African Journal of Mathematics and Statistics Studies

ISSN: 2689-5323

Volume 5, Issue 3, 2022 (pp. 16-32)

www.abjournals.org

MODELLING UNDERDISPERSED COUNT DATA: RELATIVE PERFORMANCE OF POISSON MODEL AND ITS ALTERNATIVES

Ndèye Khady Guissé Seck[1*], Ablaye Ngom[1], Kandioura Noba[1]

Institut Supérieur d'Agriculture et Entreprenariat, Faculté des Sciences et Techniques, Université Cheikh Anta Diop de Dakar, Sénégal.

Corresponding author: khadijaseck7@gmail.com

ABSTRACT: *Count data are common in many fields and often modelled with the Poisson model. However, the equidispersion assumption (variance = mean) related to the Poisson model is often violated in practice. While much research has focused on modelling overdispersed count data, underdispersion has received relatively little attention. Alternative models are therefore needed to handle overdispersion (variance > mean) and underdispersion (variance < mean). This study assessed the relative performance of the Poisson model and its alternatives (COM-Poisson, Generalized Poisson Regression, Double Poisson and Gamma Count) to model underdispersed count data. Using a Monte Carlo experiment, the simulation plan considered various underdispersion levels (k (variance/mean) =* 0.2*,* 0.5 *and* 0.81*),* $k = 1$ *as a control, and sample sizes (*$n = 20, 50, 100, 300$ *and* 500*). Results showed that the Poisson model is not robust to handle underdispersion but it is the best performer when* $k = 1$*. The COM-Poisson model best fitted severe underdispersed data (*$k = 0.2$*). It is also the best performer model for moderate underdispersed count data (*$k = 0.81$*). However, when* $k = 0.5$*, the Double Poisson model and Generalized Poisson model outperformed other models for relatively large sample sizes (*$n = 100, 300$ *and* 500*). Our finding suggests that none of the models suits all situations. Therefore, in practice, several of these models need to be tested to select the best one.*

KEYWORDS: Poisson model, Underdispersion models, Count data, COM-Poisson, Gamma Count, Double Poisson, Generalized Poisson Regression.

Cite this article:

Seck N.K.G., Ngom A., Noba K. (2022), Modelling Underdispersed Count Data: Relative Performance of Poisson Model and its Alternatives. African Journal of Mathematics and Statistics Studies 5(3), 16-32. DOI: 10.52589/AJMSS-1WPJQHYT

Manuscript History

Received: 23 June 2022

Accepted: 29 July 2022

Published: 23 Aug 2022

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

INTRODUCTION

Count data are commonly produced in many scientific disciplines from both experimental and observational study designs. The Poisson distribution is appropriately used in the regression analysis of count responses for which the mean and variance are almost equal (Famoye, 1993). The resulting Poisson regression model indeed strongly relies on the equidispersion assumption (variance equals mean). However, due to variability of experimental material, omitted unobserved variables, or lack of independence between individual item responses (Kokonendji *et al*., 2008), real count data often exhibit a variance greater than mean (overdispersion) or smaller than mean (underdispersion). In either of these cases, the standard Poisson model is no longer applicable.

Overdispersion is more frequently reported and handled using overdispersion models (Kokonendji, 2014). While underdispersion is considered rarer and has received much less attention, it may have been prominently hidden by the use of popular equidispersion and overdispersion models (Sellers & Morris, 2017). Many record processes can actually lead to underdispersed data. Indeed, underdispersion can result from the data generating mechanism (Sellers *et al*., 2012). For instance, the condensed Poisson model results from reporting only every second event in a Poisson process (Chatfield & Goodhardt, 1973). In addition to such inherently underdispersed populations, observed underdispersion can be caused by a small sample size (Kokonendji *et al*., 2008), the mechanism of data collection (Kokonendji, 2014) or low sample mean value (Lord & Mannering, 2010). Underdispersion can also be a sign of over-fitting, meaning that the count model contains too many explanatory variables, leading to deficient variation (Sellers *et al*., 2012; Sellers & Morris, 2017).

Underdispersion mostly impacts estimated standard errors, although it can also induce incorrect estimation of regression parameters (Kokonendji *et al*., 2008). Indeed, it is well known that ignoring underdispersion in the regression analysis of count data leads to upward-biased standard errors, thus under estimating the statistical significance of associated explanatory variables (Sellers & Premeaux, 2020; Forthmann *et al*., 2020). Therefore, alternative methods have been proposed and used to deal with underdispersed count data. Pure underdispersion models, such as the continuous parameter binomial model (King, 1989) and the condensed Poisson model (Sellers & Morris, 2017), are rarely used in practice. Indeed, applied scientists often turn to flexible count models which can account for both underdispersion and overdispersion in observed count data. Most popular examples include, among others, the Conway-Maxwell-Poisson (CMP or COM–Poisson) model (Conway & Maxwell, 1962), the Generalized Poisson (GP) model (Consul & Jain, 1973), the Double Poisson (DP) model (Efron, 1986) and the Gamma Count (GC) model (Oh *et al*., 2006).

Several studies have compared alternatives to the Poisson model for handling underdispersed count data. For instance, Wang and Famoye (1997) and Husai and Bagmar (2015) have compared the Poisson model with the GP model using data related to household fertility decisions. The estimated parameters from both Poisson and GP model are quite similar, but as expected, the standard errors for parameter estimates from the Poisson model are smaller than those from the GP model (Husain & Bagmar, 2015). Lord *et al*. (2010) have compared the CMP model with the Poisson and GC model. They found that the CMP model fit is not significantly different from both other model fits. Instead, CMP provides a practical tool for modeling count data that have various levels of dispersion. Zou *et al*. (2013) showed that the performance of DP model is comparable to that of the CMP model in terms of goodness of fit, but the CMP

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

model provides a slightly better fit in all the considered datasets.

