.500762	0.000762	0.013577
500	γ_0	0.494108	0.005892	0.006232	0.193702	0.006298	0.006675
	γ_1	0.098186	0.001814	0.007140	0.702319	0.002319	0.007309
	γ_2	0.401727	0.001727	0.007149	0.500547	0.000547	0.006969
1000	γ_0	0.496372	0.003628	0.002971	0.197913	0.002087	0.003236
	γ_1	0.099694	0.000307	0.003512	0.702143	0.002143	0.003694
	γ_2	0.400233	0.000233	0.003518	0.500348	0.000348	0.003456

From Table 4, it is clear that as sample size increases, the bias and MSEs are nearer to 0, implying the consistency property of the MLEs for estimating the regression parameters.

5. INAR(1) PROCESS WITH P2S-L INNOVATIONS

In this section, we use the P2S-L distribution as an innovation distribution for the INAR(1) process for modelling over-dispersed count data sets. We have seen in the earlier sections that the P2S-L distribution is appropriate for over-dispersed count data sets. Also, in the empirical results section discussed later, we clearly prove that the P2S-L model provides better fits and results than some other recently developed overdispersed models. So the P2S-L distribution is applied here as an innovation distribution, and then the statistical properties of the process are investigated. A simulation study is included,

and the parameters are estimated using the conditional maximum likelihood.
The process $\{X_t\}_{t\in\mathbb{Z}}$ defined by

$$X_t = p \circ X_{t-1} + \varepsilon_t, \tag{5.1}$$

with $p \in [0,1)$, is called the INAR(1) process if the innovations $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ are independent and identically distributed (iid). The symbol $\circ$ in (5.1) denotes the binomial thinning operator, which is described as

$$p \circ X_{t-1} = \sum_{j=1}^{X_{t-1}} C_j,$$

where $\{C_j\}_{j\in\mathbb{Z}}$ is a sequence of iid Bernoulli random variables with parameter p. Therefore, the corresponding INAR(1) process has the following one step transition probability:

$$\Pr\left(X_t = k \mid X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \Pr\left(B(l,p) = i\right) \Pr\left(\varepsilon_t = k - i\right), \quad k, l = 0, 1, \ldots,$$

where $B(l,p) \sim \text{Binomial}(l,p)$ and $p \in [0,1)$. Historically, the P distribution was the first to be used with the INAR(1) process (for more information, see [25] and [2]).
If $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ has a P2S-L distribution with parameter θ, then the corresponding INAR(1)P2S-L process has the one step transition probability given as

$$\begin{aligned}\Pr\left(X_t = k \mid X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} & \binom{l}{i} p^i (1-p)^{l-i} \times \\ & \frac{\theta^4\left[1 + (k-i)\right]}{6(1+\theta)^{6+(k-i)}} \left[(k-i)^2 + 6(\theta+2)^2 + (k-i)(11+6\theta)\right].\end{aligned}$$

By using the expressions for the mean, variance, DI, conditional mean and conditional variance of $\{X_t\}_{t\in\mathbb{Z}}$ (see [30] and [2]), the mean, variance, DI, conditional mean and conditional variance of the INAR(1)P2S-L process are derived as

$$\begin{aligned}\mu_* = \mathrm{E}\left(X_t\right) &= \frac{2(2+\theta)}{(1-p)\theta(1+\theta)},\\ \sigma_*^2 = \mathrm{Var}\left(X_t\right) &= \frac{2\left(\theta^3(p+1) + \theta^2(3p+4) + 2\theta(p+3) + 2\right)}{\theta^2(\theta+1)^2\left(1-p^2\right)},\\ \mathrm{DI}_* = \mathrm{DI}(X_t) &= \frac{\theta^3(p+1) + \theta^2(3p+4) + 2\theta(p+3) + 2}{\theta(\theta+1)(\theta+2)(p+1)},\end{aligned}$$

$$\mathrm{E}\left(X_t \mid X_{t-1}\right) = pX_{t-1} + \frac{2(2+\theta)}{\theta(1+\theta)} \tag{5.2}$$

and

$$\mathrm{Var}\left(X_t \mid X_{t-1}\right) = p(1-p)X_{t-1} + \frac{2\left(\theta^3 + 4\theta^2 + 6\theta + 2\right)}{\theta^2(\theta+1)^2}. \tag{5.3}$$

