Brazilian Journal of Biometrics, **41**, 274–286 (2023)
doi:10.28951/bjb.v41i3.635

BRAZILIAN JOURNAL OF BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Counting models for overdispersed data: A review with application to *tuberculosis* data

Alcinei Místico Azevedo,[1] Ítallo Jesus Silva,[2] Marcela Carlota Nery,[2] Honovan Paz Rocha,[3] and Rogério Alves Santana*,[4]

[1] Instituto de Ciências Agrarias, Universidade Federal de Minas Gerais, Montes Claros-MG, Brasil.
[2] Produção Vegetal, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina-MG, Brasil.
[3] Instituto de Engenharia Ciência e Tecnologia, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Janaúba-MG, Brasil.
[4] Faculdade de Ciências Sociais Aplicadas e Exatas, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Teófilo Otoni-MG, Brasil.
*Corresponding author. Email: rogerio.santana@ufvjm.edu.br

(Received: September 14, 2022; Revised: January 25, 2023; Accepted: March 24, 2023; Published: September 25, 2023)

Abstract

The present work reviews distributions for counting data: Poisson; Negative Binomial; COM-Poisson and Generalized Poisson, and their regression models. Aspects such as parameter estimation and model choice criteria are presented. And as an application example, we use the regression models of these distributions to explain the relationship between *tuberculosis* notifications with the HDI Human Development Index of the 102 cities in the state of Alagoas. The existing relationship between notifications of *tuberculosis* with HDI is significant and overdispersion at the level $\alpha = 5\%$ of probability, and the COM-Poisson distribution regression model was the best fit data, according to the Akaike AIC and Bayesian BIC information criteria.

Keywords:Tuberculosis; Negative Binomial; COM-Poisson; Generalized Poisson.

1. Introduction

Counting data is common in areas of agriculture (Carvalho *et al.,* 2018; Hall, 2000); education Desjardins (2016); engineering Anastasopoulos & Mannering (2009); industry Lambert (1992); psychology Atkins & Gallop (2007); public health Khan *et al.* (2011), among others. The linear regression model of the Poisson distribution is one of the most used models in the areas of science, since most of the phenomena meet the postulates of the Poisson distribution Shmueli *et al.* (2005). However, in cases where the data are overdispersed, that is, when the data variance is greater than the mean, the Poisson model is not indicated because it accommodates only equidispersion, when the variance is equal to the mean. In this case, the linear regression models of the Negative Binomial, COM-Poisson and Generalized Poisson distributions are the most indicated because they have a dispersion parameter that can accommodate overdispersion. Data overdispersion can be caused by too many zeros; *outliers*; the way the sampling process was conducted; the use of inappropriate link function in the model; non-linear effects considered as linear in the systematic component of the model; omitted covariates in the model, among others (Ridout & Besbeas, 2004; Hilbe, 2011). In any case, overdispersion cannot be neglected, otherwise the model estimates will be biased Ridout & Besbeas (2004).

Underdispersion may also occur in the data, a less common situation when the data variance is less than the mean. Works involving underdispersion data are found in (Shmueli *et al.,* 2005; Sellers & Morris, 2017; Barlow & Proschan, 1996; citeridout2004; Hayati *et al.,* 2018). For these types of data, the COM-Poisson and Generalized Poisson distribution models, as they have a more flexible dispersion parameter to accommodate underdispersion, are more suitable than the Negative Binomial distribution Santana (2019). Works involving these types of models are still scarce in some literature, such as in the agricultural sciences. For example, in a search on the SciELO base page in SciELO (2022a), of the 115,681 published articles, only 59 articles were found involving the distribution Poisson, 15 papers involving the Negative Binomial distribution and no papers involving the COM-Poisson and Generalized Poisson distributions. This problem, due to the lack of scientific articles with statistical rigor in the agricultural sciences, is reported in Carvalho *et al.* (2019) in a bibliographic review survey.

This problem also spreads to other areas, such as health case. In a search on the SciELO database in SciELO (2022b), of the 506,199 published articles, no article was found involving the COM-Poisson and Generalized Poisson distributions. This demonstrates the need to publish scientific papers involving more sophisticated distributions in related areas of statistics.

In view of the above, this work makes a bibliographic review of Poisson distributions; Negative Binomial; COM-Poisson and Generalized Poisson and their regression models, with the aim of providing information to professionals in sciences related to statistics who are not familiar with regression models and the criteria for choosing these distributions. As an application example, we used health data from the profile of municipalities in 2020 on notifications of *tuberculosis* and the Human Development Index of the 102 cities in the state of Alagoas, available at (de Saúde do Estado de Alagoas (2017); BRASIL (2021); de Saúde Perfil dos Municípios Alagoanos (2022)).

This article is organized as follows: in Section 2, we review regression models for Poisson distributions; Negative Binomial; COM-Poisson and Generalized Poisson and selection criteria. In Section 3, we present an application with these models to explain the relationship between *tuberculosis* notifications with the Human Development Index for all cities in the state of Alagoas. Finally, we present the conclusion in Section 4.