However, studies comparing the performance of alternatives to the Poisson model for underdispersed count data were limited to one or two approaches, in addition to the basic Poisson model. To guide applied scientists in the choice of the appropriate model to handle underdispersed count data, this work targets four of the most popular flexible models in addition to the Poisson model. Our purpose is to assess the robustness of the Poisson model to underdispersion and compare it with its alternatives through Monte Carlo simulations.

Specifically, this study assesses the relative performance of the Poisson, the Generalized Poisson, the Conway-Maxwell-Poisson, the Gamma Count and the Double Poisson regression models in underdispersed count data.

THEORETICAL UNDERPINNING

Poisson Model

Poisson model is the basic regression technique on which a variety of count models are based. Let λ be a positive real and Y a random variable. The probability distribution function of the Poisson law is:

$$P(\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, y = 0,1,2,\ldots \quad (1)$$

The Poisson regression model is defined for each count y_i through a log link function which expresses the Poisson distribution parameter λ_i (Expected number of counts, $i = 1,2,\ldots,n$) in terms of a linear function of a matrix of explanatory variables X_i (Frome, 1983):

$$ln(\lambda_i) = \beta X_i \quad (2)$$

X_i is the i^{th} row of the regression covariate matrix, and $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$ is the unknown k-dimensional vector of partial regression coefficients. The mean and the variance of Y_i are given by $E(Y_i|X_i) = Var(Y_i|X_i) = \lambda_i$. The regression parameters in $\beta_j (j = 1,2,\ldots,k)$ express the effect of the explanatory variable X_i on the expected count. More specifically, β_j gives the relative change (%) in the mean count for every unit increase in X_i.

The relationship between the mean and variance implies a goodness-of-fit index, $GOF = \frac{Var(Y)}{E(Y)} = 1$, i.e., equidispersion is established (Sellers *et al.*, 2017). If GOF is greater than 1, data are overdispersed and when it is less than 1, data are underdispersed. For these cases, the poisson model is no longer suitable.

The parameter β can be estimated by the maximum likelihood approach. The likelihood function is given by:

$$L(\beta) = \prod_{i=1}^{n} \quad p(Y_i = y_i|\lambda_i) = \prod_{i=1}^{n} \quad \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}. \qquad (3)$$

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

The log-likelihood function is given by:

$$L(\beta) = \sum_{i=1}^{n} \quad [-\lambda_i + y_i ln\lambda_i - ln(y_i!)] \qquad (4)$$

$$= \sum_{i=1}^{n} \quad [-exp(X_i^{\top}\beta) + y_i X_i^{\top}\beta - y_i!].$$

By differentiating the equation (4) with respect to β_j and equating each of the results to zero we get

$$g_i = \frac{\partial lnL}{\partial \beta_j} = \sum_{i=1}^{n} \quad \left[-exp(X_i^{\top}\beta)X_{ij} + y_i X_{ij}\right] = 0, j = 1,2,\ldots \qquad (5)$$

The Hessian is the matrix of second derivatives of the likelihood with respect to the parameter:

$$H = \frac{\partial^2 lnL}{\partial \beta_j \partial \beta_j^{\top}} = -\sum_{i=1}^{n} \quad (X_i^{\top} X_i e^{X_i^{\top}\beta}). \qquad (6)$$

The iterative algorithm of Newton-Raphson is used to find the estimated β:

$$\beta(i+1) = \beta(i))H^{-1}(i)g_i. \quad (7)$$

The asymptotic variance-covariance matrix of estimator is the inverse of the observed information matrix (Hessian matrix):

$$\hat{v}ar_{asy}(\hat{\beta}) = \left[-\sum_{i=1}^{n} \quad (X_i^{\top} X_i e^{X_i^{\top}\beta})\right]^{-1}. \qquad (8)$$

Hence, standard errors are square roots of the diagonal elements of the inverse of the information matrix.

Generalized Poisson Regression (GPR)

The generalized Poisson regression (GPR) model is often used to deal with overdispersed count data, although it can as well be used to model underdispersed data. This method was introduced by Consul and Jain (1973). GPR can model both overdispersion and underdispersion. Suppose Y_i is a count response variable that follows a generalized Poisson distribution, the probability

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

mass function of Y_i, $i = 1,2\ldots,n$ is given as follows (Famoye, 1993):

$$P(Y_i = y_i) = \left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}exp\left[-\frac{\lambda_i(1+\alpha y_i)}{1+\alpha\lambda_i}\right], y_i = 0,1,2,\ldots, \quad (9)$$

where $\lambda_i = exp(X_i\beta)$, X_i is the matrix of explanatory variables, β is a vector of regression parameters and α is the dispersion parameter.

The mean and variance of Y_i are given by:

$$E(Y_i) = \lambda_i, \quad (10)$$

$$Var(Y_i) = \lambda_i(1 + \alpha\lambda_i)^2. \quad (11)$$

When $\alpha > 0$, then the variance is larger than the mean, and this represents the situation of overdispersion. However, when $\alpha < 0$, the variance is smaller than the mean, and this represents the situation of underdispersion. The estimates of α and β in the GPR are obtained using the method of maximum likelihood.