5.1. Estimation of the Process

Here, the estimation procedure using the conditional maximum likelihood (CML) method for the INAR(1)P2S-L process is examined. Let us consider T observations $x_1, x_2, \ldots, x_T$ of the INAR(1)P2S-L process from the random sample $X_1, X_2, \ldots, X_T$. Then the conditional log likelihood function is

$$\begin{aligned}\ell(\Theta) &= \sum_{t=2}^{T} \log\left[\Pr\left(X_t = x_t \mid X_{t-1} = x_{t-1}\right)\right] \\ &= \sum_{t=2}^{T} \log\left\{\sum_{i=0}^{\min(x_t, x_{t-1})} \binom{x_{t-1}}{i} p^i (1-p)^{x_{t-1}-i} \times \right. \\ &\qquad \left. \frac{\theta^4\left[1+(x_t-i)\right]}{6(1+\theta)^{6+(x_t-i)}}\left[(x_t-i)^2 + 6(\theta+2)^2 + (x_t-i)(11+6\theta)\right]\right\},\end{aligned}$$

where $\Theta = (\theta, p)^T$ is the unknown parametric vector to be estimated. The CML estimates (CMLEs) are obtained by

$$\hat{\Theta} = \text{argmax}_{\Theta}\, \ell(\Theta).$$

The optim and fdHess functions of the R program are used to obtain the CMLEs, observed information matrix, and therefore the SEs of the parameter estimates.

5.2. Simulation of the INAR(1)P2S-L Process

A simulation study is used to thoroughly investigate the CML method for estimating the parameters of the INAR(1)P2S-L process. Hence, we generated $N = 1000$ samples, each of sizes $n = 20, 50, 100, 500, 1000$, from the INAR(1)P2S-L process for the following four sets of parameter values: $(\theta = 0.2, p = 0.3)$, $(\theta = 0.25, p = 0.5)$, $(\theta = 1.2, p = 0.7)$ and $(\theta = 1.6, p = 0.9)$. For each n, the average bias and MSE for the parameters were calculated, and the simulation results are presented in Table 5.

Table 5: Simulation results for the INAR(1)P2S-L process.

n	Parameters	$\theta = 0.2,\ p = 0.3$			$\theta = 0.25,\ p = 0.5$		
		Estimates	Bias	MSE	Estimates	Bias	MSE
20	θ	0.219149	0.019149	0.002210	0.265957	0.015957	0.003318
	p	0.337932	0.037932	0.008395	0.510654	0.010654	0.006919
50	θ	0.210206	0.010206	0.000677	0.258585	0.008585	0.001116
	p	0.323072	0.023072	0.003635	0.507034	0.007034	0.002768
100	θ	0.205219	0.005219	0.000241	0.251768	0.001768	0.000489
	p	0.308560	0.008560	0.001226	0.498929	-0.001071	0.001371
500	θ	0.203217	0.003217	0.000062	0.250827	0.000827	0.000094
	p	0.311741	0.011741	0.000414	0.500856	0.000856	0.000272
1000	θ	0.204506	0.004506	0.000043	0.250165	0.000165	0.000045
	p	0.312342	0.012342	0.000288	0.500012	0.000012	0.000136
n	Parameters	$\theta = 1.2,\ p = 0.7$			$\theta = 1.6,\ p = 0.9$		
		Estimates	Bias	MSE	Estimates	Bias	MSE
20	θ	1.326347	0.126347	0.134854	1.751818	0.151818	0.402999
	p	0.724984	0.024984	0.006737	0.897135	-0.002865	0.001089
50	θ	1.289355	0.089355	0.055774	1.647628	0.047628	0.097429
	p	0.716579	0.016579	0.002509	0.897772	-0.002228	0.000349
100	θ	1.282011	0.082011	0.014919	1.625991	0.025991	0.048762
	p	0.711249	0.011249	0.000982	0.899145	-0.000855	0.000189
500	θ	1.243593	0.043593	0.008646	1.600812	0.000812	0.008673
	p	0.705575	0.005575	0.000652	0.899731	-0.000269	0.000034
1000	θ	1.220952	0.020952	0.008867	1.602889	0.002889	0.004258
	p	0.702504	0.002504	0.000261	0.900023	0.000023	0.000018

Table 5 makes it clear that the CML method is suitable for the estimation of the parameters since the bias and MSEs decrease to 0 quickly for small as well as large sample sizes.