2. Materials and methods

2.1 Poisson distribution model (Po)

We start counting models with the Poisson distribution model. Consider Y_i as a random variable with Poisson distribution and we write $(Y_i \sim \mathrm{Po}(\lambda_i))$, if Y_i has the following probability mass function (*fmp*)

$$\Pr(Y_i = y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots, \tag{1}$$

where $\lambda_i > 0$ corresponds to the average number of occurrences of a given event, y_i is the realization of the random variable Y_i, with mean $\mathrm{E}(Y_i) = \lambda_i$ and variance $\mathrm{Var}(Y_i) = \lambda_i$. In the regression structure of the Poisson distribution, the parameter λ_i is related to the covariable $\mathbf{x}_i$ through the logarithmic link function (Paula, 2004) that is,

$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{2}$$

where $\mathbf{x}_i = (1, x_1, \ldots, x_p)$ is a vector of explanatory variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\top$ is a vector of unknown model parameters, both of dimension (p+1). Then, for a random sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, with n independent observations, the log-likelihood function of the Poisson model is given by

$$\ell(\boldsymbol{\lambda}; \mathbf{y}) = \sum_{i=1}^{n} \left\{ y_i \log(\lambda_i) - \lambda_i - \log(y_i!) \right\}. \tag{3}$$

For details on the log-likelihood function, and the various link functions of the Poisson distribution, see (Nelder & Wedderburn, 1972; Cameron & Trivedi, 2013).

2.2 Negative Binomial Distribution Model (NB)

Let Y_i be a random variable representing the number of trials before the rth success in n Bernoulli trials. We say that the random variable Y_i has a Negative Binomial distribution with parameters r and p and we write $(Y_i \sim \text{NB}(r, p))$, if Y_i has the following *fmp*

$$\Pr(Y_i = y_i) = \binom{y_i + r - 1}{y_i} p^r (1-p)^{y_i}, \qquad y_i = 0, 1, 2, \ldots, \tag{4}$$

where r>0 is a fixed integer; p is the probability of success on the rth success and y_i is a realization of the random variable Y_i with $\text{E}(Y_i) = r(1-p)/p$ and $\text{Var}(Y_i) = r(1-p)/p^2$ (Bartko, 1960). Specifically, the Negative Binomial distribution is equivalent to the Geometric distribution with parameter p, that is, $(Y_i \sim \text{Geom}(p))$ when $r = 1$ (Hilbe, 2011).

There are several ways to represent the Negative Binomial distribution, depending on the parameterization to be used, one of them can be obtained through the parameterization $p = \mu_i/(\phi + \mu_i)$ and $\phi = r$ used in Nelder & Wedderburn, 1972, and by replacing it in Equation (4) we obtain

$$\Pr(Y_i = y_i) = \frac{\Gamma(y_i + \phi)}{\Gamma(\phi)\Gamma(y_i + 1)} \left(\frac{\mu_i}{\phi + \mu_i} \right)^{y_i} \left(\frac{\phi}{\phi + \mu_i} \right)^{\phi}, \qquad y_i = 1, 2, \ldots, . \tag{5}$$

Where $\mu_i > 0$ and $\phi > 0$ are distribution parameters $\text{NB}(\mu_i, \phi)$, $\Gamma(\cdot) = \int_0^\infty x^{\nu-1} e^{-\nu} dx$ is the gamma function, and y_i is a realization of the random variable Y_i with mean $\text{E}(Y_i) = \mu_i$ and variance $\text{Var}(Y_i) = \mu_i + \mu_i^2/\phi$. Thus, the dispersion index $\text{Var}(Y_i)/\text{E}(Y_i) > 1$, that is, the NB distribution can be used to model the overdispersion in the data and ϕ^{-1} corresponds to the dispersion parameter, for details see (Hinde & Demétrio, 1998; Park & Lord, 2009; Nelder & Wedderburn, 1972; Hilbe, 2011). Specifically, the NB distribution is equivalent to the Poisson distribution when $\phi \to \infty$ (Cameron & Trivedi, 2013).

In the regression structure of the NB distribution with regression on the mean, the parameters μ_i and ϕ are related to the covariate $\mathbf{x}_i$ and to the parameter τ , through the logarithmic binding function (Hilbe 2011), that is:

$$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \qquad \text{and} \qquad \log(\phi) = \tau, \tag{6}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\top$ and $\mathbf{x}_i = (1, x_1, \ldots, x_p)$, are the vectors of parameters and explanatory variables, both of dimension $p + 1$. Thus, for a random sample with n independent observations, the log-likelihood function of the NB distribution model is given by

$$\ell(\mu, \phi; \mathrm{y}) = \sum_{i=1}^{n} \left\{ \log(\Gamma(y_i + \phi)) - \log(\Gamma(\phi)) - \log(\Gamma(y_i + 1)) + y_i \log(\mu_i) + \phi \log(\phi) - (\phi + y_i) \log(\phi + \mu_i) \right\}. \tag{7}$$

For details on the log-likelihood function, and the various link functions of the Negative Binomial distribution, see(Paula, 2004, p. 306; Nelder & Wedderburn, 1972; Hilbe, 2011, p. 190; Cameron & Trivedi, 2013; Ripley *et al.,* 2013).