The log-likelihood function is given by:

$$L(\alpha,\beta) = \sum_{i=1}^{n} \left[-ln\left(\frac{\lambda_i}{1+\alpha\lambda_i}\right) + (y_i - 1)ln(1 + \alpha y_i) - \frac{\lambda_i(1+\alpha y_i)}{1+\alpha\lambda_i} - ln(y_i!)\right]. \quad (12)$$

The maximum likelihood equations for estimating α and β are obtained by taking the partial derivatives of the log-likelihood function and equating to zero, giving:

$$\frac{\partial lnL}{\partial\alpha} = \sum_{i=1}^{n} \left[\frac{-\lambda_i y_i}{1+\alpha\lambda_i} + \frac{y_i(y_i-1)}{(1+\alpha y_i)} - \frac{\lambda_i(y_i-\lambda_i)}{(1+\alpha\lambda_i)^2}\right] = 0, \quad (13)$$

$$\frac{\partial lnL}{\partial\beta_j} = \sum_{i=1}^{n} \left[\frac{(y_i-\lambda_i)}{\lambda_i(1+\alpha\lambda_i)^2}\frac{d\lambda_i}{d\beta_j}\right] = 0, j = 1,2,\ldots,q. \quad (14)$$

Right from generalized linear model the link function for Poisson regression is $log(\lambda_i) = (X_i^{\top}\beta)$. Then substituting $\lambda_i = exp(X_i^{\top}\beta)$ on equations (13) and (14) can be rewritten as:

$$\frac{\partial lnL}{\partial\alpha} = \sum_{i=1}^{n} \left[\frac{(y_i-\lambda_i)}{(1+\alpha\lambda_i)^2}\right] = 0, \quad (15)$$

$$\frac{\partial lnL}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{(y_i - \lambda_i)x_j}{(1+\alpha\lambda_i)^2}\right] = 0 \quad j = 1,2,\dots,q. \qquad (16)$$

The parameters α and β are estimated by the Newton-Raphson method.

Estimation of α can also be done by using method of moments where α may be estimated by equating the Pearson chi-squared with $(n - q)$ degree of freedom (Husain & Bagmar, 2015), and is given by:

$$\frac{\partial lnL}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{(y_i - \lambda_i)x_j}{(1+\alpha\lambda_i)^2}\right] = n - q, \quad (17)$$

where n denotes the number of values and q the number of regression parameters.

Conway Maxwell Poisson Regression (COM-Poisson)

The conway Maxwell Poisson Regression, commonly named the COM-Poisson distribution, is a generalization of the Poisson distribution, first introduced by Conway and Maxwell (1962) for modelling queues and service rates. The COM-Poisson distribution has recently been reintroduced by statisticians for analyzing count data subjected to either over or underdispersion (Lord *et al*., 2010; Lord & Mannering, 2010). Its probability mass function is:

$$P(Y = y) = \frac{1}{Z(\lambda,v)} \frac{\lambda^y}{(y!)^v} \;, \quad y = 0,1,2\dots \qquad (18)$$

where Y is a discrete count; λ is a centering parameter that is approximately the mean of the observation in many cases ($\lambda_i = exp(X_i\beta)$ with β a vector of regression parameters, X_i the matrix of explanatory variables); and v is the dispersion parameter of the COM-Poisson v, distribution.

Thus, $Z(\lambda, v)$ is given by:

$$Z(\lambda, v) = \sum_{n=0}^{\infty} \frac{\lambda^n}{(n!)^v} \;. \qquad (19)$$

The mean and the variance of Y are given by:

$$E(Y) = \frac{\partial logZ(\lambda_i,v)}{\partial log\lambda_i}, \quad (20)$$

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

$$Var(Y) = \frac{\partial E(Y)}{\partial log\lambda_i}. \quad (21)$$

The λ is approximately the mean when ν is close to one and it differs substantially from the mean for small ν. These approximations fail for either $\nu > 1$ (i.e., underdispersion) or $\lambda^{1/\nu} < 10$ (i.e., low expected values). In many cases, a more useful parameterization for the COM-Poisson involves substituting $\mu = \lambda^{1/\nu}$, where μ is approximately the mode of Y (Lord *et al*., 2010; Lynch *et al*., 2014).

The log-likelihood function is given by:

$$\mathcal{L}(\lambda, \nu|\mathbf{Y}) = \sum_{i=1}^{n} \mathbf{y}_i \log\lambda_i - \nu \sum_{i=1}^{n} log y_i! - \sum_{i=1}^{n} \log Z(\lambda_i, \nu). \quad (22)$$

In $\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$; $\overline{\log(\mathbf{Y}!)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}!$

which is acquired as:

$$\log\mathcal{L}(\lambda, \nu|\mathbf{Y}) = n\overline{\mathbf{Y}}log\lambda - vlog(\mathbf{Y}!) - nlogZ(\lambda, \nu) \quad (23)$$

Then, the likelihood is a function of sufficient statistic $\underline{Y}$ and $log(Y!)$, for example, $\theta = (log(\lambda, \nu)$. Parameter Estimation can be determined by applying Newton Rhapson method. Loglikelihood gradient is measured as (Hayati *et al*., 2018):

$$\Delta\mathcal{L}(\theta) = n\begin{bmatrix} \overline{\mathbf{Y}} - E(\mathbf{X}) \\ -(\overline{\log(\mathbf{Y}!)}) - E[\log(\mathbf{Y}!)] \end{bmatrix} \quad (24)$$

and the second derivation is:

$$\Delta^2\mathcal{L}(\theta) = n\begin{bmatrix} -var(\mathbf{Y}) & cov(\mathbf{Y}, \log(\mathbf{Y}!)) \\ cov(\mathbf{Y}, \log(\mathbf{Y}!)) & var(\log(\mathbf{Y})) \end{bmatrix} \quad (25)$$

Newton Rhapson formula is estimated as:

$$\theta^{t+1} = \theta^t - (\Delta^2\mathcal{L}(\theta))^{-1}\Delta\mathcal{L}(\theta) \quad (26)$$

Initial value probably appears for iteration that MLE from Poisson is acquired as

$(\lambda = \bar{Y}, \nu = 1)$ so $\theta = (\log(\mathbf{Y}),1)$.