6. EMPIRICAL STUDY

In this section, the proposed models under the P2S-L distribution are compared with some discrete models using four real data sets to prove their efficiency. The first two data sets illustrate the performance of the P2S-L distribution. The new P weighted exponential (NPWE) distribution (see [3]), P xgamma (PX) distribution (see [8]), P Bilal (PBl) distribution (see [4]), P Lindley (PL) distribution (see [28]) and P distribution are considered for this purpose. The estimates of the involved parameters are computed along with their SEs, CIs (in the form of lower and upper bounds (LCI and UCI, respectively)), goodness of fit statistic (χ^2 statistic), degrees of freedom, p-values, and model adequacy measures such as the minus maximized log likelihood function (-Log L) Akaike information criterion (AIC) and Bayesian information criterion (BIC). The third data set is used to check the efficiency of the P2S-L regression model by comparing it with the new P generalized Lindley (NPGL) regression model (see [5]), the PX regression model, the PL regression model, and the P regression model by using the measures -Log L, AIC, and BIC. The last data set is used to check the capability of an integer-valued auto-regressive process with P2S-L innovations, the INAR(1)P2S-L process. The INAR(1)P2S-L process is compared with the INAR(1)PTE process (see [6]), INAR(1)PX process (see [8]), INAR(1)PL process (see [22]), INAR(1)P process, and the INAR(1)G process by means of -Log L, AIC, BIC, and fitted measures such as mean, variance, and DI.

6.1. Bacterial Clumps

This data set contains measures of the distribution of bacterial clumps per field in a milk film. A microscope slide was split into 400 regions of equal area, and the number of bacterial clumps on each was counted. The authors in [11] considered this data set and fitted it by the NB distribution rather than the P distribution. Table 6 contains some descriptive measures of the fitted P2S-L distribution using this data set. Hence, it is evident that the fitted P2S-L distribution is over-dispersed, right-skewed, and leptokurtic.

Table 6: Values of some descriptive statistics of the P2S-L distribution for the bacterial clumps data set.

Mean	Variance	DI	Skewness	Kurtosis
2.4625	5.7029	2.3159	2.1017	10.7154

The MLEs with their corresponding SEs, CIs under the form (LCI,UCI) for the parameter(s) and goodness-of-fit statistic for the bacterial clumps data set are given in Table 7.

Table 7: The MLE, LCI, UCI, -Log L, AIC, BIC, χ^2 and p-values for the distributions considered using the bacterial clumps data set.

X	Observed frequency	Expected frequency					
		P2S-L	NPWE	PX	PBl	PL	P
0	56	74.13	116.28	105.96	82.04	100.07	34.86
1	104	88.85	82.48	79.90	89.32	83.40	85.07
2	80	76.84	58.50	61.50	73.14	64.29	103.78
3	62	57.42	41.49	46.40	53.38	47.28	84.41
4	42	39.36	29.43	33.95	36.63	33.68	51.49
5	27	25.46	20.88	24.09	24.20	23.44	25.13
6	9	15.79	14.81	16.63	15.58	16.04	10.22
7	9	9.48	10.50	11.21	9.85	10.83	3.56
8	5	5.55	7.45	7.40	6.15	7.24	1.09
9	3	3.18	5.28	4.81	3.80	4.79	0.29
10	2	1.79	3.75	3.08	2.33	3.15	0.07
11	0	0.99	2.66	1.94	1.42	2.06	0.02
12	1	1.17	6.49	3.13	2.17	3.74	0.00
Total	400	400	400	400	400	400	400
θ	MLE	1.1915	0.3724	0.8540	0.3411	0.6523	2.4400
	SE	0.0023	0.2782	0.0018	0.0008	0.0015	0.0039
	LCI	1.1870	0	0.8505	0.3395	0.6493	2.4323
	UCI	1.1960	0.9177	0.8576	0.3427	0.6552	2.4477
α	MLE	-	0.1004	-	-	-	-
	SE	-	0.8219	-	-	-	-
	LCI	-	0	-	-	-	-
	UCI	-	1.7114	-	-	-	-
χ^2		10.9593	73.5180	52.3447	18.3150	43.9591	42.7049
degrees of freedom		8	7	8	8	8	8
p-value		0.2785	<0.001	<0.001	0.0317	<0.001	<0.001
-Log L		-795.5053	-829.4187	-817.7587	-798.9790	-813.0958	-831.6082
AIC		1593.0106	1660.8373	1637.5174	1599.9580	1628.1917	1665.2165
BIC		1597.0021	1664.8288	1641.5089	1603.9495	1632.1831	1669.2079

According to Table 7, the P2S-L distribution performs well because it has the lowest AIC, BIC, and χ^2 values with a higher p-value. Also, the estimated pmfs of the fitted distributions are diagrammatically represented in Figure 2.