2.3 COM-Poisson distribution model (COMP)

The COM-Poisson distribution was proposed in the 1960s by Richard W. Conway and William L. Maxwell in (Conway & Maxwell, 1962) to model service fees, and then forgotten. Four decades later, COM-Poisson is revived with the paper by Shmueli et. al. in (Shmueli *et al.,* 2005). Since then, numerous works involving COM-Poisson extensions have been developed (Santana, 2019). Thus, we say that the random variable Y_i has a COM-Poisson distribution and we write $Y_i \sim \mathrm{COMP}(\lambda_i, \nu)$, if it has the following *fmp*

$$\Pr(Y_i = y_i) = \frac{\lambda_i^{y_i}}{S(\lambda_i, \nu)(y!)^{\nu}}, \quad y_i = 0, 1, 2, ..., \tag{8}$$

where $\lambda_i > 0$ is a center parameter that is approximately the mean when $\nu \approx 1$ (Barriga & Louzada, 2014); $\nu \geq 0$ is the dispersion parameter of the COM-Poisson distribution with under-equi-overdispersion when ($\nu > 1$, $\nu = 1$ and $\nu < 1$), and $S(\lambda_i, \nu)$ is the COMP normalization constant (Shmueli *et al.,* 2005), which is given by

$$S(\lambda_i, \nu) = \sum_{j=0}^{\infty} \left(\frac{\lambda^j}{j!} \right)^{\nu}. \tag{9}$$

The COM-Poisson distribution has three distributions as particular cases, that is, it is equivalent to Poisson when $\nu = 1$, the geometric when ($\lambda_i < 1$ and $\nu = 0$) , and the Bernoulli with probability of success $\lambda_i/(1+\lambda_i)$ when $\nu \to \infty$, (Boatwright *et al.,* 2006). The COM-Poisson distribution does not have a closed expression for its moments concerning parameters λ and ν, Guikema & Goffelt (2008). However, an approximation for its mean $\mathrm{E}(Y_i)$ and its variance $\mathrm{Var}(Y_i)$, is presented in Shmueli *et al.* (2005), according to Equations10 and 11:

$$\mathrm{E}(Y_i) = \lambda_i \frac{\partial \log S(\lambda_i, \nu)}{\partial \lambda_i} \approx \lambda_i^{1/\nu} - \frac{\nu - 1}{2\nu} \tag{10}$$

$$\mathrm{Var}(Y_i) = \lambda_i \frac{\partial E(Y_i)}{\partial \lambda_i} \approx \frac{1}{\nu} \lambda_i^{1/\nu}, \tag{11}$$

another structure for the COM-Poisson mean and variance is presented in Guikema & Goffelt (2008) with the parameterization of COM-Poisson for the GLM structure. For details, see (Lord *et al.,* 2008; Huang, 2017; Ribeiro Junior, 2019).

In the regression structure of the COM-Poisson distribution, with regression on the mean, the parameters λ_i and ν are related to the covariate $\mathbf{x}_i$ and to the parameter ξ, through the logarithmic binding function Sellers & Shmueli (2010), that is:

$$\log(\lambda_i) = \mathbf{x}_i^{\top} \boldsymbol{\beta}, \qquad \text{and} \qquad \log(\nu) = \xi, \tag{12}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\top$ and $\mathbf{x}_i$ are vectors of parameters and covariates of dimensions $p+1$. For a random sample with n independent observations, the log-likelihood function of the COM-Poisson model is given by

$$\ell(\boldsymbol{\lambda}, \nu; \mathbf{y}) = \sum_{i=1}^{n} \left\{ y_i \log(\lambda_i) - \log S(\lambda_i, \nu) - \nu \log(y_i!) \right\}. \tag{13}$$

For details on the log-likelihood function and the linkage function of the COM-Poisson distribution, see (Sellers & Shmueli, 2013; Guikema & Goffelt, 2008; Sellers & Shmueli, 2010).

2.4 Generalized Poisson distribution model (GP)

Let Y_i be a random variable with Generalized Poisson distribution (GP) in Consul (1989), and denote $Y_i \sim \text{GP}(\lambda_i, \varphi)$, if Y_i has the following *fmp*

$$\Pr(Y_i = y_i) = \frac{\lambda_i(\lambda_i + \varphi y_i)^{y_i - 1} e^{-(\lambda_i + \varphi y_i)}}{y_i!}, \qquad y_i = 0, 1, \ldots, \tag{14}$$

where $\lambda_i > 0$ and $\max(-1 - \lambda_i/4) < \varphi < 1$ are the center and dispersion parameters of the GP distribution, with sub-equi-overdispersion when ($\varphi < 0$, $\varphi = 0$ and $\varphi > 0$), specifically the GP distribution is equivalent to the Poisson distribution when $\varphi = 0$. And y_i is a realization of the random variable Y_i with mean $\text{E}(Y_i) = \lambda_i/(1-\varphi)$ and variance $\text{Var}(Y_i) = \lambda_i/(1-\varphi)^3$. For details, see Ridout *et al.* (2001).