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

Gamma Count Model

Gamma model was proposed by Oh *et al.* (2006) to analyze crash data exhibiting underdispersion. Lord and Mannering (2010) once employed this model as an alternative to handle underdispersion. The Gamma Count probability model for count data is given as:

$$P(y_i = j) = Gamma(\alpha j, \lambda_i) - Gamma(\alpha j + \alpha, \lambda_i), \quad j = 0,1,2, \dots, \quad i = 1, \dots n, \qquad (27)$$

where $\lambda_i = exp(X_i\beta)$, β is a vector of regression parameters, X is the matrix of explanatory varibles and α is the dispersion parameter,

$$Gamma(\alpha j, \lambda_i) = 1, \quad if \quad j = 0 \qquad (28)$$

$$Gamma(\alpha j, \lambda_i) = \frac{1}{\Gamma(\alpha j)} \int_0^{\lambda_i} \quad t^{\alpha j - 1} e^{-t} dt, \quad if \quad j > 0. \qquad (29)$$

The conditional mean and variance functions are given by:

$$E(y_i) = \sum_{i=1}^{\infty} \quad jGamma(\alpha j, \lambda_i), \qquad (30)$$

$$var(y_i) = \sum_{i=1}^{\infty} \quad j^2[Gamma(\alpha j, \lambda_i) - Gamma(\alpha j + \alpha, \lambda_i)] - E(y_i|X_i)^2 \qquad (31)$$

For $\alpha > 1$, the model shows underdispersion; for $\alpha < 1$, the model exhibits overdispersion; for $\alpha = 1$, it is equidispersion meaning the gamma model reduces to Poisson model.

The log-likelihood function is given by:

$$L(y_i|\alpha, \lambda) = \sum_{i=1}^{n} \quad log(Gamma\left(y_i\alpha, \alpha exp(\lambda_i) - Gamma\left(y_i(\alpha + 1), \alpha exp(\lambda_i)\right)\right) \qquad (32)$$

Parameter estimation requires numerical maximization of equation (32).

Double Poisson Model (DP)

Based on the double exponential family, Efron (1986) proposed the Double Poisson distribution. The Double Poisson model, based on the distribution, has two parameters μ and θ. Zou *et al.* (2013) and Zou *et al.* (2011) give the approximate probability mass function (p.m.f) as per the following equation:

$$P(Y = y) = f_{\mu,\theta}(y) = (\theta^{\frac{1}{2}} e^{-\theta\mu})(\frac{e^{-y}y^y}{y!})(\frac{e^{\mu}}{y})^{\theta y}, \quad y = 0,1,2\dots, \qquad (33)$$

where θ is the dispersion parameter, $\mu = exp(X_i\beta)$, β is a vector of regression parameters, and X is the covariate matrix. The exact double Poisson density is given as:

$$P(Y = y) = \tilde{f}_{\mu,\theta}(y) = c(\mu, \theta) f_{\mu,\theta}(y), \qquad (34)$$

where the factor $c(\mu, \theta)$ can be calculated as:

$$\frac{1}{c(\mu,\theta)} = \sum_{y=0}^{\infty} \quad f_{\mu,\theta}(y) \approx 1 + \frac{1-\theta}{12\mu\theta}(1 + \frac{1}{\mu\theta}). \qquad (35)$$

with $c(\mu, \theta)$ being the normalizing constant nearly equal to 1. The constant $c(\mu, \theta)$ ensures that the density sums to unity. The expected value and the standard deviation (SD), referring to the exact density $\tilde{f}_{\mu,\theta}$, are estimated as follows:

$$E(Y) \approx \mu, \qquad (36)$$

$$SD(Y) \approx \left(\frac{\mu}{\theta}\right)^{\frac{1}{2}}. \qquad (37)$$

Thus, the Double Poisson model allows for both overdispersion $(\theta < 1)$ and underdispersion $(\theta > 1)$. When $\theta = 1$, the Double Poisson distribution collapses to the Poisson distribution. In DP model, particular focus should be given to the use of the normalizing constant which includes an infinite series (Zou *et al*., 2013). The infinite series is as follows:

$$\sum_{y=0}^{\infty} \quad f_{\mu,\theta}(y). \qquad (38)$$

Therefore, the log-likelihood function is given by:

$$L(\mu, \theta|Y) = \sum_{i=0}^{n} \quad \left\{\frac{1}{2}(\theta) - \theta\mu - y_i + y_i ln y_i - ln\Gamma(y_i + 1) + \theta y_i(ln\mu - ln y_i + 1) - ln(c(\mu, \theta))\right\}. \qquad (39)$$

Parameter estimations requires numerical maximization of equation (39).

METHODOLOGY

Data Generation Process

To evaluate the robustness of Poisson model and its alternatives to handle underdispersed count data, a simulation study was conducted. The simulation was done by mimicking the work of Nkegbe and Shankar (2014) on the adoption intensity of soil and water conservation practices by smallholders where their data exhibited underdispersion.

To conduct the simulation, we used SHLAB (Total self-help labour for 2008/09 agricultural year (in man-days)) as covariate (x_1) because it is significant in the model, and the response variable (y) was the number of conservation practices adopted. Thereby, the coefficients β_0

and β_1 take respectively the values -0.2088 and 0.0029 following the result of (Nkegbe & Shankar, 2014). The mean μ is expressed as:

$$E(y|SHLAB) = \mu = exp(-0.2088 + 0.0029 \times SHLAB) \quad (40)$$

In R software, the response variable y was generated using the Generalized Condensed Poisson distribution (GCP) which is a mixture of the Asynchronous counting distribution introduced by Whittlesey and Haight (1961). The Erlang count (asynchroneous Poisson) distribution, with mean μ and condensing coeffcient (shape) m, is denoted $AP(\mu, m)$ and has probability mass function (Whittlesey & Haight, 1961):

$$f_{ap}(y|\mu, m) = \sum_{t=1-m}^{m-1} \quad \frac{m-|t|}{m} f_p(my + t|\lambda), \qquad (41)$$

where $f_p = \frac{e^{-\mu}\mu^y}{y!}$, $\lambda = m\mu$ and y is the response variable.