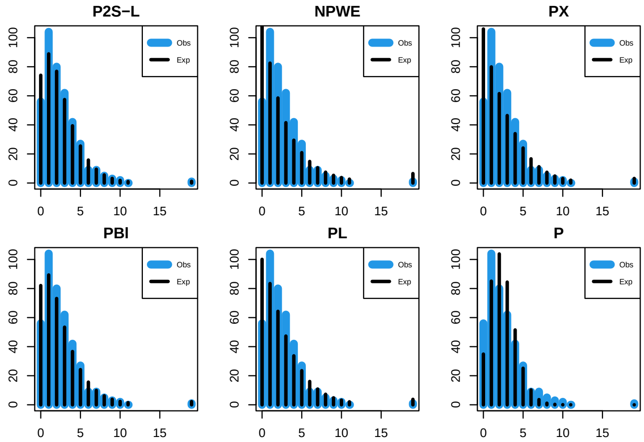

Figure 2: The estimated pmfs of the fitted distributions for the bacterial clumps data set.

From Figure 2 also, we can infer that the P2S-L distribution yields the best fit among other fitted distributions.

6.2. Yeast Cells

The second data set contains measurements of yeast cell distribution per square in a hemocytometer (see [11]). The authors of this reference also used this data set to claim the fit of the NB distribution because it lacks randomness. Table 8 contains some descriptive measures of the fitted P2S-L distribution for this data set. Hence, here also, it is evident that the fitted P2S-L distribution is over-dispersed, right-skewed, and leptokurtic.

Table 8: Values of some descriptive statistics of the P2S-L distribution for the yeast cells data set.

Mean	Variance	DI	Skewness	Kurtosis
0.69	0.91118	1.3205	1.7254	6.7372

Table 9 shows the MLEs with their corresponding SEs, CIs in the form (LCI, UCI) for the parameter(s), and goodness-of-fit statistic for the yeast cells data set.

Table 9: The MLE, LCI, UCI, -Log L, AIC, BIC, χ^2 and p-values for distributions considered using the yeast cells data set.

X	Observed frequency	P2S-L	NWPE	PX	PBl	PL	P
1	213	221.22	237.74	237.14	199.88	234.04	202.14
2	128	114.24	96.44	95.85	116.68	99.41	137.96
3	37	43.56	39.12	39.87	51.43	40.50	47.08
4	18	14.58	15.87	16.41	20.29	16.04	10.71
5	3	4.53	6.44	6.59	7.55	6.22	1.83
6	1	1.86	4.39	4.14	4.17	3.79	0.28
Total	400	400	400	400	400	400	400
θ	MLE	3.5683	1.1235	2.3330	0.9990	1.9502	0.6825
	SE	0.0109	2.2302	0.0072	0.0031	0.0064	0.0021
	LCI	3.5470	0	2.3189	0.9930	1.9377	0.6785
	UCI	3.5897	5.4947	2.3470	1.0050	1.9627	0.6865
α	MLE	-	0.3041	-	-	-	-
	SE	-	2.5888	-	-	-	-
	LCI	-	0	-	-	-	-
	UCI	-	5.3782	-	-	-	-
χ^2		3.0015	13.8457	14.4230	9.1358	11.0452	10.0362
degrees of freedom		2	1	2	2	2	2
p-value		0.3914	0.0031	0.0024	0.0275	0.0115	$<$0.001
-Log L		-447.3560	-454.4330	-454.3928	-451.9304	-452.6185	-449.5038
AIC		896.7120	910.8660	910.7857	905.8609	907.2370	901.0076
BIC		900.7035	914.8574	914.7771	909.8524	911.2285	904.9990

From Table 9, it is clear that the P2S-L distribution is the best among the considered competitive models since it has the lowest -Log L, AIC, BIC and χ^2 value with the highest p-value.

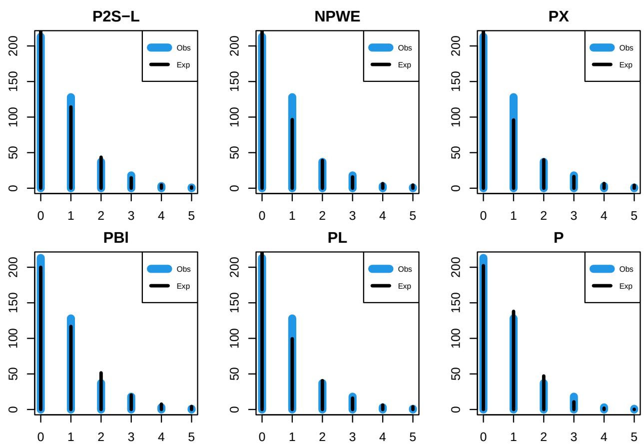

Figure 3: The estimated pmfs of the fitted distributions for the yeast cells data set.