In the regression structure of the GP distribution, with regression on the mean, the parameters λ_i and φ are related to the covariate $\mathbf{x}_i$ and the parameter ζ , through the logarithmic binding function (Consul, 1989), that is:

$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \qquad \text{and} \qquad \log(\varphi) = \zeta, \tag{15}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ and $\mathbf{x}_i = (1, x_1, \ldots, x_p)$, are the vectors of parameters and explanatory variables, both of dimension $p + 1$. ζ is one more parameter of the model to be estimated. For a random sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, with n independent observations the log-likelihood function of the GP distribution model is given by

$$\ell(\boldsymbol{\lambda}, \varphi; \mathbf{y}) = \sum_{i=1}^{n} \left\{ \log(\lambda_i) + (y_i - 1)\log(\lambda_i + \varphi y_i) - (\lambda_i + \varphi y_i) - \log(y_i!) \right\}. \tag{16}$$

For details on the log-likelihood function, and the various link functions of the Generalized Poisson distribution, see (Yee, 2017; Consul, 1989).

2.5 Estimation

To obtain the estimates of the models' parameters, we maximized the log-likelihood functions of the Equations (3; 7; 13 and 16), concerning the parameters of proposed models. Due to the complexity of the log-likelihood functions of the Equations (7, 13 and 16) estimators are obtained via numerical methods, such as the Quasi-Newton used in *BFGS* method, implemented in the *optim* function of *software* R Team *et al.* (2013). In implementing the codes for the NB, COMP and GP models, we multiplied the log-likelihood functions by -1, since *optim* minimizes the objective function, and as initial estimates, we used the Poisson model estimates for the vector of parameters $_\iota$ and ($\tau = 1$, $\xi = 1$ and $\zeta = 1$) for the dispersion parameters of the NB, COMP and GP models. To obtain the observed information matrix, we use the option *hessian* = TRUE of the *optim* function.

A similar alternative to find the parameter estimates of the proposed models without implementing codes is to use the *glm* function for the Poisson model, and the *MASS* packages for the Negative Binomial; *COMPoissonReg* for COM-Poisson and *VGAM* for Generalized Poisson. For details, see (Ripley *et al.,* 2013; Yee, 2017; Sellers *et al.,* 2019).

2.6 Confidence Interval

Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator of the generic vector $\boldsymbol{\beta}$ of parameters for the regression models, with $\boldsymbol{\theta} = \boldsymbol{\beta}$; $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$; $\boldsymbol{\theta} = (\boldsymbol{\beta}, \nu)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \varphi)$ for Poisson models; Negative Binomial; COM-Poisson and Generalized Poisson. The Asymptotic Confidence Interval (ICa) with confidence level $(1-\alpha)100\%$ for the parameter vector $\boldsymbol{\theta}$, is given by

$$\mathrm{ICa}_{1-\alpha}(\boldsymbol{\theta}) = \left[\hat{\boldsymbol{\theta}} - z_{\alpha/2}\sqrt{[\mathbf{J}(\hat{\boldsymbol{\theta}})]^{-1}}, \hat{\boldsymbol{\theta}} + z_{\alpha/2}\sqrt{[\mathbf{J}(\hat{\boldsymbol{\theta}})]^{-1}}\right], \tag{17}$$

where $[\mathbf{J}(\hat{\boldsymbol{\theta}})]^{-1}$ is the inverse matrix of the observed information, $z_{\alpha/2}$ is the $\alpha/2$th quantile of the standard normal distribution (Santana *et al.,* 2022). The ICa in Equation 17 is recommended for large sample sizes, that is, when the distribution of $\hat{\boldsymbol{\theta}}$ approaches the normal distribution. For small or moderate sample sizes, the ICa may not provide accurate results. The *Bootstrap* Confidence Interval (ICb) can solve statistical inference problems related to sample size. There are several types of *bootstrap* confidence intervals, however, in this work we restrict them only to the non-parametric method, for details see (Efron, 1979).