We considered a positive real $m \in R \quad (m \geqslant 1)$ and let $m_0 = \lfloor m \rfloor$ the integral part of m. The Generalized Condensed Poisson distribution (GCP) with mean μ and condensing coefficient (shape) m was denoted $GCP(\mu, m)$ defined as the mixture of $AP(\mu, m_0)$ and $AP(\mu, m_0 + 1)$ with respective mixing probabilities $1 - \rho$ and ρ, where $\rho = (m - m_0)\frac{m_0+1}{m}$. This implicit definition actually gives an algorithm to generate variates from $GCP(\mu, m)$. The $GCP(\mu, m)$ has probability distribution function:

$$f(y|\mu, m) = 1 - \rho f_{ap}(y|\mu, m) + \rho f_{ap}(y|\mu, m_0 + 1) \qquad (42)$$

The core advantage of this distribution over other underdispersion distributions is that it is easy to simulate. If the shape parameter $m = 1$, the Asynchronous/Condensed Poisson distribution reduces to Poisson distribution. The parameter m can be set in such a way that the underdispersion level or parameter of underdispersion ($k = var/mean$) takes a fixed value ($0 < k \leqslant 1$) given a fixed expectation value.

In our simulation, we have considered three different values of k (0.81, 0.5 and 0.2) and $k = 1$ as a control. These three values were selected from the underdispersion levels that we found from the review (Nkegbe & Shankar, 2014; Min *et al.*, 2017). Five sample sizes were considered in our simulation (20, 50, 100, 300 and 500) inspired by the works of Hayati *et al.* (2018).

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

Table 1: Simulation Scenario

Source of data	Parameters	Sample size	Level of underdispersion	Model Comparison	Number of replications
Y generates from Asynchronous/Condensed Poisson distribution	$\beta_0 = -0.2088$ $\beta_1 = 0.0029$	$n = 20,$ 50, 100, 300 and 500	$k = 0.81, 0.5,$ and 0.2 A control $k = 1$	Poisson, COM-Poisson Double Poisson and Gamma count	1000 times

Fitting Models Studied to Simulated Data

In this study, the software R version 3.5.2 was used for computations. We used the *glm* function of the package *MASS* with the option $family = poisson(log)$ for Poisson model, the *glm.cmp* function of the package *COMPoissonReg* for COM-Poisson model, the *vglm* function of the package *VGAM* for Generalized Poisson Regression, the *gamlss* function of the package *gamlss* with the option *family = DPO* for Double Poisson regression and the *renewalCount* function of the package *Count* for Gamma Count regression.

After each model estimation, parameter estimates as well as their standard deviations and 95% confidence intervals were extracted.

Performance Measures

For each combination n and k, models were compared following performance measures (Morris *et al.*, 2019). The 5 different models were compared using Relative Bias (RBias) and Root Mean Squared Error (RMSE). The slope indicates the relative relationship between simulated and measured values. For the slope β_1, we get:

$$Bias = \frac{1}{N}\sum_{i=1}^{N} \quad \hat{\beta}_{1i} - \beta_1; \quad RBias = \frac{Bias}{\beta_1} \times 100, \quad (43)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \quad (\hat{\beta}_{1i} - \beta_1)^2}, \qquad (44)$$

where $\hat{\beta}_{1i}$ is the estimated parameter, β_1 is the true value and i=1...N, N is the number of simulations.

The model showing the low values of these statistics is the best. Relative Bias and RMSE were plotted and analysed.

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

RESULTS

Performance of Poisson Model Against Alternatives

Figure 1 shows boxplots of relative bias of Poisson model and its alternatives. The Generalized Poisson Regression, Poisson, Double Poisson and COM-Poisson model showed the lowest values of the Relative Bias. The Gamma Count model has the highest median value. Moreover, the dispersion around the median values of relative bias was large for the Gamma Count model.

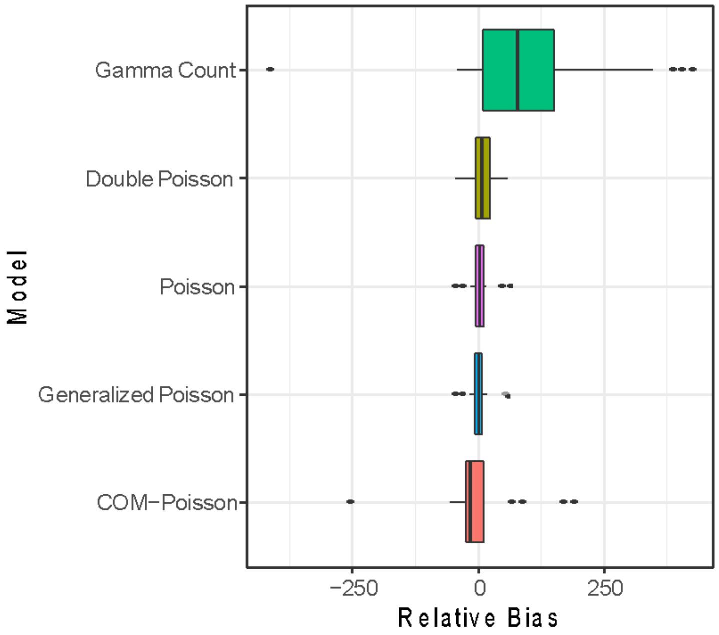

Figure 1: Relative performance of Poisson model and its alternatives irrespective of underdispersion levels and sample size

The Effect of Underdispersion and Sample Size

Figure 2 shows relative bias of slopes of the five models according to the values of k for each sample size. The relative bias is close to zero for Double Poisson, Generalized Poisson Regression and Poisson model along the level of underdispersion k for all sample sizes. However, relative bias for COM-Poisson model were close to zero for sample size $n = 50$. It also shows relative constant values of relative bias which is close to zero when k is between 0.2 and 0.5 for low sample size $(n = 20)$. The COM-Poisson model has relative bias close to zero when the dispersion is 0.2, 0.81 or 1. The relative bias of Gamma Count model is close to zero when the dispersion is between 0.81 to 1. In summary, Gamma Count model is more biased compared to other alternative models.