From Figure 3, we can infer that the P2S-L distribution yields the best fit among other fitted distributions.

6.3. Length of Hospital Stay

The third data set is utilized to examine the proficiency of the count regression model proposed under the P2S-L distribution. The data set consists of 3589 observations from the files of 1991 Arizona cardiovascular patients found in the COUNT package in R software. The P2S-L regression model is used to model the length of stay (y_i) by using the covariates: cardiovascular procedure (x_{1i}) (1 = CABG, 0 = PTCA), sex (x_{2i}) (1 = male, 0 = female), type of admission (x_{3i}) (1 = urgent, 0 = elective) and age (x_{4i}) (1 = age > 75, 0 = age ≤ 75). Given below is the regression structure that will be fitted by the P2S-L, NPGL, PX, PL, and P regression models:

$$\mu_i = e^{\gamma_0+\gamma_1 x_{1i}+\gamma_2 x_{2i}+\gamma_3 x_{3i}+\gamma_4 x_{4i}}.$$

The mean and variance of the dependent variable are calculated as 8.831 and 47.973, respectively, stating the clear over-dispersion. Table 10 gives the parameter estimates and results of the information criterion.

Table 10: The MLE, -Log L, AIC and BIC of the fitted regression models for the length of stay data set.

Covariates	P		PL		PX		NPGL		P2SL	
	Estimate SE	p-value	Estimate SE	p-value	Estimate SE	p- value	Estimate SE	p-value	Estimate SE	p-value
γ_0	1.4560 0.0158	<0.001	1.4133 0.0372	<0.001	1.3996 0.0349	<0.001	1.4044 0.0353	<0.001	1.4159 0.0298	<0.001
γ_1	0.9606 0.0122	<0.001	0.9843 0.0291	<0.001	0.9725 0.0270	<0.001	0.9761 0.0274	<0.001	0.9835 0.0231	<0.001
γ_2	-0.1240 0.0118	<0.001	-0.1288 0.0304	<0.001	-0.1269 0.0280	<0.001	-0.1267 0.0285	<0.001	-0.1262 0.0240	<0.001
γ_3	0.3266 0.0121	<0.001	0.3843 0.0302	<0.001	0.3732 0.0280	<0.001	0.3759 0.0284	<0.001	0.3746 0.0240	<0.001
γ_4	0.1224 0.0124	<0.001	0.1193 0.0323	<0.001	0.1202 0.0298	<0.001	0.1198 0.0303	<0.001	0.1197 0.0255	<0.001
-Log L AIC BIC	-11189.8976 22389.7952 22420.7233		-10625.5957 21239.1913 21202.0775		-10569.8162 21127.6324 21090.5187		-10563.2551 21114.5102 21077.3964		-10164.8969 20317.7938 20280.6801	

The authors in [5] used this data set to prove the better fit of the NPGL regression model. Hence, the obtained -Log L, AIC and BIC were better than those of the NB regression model. From Table 10, it is clear that the P2S-L regression model is better than the NB regression model. Now, since the P2S-L regression model has minimized values for its -Log L, AIC, and BIC, we conclude that it will be a more appropriate model than the other models for modelling this data set. As a result, we can say that the length of hospital stay increases when people have CABG cardiovascular surgery, are admitted urgently, and are over the age of 75. Additionally, female individuals have a longer hospital stay than male individuals.

6.4. Number of Weakly Sales

The fourth data set that we consider is the series of weekly sales (in integer units) of a particular soap product in a supermarket. The information comes from the Kilts Center for Marketing, Graduate School of Business, University of Chicago, and can be found at: http://gbswww. uchicago.edu/kilts/research/db/dominicks. (The product is 'Zest White Water 15 oz.', with code 3700031165). We consider it to prove the applicability of the P2S-L distribution as an innovation distribution for the INAR(1) process. The one step transition probabilities of the competitive INAR(1) models used here for comparison are given in Appendix I.

The mean, variance, and DI of the data set are 242, 5.44, and 15.40, respectively. The test proposed in [29] for determining statistically significant over-dispersion shows a p-value less than 0.001, indicating the data possess significant over-dispersion. The autocorrelation function (ACF), partial ACF (PACF), histogram, and time series plots are displayed in Figure 4 and in the PACF plot, the only first-lag significance proves that this data set can be used for modelling the INAR(1) process.