Consider $\mathbf{y} = \{y_1 \ldots, y_n\}$ a random sample of the variable Y with cumulative distribution F. And let $\boldsymbol{\theta} = h(\mathrm{F})$, a vector of unknown parameters of F and $\hat{\boldsymbol{\theta}} = s(\mathbf{y})$, your estimator. Then by randomly selecting B independent samples with replacement, $\mathbf{y}^* = \{y_1^* \ldots, y_n^*\}$ of $\mathbf{y}$, we get $\{\hat{\boldsymbol{\theta}}^*\}_{b=1}^{B}$ estimates of $\hat{\boldsymbol{\theta}}$. Consequently, the distribution $\hat{\boldsymbol{\theta}}$ is obtained by the empirical distribution of $\hat{\boldsymbol{\theta}}^*$. And then, the standard error $\mathrm{ep}(\hat{\boldsymbol{\theta}})$, of $\hat{\boldsymbol{\theta}}$ can be estimated by the standard deviation of $\{\ hat\boldsymbol{\theta}^*\}_{b=1}^{B}$, (Efron & Tibshirani, 1994, p. 45; Kehler, 2018, p. 22) this is,

$$\hat{\mathrm{ep}}(\hat{\boldsymbol{\theta}}) = \left\{\frac{1}{B-1}\sum_{b=1}^{B}[\hat{\boldsymbol{\theta}}_b^* - (\hat{\boldsymbol{\theta}}_1^* + \cdots + \hat{\boldsymbol{\theta}}_B^*)/B]^2\right\}^{1/2}. \tag{18}$$

So, if the *B bootstrap* samples of $\{\hat{\boldsymbol{\theta}}^*\}_{b=1}^{B}$ have distribution close to normal the ICb (Liu & Tian, 2015, p. 206; Alves, 2013, p. 43), with confidence level $(1-\alpha)100\%$ for $\boldsymbol{\theta}$ is given by

$$\mathrm{ICb}_{1-\alpha}(\boldsymbol{\theta}) = [\hat{\boldsymbol{\theta}} - z_{\alpha/2}\hat{\mathrm{ep}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}} + z_{\alpha/2}\hat{\mathrm{ep}}(\hat{\boldsymbol{\theta}})]. \tag{19}$$

2.7 Choosing the Best Model

To choose the best model that fits the data, we used the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which are given by

$$\begin{cases} \mathrm{AIC} = -2\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) + 2\mathrm{k} \\ \\ \mathrm{BIC} = -2\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) + 2\mathrm{k}\log(n). \end{cases} \tag{20}$$

where $\ell(\hat{\boldsymbol{\theta}}; \mathbf{y})$ is the maximized log-likelihood function of the model; k is the number of parameters of the proposed model and n the number of observations in the sample. For details, see (Akaike, 1974; Schwarz, 1978). The best model will be the one that presents the lowest value for the AIC and BIC criteria (Paula, 2004).

3. Application

The variables under study correspond to confirmed cases of *tuberculosis* (all forms) for the year 2020 and the Human Development Index (HDI) for the 102 municipalities in the state of Alagoas. Data were collected by the authors in the statistical yearbook of the state of Alagoas for the year 2020 and on the *Datasus* page available at (BRASIL, 2021; de Saúde do Estado de Alagoas, 2017; *Tabnet* 2022). Figure 1 presents the map of the state of Alagoas with notifications of *tuberculosis* for its 102 municipalities.

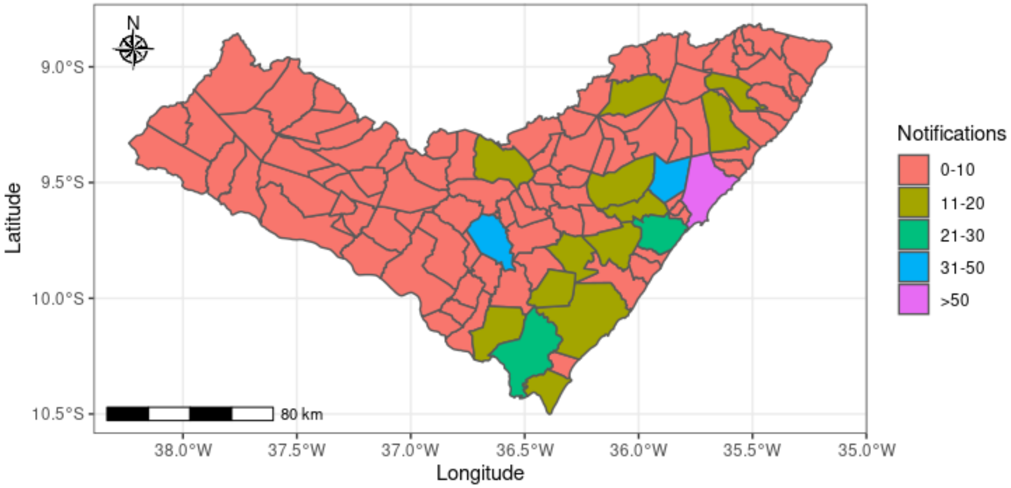

Figure 1. Map of the state of Alagoas with *tuberculosis* notifications for the year 2020.

In Figure 1, we can observe that the data are scattered and asymmetrical, for example, the class >50 corresponds to the 584 cases of *tuberculosis* for the capital Maceió, with an estimated population of 1,031,597 inhabitants and HDI of 0.721 (*IBGE* 2022).