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

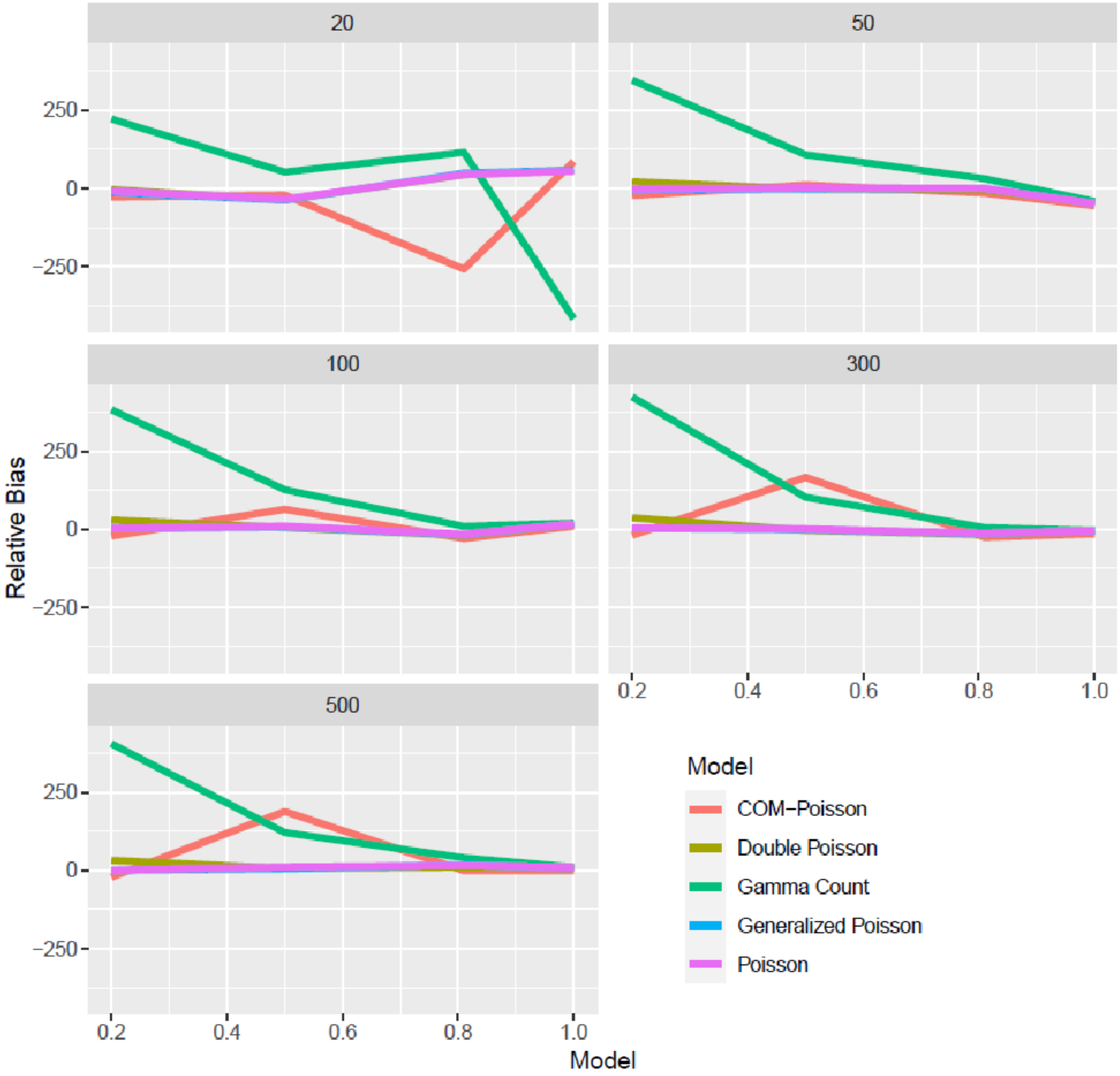

Figure 2: Plots of relative bias against sample size for Poisson model and its alternative

Figure 3 shows values of RMSE for Poisson model and its alternatives. The COM-Poisson model is the best performer model when the level of underdispersion is 0.2. The performance of Poisson model is quite similar that of the Double Poisson and Generalized Poisson for all combination of n and k. The COM-Poisson model is also the best performer model when k is 0.81 for most sample sizes ($n = 50, 100, 300$ and 500). When $k = 1$ for sample sizes ($n = 50, 100, 300$ and 500), we note the best behavior for all models but the Poisson model is the best performer.