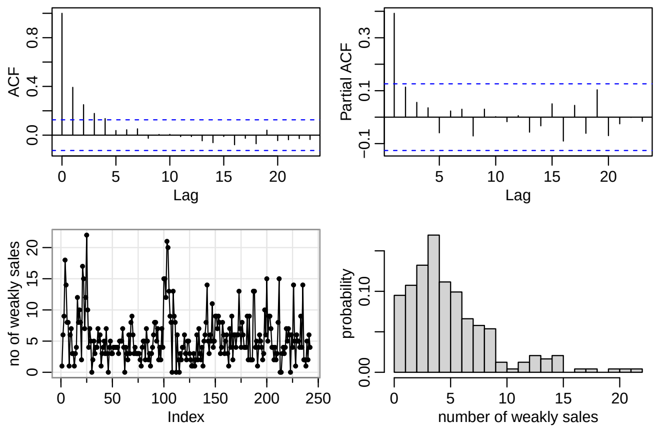

Figure 4: ACF, PACF, time series and histogram plots for the numbers of weekly sales data.

The fitted INAR(1) processes with the P2S-L innovation and other corresponding innovations used for comparison yield the parameter estimates along with SE, AIC, BIC, theoretical mean, variance, and DI as given in Table 11. The minimum values of AIC and BIC statistics for the INAR(1)P2S-L process prove that it provides a better fit than the INAR(1)PTE, INAR(1)PX, INAR(1)PL, INAR(1)P and INAR(1)G processes. Furthermore, the theoretical DI value of the INAR(1)P2S-L process is close to the empirical DI value. As a result, it is convincing that the INAR(1)P2S-L process effectively explains the data set's characteristics.

Table 11: The estimates and modelling adequacy statistics of the fitted distributions for the number of weekly sales data set.

Model	Parameters	Estimates	SE	AIC	BIC	μ_*	σ_*^2	DI_*
INAR(1)P2S-L	θ	0.7569	0.0485	1244.4780	1251.4559	5.3926	12.6378	2.3436
	p	0.2407	0.0426					
INAR(1)PTE	α	-0.9999	0.2946	1246.4137	1256.8805	1.6240	3.1983	1.9694
	θ	0.3839	0.0353					
	p	0.2860	0.0516					
INAR(1)PX	θ	0.6086	0.0397	1250.5878	1257.5656	3.1074	7.1419	2.2983
	p	0.3241	0.0369					
INAR(1)PL	θ	0.4533	0.0325	1249.0689	1256.0467	5.4545	15.2780	2.8010
	p	0.3202	0.0369					
INAR(1)P	θ	4.1853	0.2110	1363.8119	1370.7898	5.5041	5.4045	0.9819
	p	0.2341	0.0324					
INAR(1)G	θ	0.2253	0.0152	1260.3368	1267.3147	0.3347	0.3315	0.9904
	p	0.3713	0.0324					
Empirical						5.4421	15.3376	2.8183

A residual analysis is done to check whether the fitted INAR(1)P2S-L process is statistically accurate. For that, the Pearson residuals $r_1, \ldots, r_T$ are calculated, where, for any $t = 1, 2, \ldots, n$,

$$r_t = \frac{x_t - \mathrm{E}\left(X_t \mid X_{t-1} = x_{t-1}\right)}{\mathrm{Var}\left(X_t \mid X_{t-1} = x_{t-1}\right)^{1/2}},$$

where $\mathrm{E}\left(X_t \mid X_{t-1}\right)$ and $\mathrm{Var}\left(X_t \mid X_{t-1}\right)$ are given in (5.2) and (5.3), respectively. The statistical validity of the fitted INAR(1) process is proved by acquiring a zero mean and unit variance for the uncorrelated Pearson residuals (see [16]). The mean and variance of the Pearson residuals of the INAR(1)P2S-L process were 0.0419 and 1.0033, respectively, which are very close to the desired values. This proves our INAR(1)P2S-L process is statistically valid for the data set. According to the findings in [19], the INAR(1)P2S-L process for the data is such that

$$X_t = 0.2407 \circ X_{t-1} + \varepsilon_t,$$

where the innovation process satisfies $\varepsilon_t \sim$ P2S-L(0.7569).

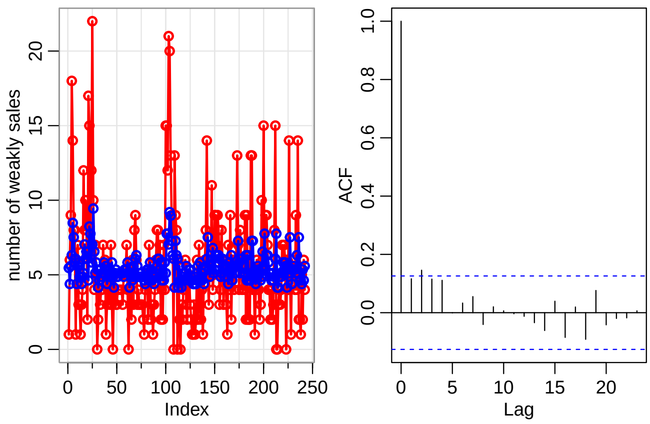

Figure 5: The predicted values of the number of weekly sales data set (left) and the ACF plot of the Pearson residuals (right).