3.1 Results and discussion

In order to better in undertand the cases of *tuberculosis*, we performed a descriptive analysis, which presented a mean of 10.83; variance 3340.38 and asymmetry coefficient 9.70 which corroborates the visual analysis of Figure 1, and the same can be observed in Figure 2.

It can be seen in the column chart in Figure 2 that the data clearly do not have normality because they are counting data, and due to the large dispersion in the data and the asymmetry, probably the Negative Binomial, COM-Poisson and Generalized Poisson will fit the data better than Poisson, as they have a dispersion parameter.

Thus, we fit the regression models of Poisson distributions; Negative Binomial; COM-Poisson and Generalized Poisson to explain the relationship between *tuberculosis* notifications and the HDI. In Table 1, we present the parameters; their maximum likelihood estimators (MLE); their asymptotic confidence intervals ICa(95%) and *bootstrap* ICb(95%) both with 95% confidence, for the confidence interval ICb(95%) 5,000 replicates were performed *bootstrap*, in addition to the (AIC) and (BIC) criteria for four models considered in the study.

We observe in Table 1 that the parameter β_1 that corresponds to the regression coefficient was significant at the level $\alpha = 5\%$ for all models, since their respective intervals of confidence ICa(95%) and ICb(95%), did not include the zero term, that is, the relationship of *tuberculosis* notifications with the HDI is significant $\alpha = 5\%$, and was identified by the four models.

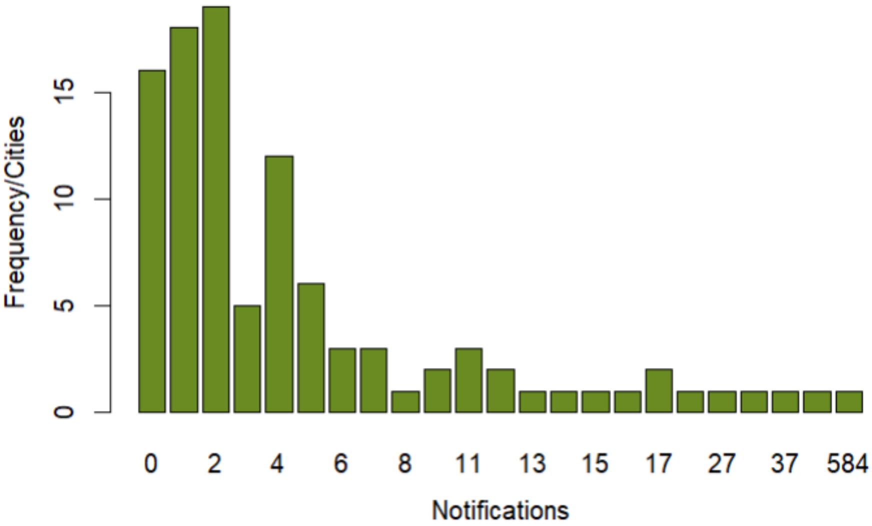

Figure 2. Column chart of *tuberculosis* notifications in the state of Alagoas for the year 2020.

Likewise, the models of the NP, COMP and GP distributions were also able to capture the overdispersion caused by the high *tuberculosis* notifications as shown in Figure 2, since the estimates of their respective dispersion parameters, presented the values: $1/\hat{\phi} = 1/1.016 \neq 0$, $\hat{\nu} = 0.046 < 1$ and $\hat{\varphi} = 0.847 > 0$. This overdispersion is also significant at the $\alpha = 5\%$ level, since their respective confidence intervals ICa(95%) and ICb(95%), did not include the zero term for ϕ and φ and term one for ν. The same did not occur for the Poisson model because it does not have a dispersion parameter, part of the overdispersion went to its mean, corroborating the eigenvalue of $\hat{\beta}_1 = 26.490$. The GP distribution presented the value for $\hat{\beta}_1$ intermediate to that of the NB and COMP distributions, in hypothesis this occurred because the GP distribution is limited by the NB distribution.

The COMP distribution regression model was the one that showed the lowest value for the estimates of the AIC and BIC criteria, followed by the NB distribution model. This fact occurred because the NB and COMP distributions have the geometric distribution as a particular case, which corroborates the values of the parameters $\hat{\phi} = 1.016 \approx 1$ for the NB model and $\hat{\nu} = 0.046 \approx 0$ and $\hat{\lambda}_i < 1$ for the COMP model. In hypothesis, the Geometric distribution regression model would fit the data better because it has a smaller number of parameters than the NB and COMP models. However, when considering the AIC and BIC criteria, we chose the COMP distribution regression model to represent the relationship between *tuberculosis* notifications and the HDI of each municipality in the state of Alagoas.

In Figure 3, we present the graph of the observed values *versus* the values predicted by the COM-Poisson distribution regression model, in order to verify the prediction quality of *tuberculosis* cases in the state of Alagoas.