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

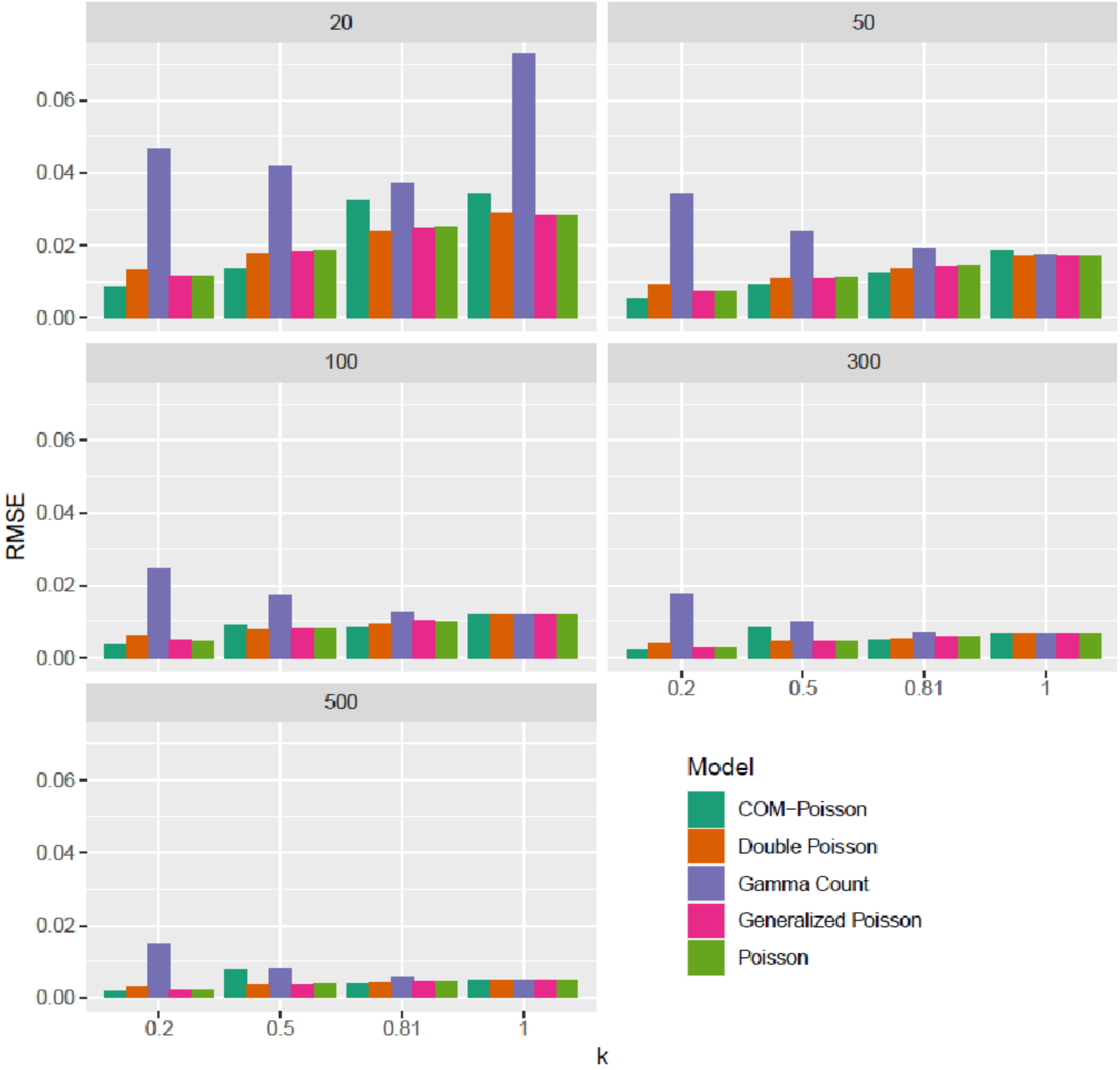

Figure 3: Variation of Root Mean Square Error against sample size for Poisson model and its alternatives.

DISCUSSION

Underdispersion occurs when the variance is lower than the mean. The dispersion index (variance to mean ratio) takes values between 0 and 1. A dispersion value of 1 means that the variance is equal to the mean and thus the Poisson model hypothesis holds. However, this is not always the case. According to the literature, several models have been developed to treat cases of underdispersion (Sellers & Shmueli, 2010; Famoye & Wang, 1997; Zou *et al*., 2013).

Four alternative models have been studied here. Results of the Monte Carlo simulation show that the Poisson model performs poorly for underdispersed count data. This supports findings of Lynch *et al*. (2014) and Barakat (2016) on the low performance of Poisson model to handle underdispersed count data. Therefore, it remains the best model in the case of equidispersion ($k = 1$).

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

For its alternatives, the results of the simulation showed that the COM-Poisson model is the best performer compared to other models when the underdispersion is severe (equal to 0.2). The COM-Poisson model yielded the best performance when the dispersion was moderate (0.81) for sample sizes ($n = 50, 100, 300$ and 500). For low sample sizes ($n = 20$ and 50) and the level of underdispersion equal to 0.5, the COM-Poisson was also found to be the best model. However, Double Poisson and Generalized Poisson Regression models outperformed COM-Poisson model for relatively large sample sizes ($n = 100$, 300 and $n = 500$) for the same level of underdispersion ($k = 0.5$). The effectiveness of the COM-Poisson model to handle different levels of underdispersion of count data was also demonstrated by Sellers & Shmueli (2010). Our results partly support the performance of COM-Poisson but highlight that COM-Poisson model is not always the best in all situations for each combination of n and k.

The elegance of the COM-Poisson regression model lies in its ability to address applications containing a wide range of dispersion in a parsimonious way. Geedipally *et al*. (2008) and Wu *et al*. (2013) used the Bayesian approach of the COM-Poisson regression model to estimate parameters. They found that the COM-Poisson is a flexible method for analyzing count data and also Bayesian estimation provides an attractive alternative for estimating the coefficients of the model compared to the method of the maximum likelihood where the likelihood equation for the COM-Poisson is complex, making analytical and numerical maximum likelihood estimation difficult. Contrary of the suggestion of Zou *et al*. (2013), the maximum likelihood estimation of the parameters of COM-Poisson was greatly simplified when compared to the Bayesian estimating method. Moreover, for Double Poisson, Generalized Poisson Regression and Gamma Count models, the maximum likelihood estimation of the parameters was often used. Famoye (1993) found that the bounded dispersion parameter when underdispersion occurs greatly diminishes the applicability of the Generalized Poisson Regression model to count data. This was shown by our results on the limits of GPR when the underdispersion is severe or moderate.