Additionally, the ACF plot of the Pearson residuals in Figure 5 specifies that there is no presence of autocorrelation for the Pearson residuals. The plot for actual and predicted values of the monthly number of claims is displayed on the left side of Figure 5.

7. CONCLUDING REMARKS

A one-parameter discrete compound distribution, namely the Poisson 2S-Lindley (P2S-L) distribution, was proposed. It was shown that it can effectively model over-dispersed data sets. Flexibility is attained by having a closed form for almost every statistical and mathematical property. Methods such as the MM, ML, LS, and WLS methods were used for estimating the unknown parameter and were compared using a broad simulation study. More importantly, a new count regression model and an INAR(1) process based on the P2S-L distribution were introduced and described in detail. Four real-life data sets were evaluated to determine the applicability of the proposed model.

Thus, the P2S-L distribution-based models will be effective in modeling count data, particularly over-dispersed count data sets.

ACKNOWLEDGMENTS

We acknowledge the valuable suggestions from the learned referees.

REFERENCES

[1] Abd El-Monsef, M.M.E. and Sohsah, N.M. (2016). Poisson–transmuted Lindley distribution, *Journal Of Advances In Mathematics*, **11**, 9, 5631–5638.

[2] Al-Osh, M.A. and Alzaid, A.A. (1987). First-order integer-valued autoregressive (INAR(1)) process, *Journal of Time Series Analysis*, **8**, 3, 261–275.

[3] Altun, E. (2020). A new generalization of geometric distribution with properties and applications, *Communications in Statistics-Simulation and Computation*, **49**, 3, 793–807.

[4] Altun, E. (2020). A new one-parameter discrete distribution with associated regression and integer-valued autoregressive models, *Mathematica Slovaca*, **70**, 4, 979–994.

[5] Altun, E. (2021). A new two-parameter discrete poisson-generalized Lindley distribution with properties and applications to healthcare data sets, *Computational Statistics*, **36**, 2841–2861.

[6] Altun, E and Khan, N. M. (2022). Modelling with the Novel INAR(1)PTE Process, *Methodology and Computing in Applied Probability*, **24**, 1735-1751.

[7] Altun, E and Bhati, D and Khan, N. M. (2021). A new approach to model the counts of earthquakes: INARPQX(1) process, *SN Applied Sciences*, **3**, 2, 1–17.

[8] Altun, E , Cordeiro, G. M and Ristić, M. M. (2022). An one-parameter compounding discrete distribution, *Journal of Applied Statistics*, **49**, 8, 1935–1956

[9] Bereta, E. M. P., Louzanda, F. and Franco, M. A. P. (2011). The Poisson-Weibull distribution, *Advances and Applications in Statistics*, **22**, 2, 107–118.

[10] Bhati, D., Kumawat, P. and Gómez–Déniz, E. (2017). A new count model generated from mixed Poisson transmuted exponential family with an application to health care data, *Communications in Statistics-Theory and Methods*, **46**, 22, 11060–11076.

[11] Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data, *Biometrics*, **9**, 2, 176–200.

[12] Cheng, L., Geedipally, S. R. and Lord, D. (2013). The Poisson–Weibull generalized linear model for analyzing motor vehicle crash data, *Safety Science*, **54**, 38–42.

[13] Chesneau, C., Tomy, L. and Gillariose, J. (2020). On a Sum and Difference of two Lindley distributions: Theory and Applications, *REVSTAT-Statistical Journal*, **18**, 5, 673–696.

[14] Eliwa, M. S., Altun, E., El-Dawoody, M. and El-Morshedy, M. (2020). A new three-parameter discrete distribution with associated INAR(1) process and applications, *IEEE access*, **8**, 91150–91162.

[15] Grine, R. and Zeghdoudi, H. (2017). On Poisson quasi-Lindley distribution and its applications, *Journal of Modern Applied Statistical Methods*, **16**, 2, 21.

[16] Harvey, A. C. and Fernandes, C. (1989). Time series models for count or qualitative observations, *Journal of Business & Economic Statistics*, **7**, 4, 407–417.