The model managed to predict the data well according to the $R^2 = 0.98$ of the observed values *versus* those predicted in Figure 3. The *outlier* point corresponds to the capital Maceió, whose prediction by the model was #580 against the #584 notifications registered in that year. Thus, the COM-Poisson distribution regression model emerges as a great analysis option for health data with

Table 1. Maximum likelihood estimators (MLE), confidence intervals ICa(95%) and ICb(95%), and the Akaike (AIC) and Bayesian (BIC) Informativity Criteria for four adjusted models in the study of *tuberculosis* with the HDI

Models	Parameters	EMV	ICa(95%)	ICb(95%)	AIC	BIC
NB	β_0	-5.423	(-6.802; -4.044)	(-5.870; -4.976)	549.280	566.400
	β_1	15.668	(12.679; 18.657)	(14.070; 17.266)		
	ϕ	1.016	(0.656; 1.376)	(0.858; 1.174)		
COMP	β_0	-1.145	(-1.424; -0.866)	(-1.380; -0.910)	**531.097**	**538.972**
	β_1	2.260	(1.617; 2.901)	(1.778; 2.740)		
	ν	0.046	(0.046; 0.046)	(0.045; 0.047)		
Po	β_0	-10.737	(-11.202; -10.271)	(-11.090; -10.384)	1,056.700	1,071.200
	β_1	26.490	(25.667; 27.313)	(24.993; 27.987)		
GP	β_0	-2.440	(-4.438; -0.442)	(-2.813; -2.067)	589.705	611.454
	β_1	6.498	(2.019; 10.977)	(5.003; 7.993)		
	φ	0.847	(0.805; 0.889)	(0.784; 0.910)		

the HDI. An explanation by the model can be given as follows

$$\begin{cases} \log(\hat{\lambda}_i) = \hat{\beta}_0 + \hat{\beta}_1 \times \mathrm{IDH} \\ \qquad\qquad = -1.145 + 2.260 \times \mathrm{IDH} \\ \qquad \hat{\lambda}_i = \exp(-1.145 + 2.260 \times \mathrm{IDH}) \\ \mathrm{E}(Y_i) \approx \hat{\lambda}_i^{1/\hat{\nu}} - \frac{\hat{\nu}-1}{2\hat{\nu}} \\ \mathrm{E}(Y_i) \approx \exp(-1.145 + 2.260 \times \mathrm{IDH})^{1/0.046} + 10.44. \end{cases} \tag{21}$$

This means that for each variation in the HDI, the expected number of *tuberculosis* cases for each city in the state of Alagoas is given by the Equation (21).

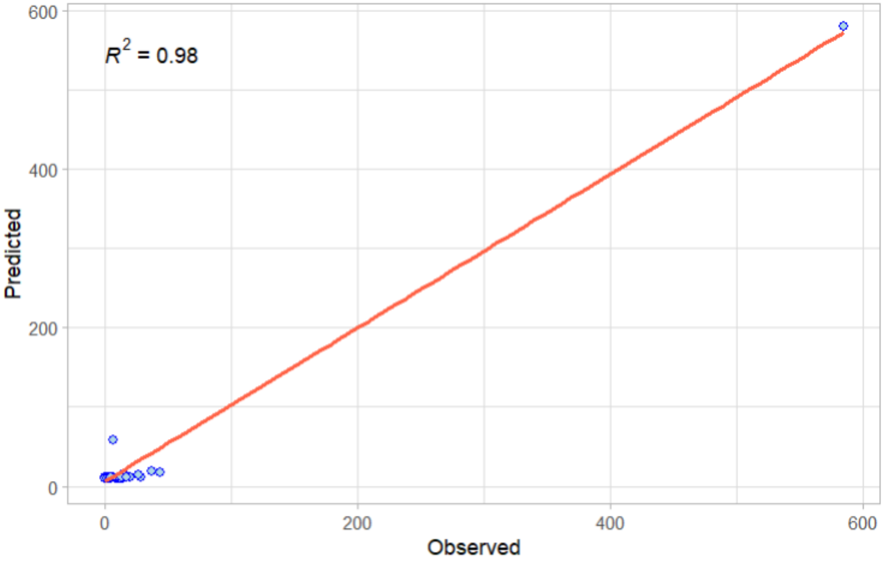

Figure 3. Predicted *versus* observed values.

4. Conclusion

In the 102 cities in the state of Alagoas, there is a relationship between *tuberculosis* notifications and the HDI human development index, which is overdispersed and significant at the probability level $\alpha = 5\%$, and can be explained by the COM-Poisson distribution regression model.

Acknowledgments

We are indebted to the editorial board and referees for their valuable comments, criticisms, and suggestions which have substantially improved the manuscript.

Conflicts of Interest

The authors have declared no conflict of interest.

References

1. Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control* **19,** 716–723 (1974).
2. Alves, E. J. *Métodos de bootstrap e aplicações em problemas biológicos* MA thesis (Universidade Estadual Paulista (Unesp), 2013).
3. Anastasopoulos, P. C. & Mannering, F. L. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention* **41,** 153–159 (2009).
4. Atkins, D. C. & Gallop, R. J. Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. *Journal of Family Psychology* **21,** 726 (2007).
5. Barlow, R. E. & Proschan, F. *Mathematical theory of reliability* (SIAM, 1996).
6. Barriga, G. D. & Louzada, F. The zero-inflated Conway–Maxwell–Poisson distribution: Bayesian inference, regression modeling and influence diagnostic. *Statistical Methodology* **21,** 23–34 (2014).