Therefore, Hayati *et al*. (2018) found that the COM-Poisson model is more flexible in dealing data with underdispersion than Generalized Poisson Regression because the underdispersion value area is wider than the Generalized Poisson Regression model. However, results showed that no model is better in all situations, so the use of a model depends on the situation we have. Therefore, in practice, all alternatives should be tested first and the best selected.

CONCLUSION AND RECOMMENDATION

To evaluate the performance of the Poisson model and its alternatives following different underdispersion parameters and sample sizes, the Monte Carlo simulation approach was used. Results show that the Poisson model is not very effective to handle underdispersion count data compared to its alternatives. All alternative models studied showed their effectiveness in handling underdispersed count data. However, the COM-Poisson model shows better statistical performance in case of severe underdispersion than other models. Moreover, the Generalized Poisson Regression and Double Poisson models have results quite similar in term of performance. It is suggested that further research should be conducted using real datasets to confirm the findings of in this research.

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT

Acknowledgements

This work was supported by intra-ACP/CSAA mobility scholarships (Intra-Academic Mobility Project: Intra-University Co-operation to Train Crop Scientists for Enhancing Agriculture in Africa).

REFERENCES

Barakat, B. F. (2016). Generalised Poisson distributions for modelling parity. Vienna Institute of Demography Working Papers.

Consul, P. C., Famoye, F. (1992). Generalized poisson regression model. Communications in Statistics - Theory and Methods, 21, 89–109.

Efron, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression.Journal of American Statistical Association, 81(295), 709-721.Famoye, F. (1993). Restricted generalized poisson regression model. Communications in Statistics - Theory and Methods, 22, 1335–1354.

Famoye F. (1993). Restricted generalized poisson regression model. Communications in Statistics - Theory and Methods, 22, 1335–1354.

Famoye F., Wang, W. (2004). Censored generalized Poisson regression model. Computational Statistics & Data Analysis, 46, 547–560.

Forthmann, B., Gühne, D., Doebler, P. (2020). Revisiting dispersion in count data item response theory models: The Conway–Maxwell–Poisson counts model. British Journal of Mathematical and Statistical Psychology, 73, 32–50.

Geedipally, S. R., Guikema, S. D., Dhavala, S. S., Lord, D. (2008). Characterizing the Performance of the Bayesian Conway-Maxwell Poisson Generalized Linear Model. In: Association, American S. (Ed.): Joint Statistical Meetings, p. 22. Citeseer.

Hayati, M., Sadik, K., Kurnia, A. (2018). Conwey-Maxwell Poisson Distribution: Approach for Over- and-Under-Dispersed Count Data Modelling. IOP Conference Series: Earth and Environmental Science, 187, 012039.

Husain, M. M., Bagmar, M. S. H. (2015). Modeling Under-dispersed Count Data Using Generalized Poisson Regression Approach. Global Journal of Quantitative Science, 2, 22–29.

Whittlesey, J. R., Haight, F. A. Counting distributions for an erlang process. Ann Inst Stat Math 13, 91–103 (1961). https://doi.org/10.1007/BF02868862

Kokonendji, C. C. (2014) – Over-and underdispersion models. Methods and Applications of Statistics in Clinical Trials, 2, 506–526.

Kokonendji, C. C., Mizère D., Balakrishnan, N. (2008). Connections of the Poisson weight function to overdispersion and underdispersion. Journal of Statistical Planning and Inference, 138, 1287–1296.

Lord, D., Geedipally, S. R., Guikema, S. D. (2010). Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion. Risk Analysis: An International Journal, 30, 1268–1276.

Lord, D., Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice, 44, 291–305. https://doi.org/10.1016/j.tra.2010.02.001

Lynch, H. J., Thorson, J. T., Shelton, A. O. (2014). Dealing with under-and over-dispersed count data in life history, spatial, and community ecology. Ecology, 95, 3173–3180.

Min, S., Huang, J., Waibel, H. (2017). Rubber specialization vs crop diversification: the roles of perceived risks. China Agricultural Economic Review, 9, 188–210.

Morris, T. P., White, I. R., Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38, 2074–2102.

Nkegbe, P. K., Kuunibe, N., Sekyi, S. (2017). Poverty and malaria morbidity in the Jirapa District of Ghana: A count regression approach (G Aye, Ed,). Cogent Economics & Finance, 5.

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

Sellers, K. F., Borle, S., Shmueli, G. (2012). The COM-Poisson model for count data: a survey of methods and applications. Applied Stochastic Models in Business and Industry, 28, 104–116.

Sellers, K. F., Morris, D. S. (2017). Underdispersion models: Models that are "under the radar." Communications in Statistics – Theory and Methods, 46, 12075–12086.

Sellers, K. F., Premeaux, B. (2020). Conway–Maxwell–Poisson regression models for dispersed count data. Wiley Interdisciplinary Reviews: Computational Statistics, pp. 1–13.

Sellers, K. F., Swift, A. W., Weems, K. S. (2017). A flexible distribution class for count data. Journal of Statistical Distributions and Applications, 4.

Wu, G., Holan, S. H., Wikle, C. K. (2013). Hierarchical Bayesian Spatio-Temporal Conway–Maxwell Poisson Models with Dynamic Dispersion. Journal of Agricultural, Biological, and Environmental Statistics, 18, 335–356.

Zou, Y., Geedipally, S. R., Lord, D. (2013). Evaluating the double Poisson generalized linear model. Accident Analysis & Prevention, 59, 497–505.

Zou, Y., Lord, D., Geedipally, S. R. (2011). Over-and under-dispersed count data: Comparing the Conway-Maxwell-Poisson and Double-Poisson distributions. In: 91 st Annual Meeting of the Transportation Research Board. Citeseer.

Article DOI: 10.52589/AJMSS-1WPJQHYT
DOI URL: https://doi.org/10.52589/AJMSS-1WPJQHYT