[17] Huang, J. and Zhu, F. (2021). A New First-Order Integer-Valued Autoregressive Model with Bell Innovations, *Entropy*, **23**, 6, 713.

[18] Irshad, M. R., Chesneau, C., D'cruz, V. and Maya, R. (2021). Discrete Pseudo Lindley Distribution: Properties, Estimation and Application on INAR(1) Process, *Mathematical and Computational Applications*, **26**, 4, 76.

[19] Jazi, M. A., Jones, G. and Lai, C. D. (2012). First-order integer valued AR processes with zero inflated Poisson innovations, *Journal of Time Series Analysis*, **33**, 6, 954–963.

[20] Karlis, D. and Xekalaki, E. (2005). Mixed Poisson distributions, *International Statistical Review/Revue Internationale de Statistique*, **73**, 1, 35–58.

[21] Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem, *Journal of the Royal Statistical Society. Series B (Methodological)*, **20**, 1, 102–107.

[22] Lívio, T., Khan, N. M., Bourguignon, M.and Bakouch, H. S. (2018). An INAR(1) model with Poisson–Lindley innovations, *Economics Bulletin*, **38**, 3, 1505–1513.

[23] Lord, D. and Geedipally, S. R. (2011). The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros, *Accident Analysis & Prevention*, **43**, 5, 1738–1742.

[24] Mahmoudi, E. and Zakerzadeh, H. (2010). Generalized poisson–Lindley distribution, *Communications in Statistics—Theory and Methods*, **39**, 10, 1785–1798.

[25] McKenzie, E. (1985). Some simple models for discrete variate time series 1, *JAWRA Journal of the American Water Resources Association*, **21**, 4, 645–650.

[26] Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., Olmo-Jiménez, M. J. and Martínez-Rodríguez, A. M. (2009). A generalized Waring regression model for count data, *Computational statistics & data analysis*, **53**, 10, 3717–3725.

[27] Sáez-Castillo, A. J. and Conde-Sánchez, A. (2013). A hyper-Poisson regression model for overdispersed and underdispersed count data, *Computational Statistics & Data Analysis*, **61**, 148–157.

[28] Sankaran, M. (1970). The discrete Poisson-Lindley distribution, *Biometrics*, **26**, 1, 145–149.

[29] Schweer, S. and Weiß, C. H. (2014). Compound Poisson INAR(1) processes: stochastic properties and testing for overdispersion, *Computational Statistics & Data Analysis*, **77**, 267–284.

[30] Weiß, C. H. (2018). *An introduction to discrete-valued time series*, John Wiley & Sons.

[31] Wongrin, W. and Bodhisuwan, W. (2017). Generalized Poisson–Lindley linear model for count data, *Journal of Applied Statistics*, **44**, 15, 2659–2671.

[32] Zamani, H., Faroughi, P. and Ismail, N. (2014). Bivariate Poisson-weighted exponential distribution with applications, *AIP Conference Proceedings*, **1602**, 1, 964–968.

APPENDIX

Transition probabilities for the processes used for comparison in Section 6.4.

1. Transition probability for the INAR(1)PTE process:

$$\Pr\left(X_t = k \mid X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \binom{l}{i} p^i (1-p)^{l-i} \theta \left(\frac{1-\alpha}{(1+\theta)^{1+k-i}} + \frac{2\alpha}{(1+2\theta)^{1+k-i}} \right),$$

with $|\alpha| \leq 1$ and $\theta > 0$.

2. Transition probability for the INAR(1)PX process:

$$\Pr\left(X_t = k \mid X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \binom{l}{i} p^i (1-p)^{l-i} \times \frac{\theta^2 \left[2(1+\theta)^2 + \theta(k-i+2)(k-i+1)\right]}{2(1+\theta)^{k-i+4}},$$

with $\theta > 0$.

3. Transition probability for the INAR(1)PL process:

$$\Pr\left(X_t = k \mid X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \binom{l}{i} p^i (1-p)^{l-i} \frac{\theta^2 (k-i+\theta+2)}{(\theta+1)^{k-i+3}},$$

with $\theta > 0$.

4. Transition probability for the INAR(1)P process:

$$\Pr\left(X_t = k \mid X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \binom{l}{i} p^i (1-p)^{l-i} \frac{e^{-\theta} \theta^{k-i}}{(k-i)!},$$

with $\theta > 0$.

5. Transition probability for the INAR(1)G process:

$$\Pr\left(X_t = k \mid X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \binom{l}{i} p^i (1-p)^{l-i} \theta (1-\theta)^{k-i},$$

with $\theta \in (0, 1)$.