7. Bartko, J. J. *The negative binomial distribution* PhD thesis (Virginia Polytechnic Institute, 1960).
8. Boatwright, P., Borle, S., Kadane, J. B., Minka, T. P. & Shmueli, G. Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian analysis* **1,** 363–374 (2006).
9. BRASIL. *Ministério da Saúde Banco de dados do Sistema Único de Saúde - DATASUS* Acessado jun. 2021. 2021. http://www2.datasus.gov.br/DATASUS/index.php?area=0206&id=8065372&VObj=http://tabnet.datasus.gov.br/.
10. Cameron, A. C. & Trivedi, P. K. *Regression analysis of count data* (Cambridge university press, 2013).
11. Carvalho, F. J., Santana, D. G. d. & Araújo, L. B. d. Why analyze germination experiments using Generalized Linear Models? *Journal of Seed Science* **40,** 281–287 (2018).
12. Carvalho, F. J. *et al. Modelos lineares generalizados na agronomia: análise de dados binomiais e de contagem, zeros inflacionados e enfoque bayesiano.* PhD thesis (Universidade Federal de Uberlândia, 2019).
13. Consul, P. C. *Generalized Poisson distributions: properties and applications* (M. Dekker, 1989).
14. Conway, R. W. & Maxwell, W. L. A queuing model with state dependent service rates. *Journal of Industrial Engineering* **12,** 132–136 (1962).
15. De Saúde do Estado de Alagoas, S. *Anuário Estatístico do Estado de Alagoas* Acesso jun. 2021. 2017. https://dados.al.gov.br/catalogo/dataset/anuario-estatistico-do-estado-de-alagoas.
16. De Saúde Perfil dos Municípios Alagoanos, D. *Casos confirmados de doenças de notificações compulsórias de 2013 a 2020* Acesso em: 12 junho 2022. 2022. https://dados.al.gov.br/catalogo/dataset/dados-de-saude-perfil-municipal/resource/68ba2469-0cdf-453c-b750-d6fa2c854b19.
17. Desjardins, C. D. Modeling zero-inflated and overdispersed count data: An empirical study of school suspensions. *The Journal of Experimental Education* **84,** 449–472 (2016).
18. Efron, B. The 1977 RIETZ lecture. *The Annals of Statistics* **7,** 1–26 (1979).
19. Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap* (CRC press, 1994).
20. Guikema, S. D. & Goffelt, J. P. A flexible count data regression model for risk analysis. *Risk Analysis: An International Journal* **28,** 213–223 (2008).
21. Hall, D. B. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56,** 1030–1039 (2000).
22. Hayati, M, Sadik, K & Kurnia, A. *Conwey-Maxwell Poisson distribution: approach for over-and-under-dispersed count data modelling* in *IOP Conference Series: Earth and Environmental Science* **187** (2018), 012039.
23. Hilbe, J. M. *Negative binomial regression* (Cambridge University Press, 2011).
24. Hinde, J. & Demétrio, C. G. Overdispersion: models and estimation. *Computational statistics & data analysis* **27,** 151–170 (1998).
25. Huang, A. Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling* **17,** 359–380 (2017).
26. *IBGE* Disponível em: https://cidades.ibge.gov.br/brasil/al/maceio/panorama. Acesso em: 12 junho 2022. 2022.
27. Kehler, A. D. *Performance of dependent bootstrap confidence intervals for generalized Gamma means* PhD thesis (The University of Regina (Canada), 2018).
28. Khan, A., Ullah, S. & Nitz, J. Statistical modelling of falls count data with excess zeros. *Injury prevention* **17,** 266–270 (2011).

29. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34,** 1–14 (1992).

30. Liu, Y. & Tian, G.-L. Type I multivariate zero-inflated Poisson distribution with applications. *Computational Statistics & Data Analysis* **83,** 200–222 (2015).

31. Lord, D., Guikema, S. D. & Geedipally, S. R. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention* **40,** 1123–1134 (2008).

32. Nelder, J. A. & Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135,** 370–384 (1972).

33. Park, B.-J. & Lord, D. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention* **41,** 683–691 (2009).

34. Paula, G. A. *Modelos de regressão: com apoio computacional* (IME-USP São Paulo, 2004).

35. Ribeiro Junior, E. E. *Contributions to the analysis of dispersed count data* MA thesis (Universidade de São Paulo, 2019).

36. Ridout, M., Hinde, J. & Demétrio, C. G. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57,** 219–223 (2001).

37. Ridout, M. S. & Besbeas, P. An empirical model for underdispersed count data. *Statistical Modelling* **4,** 77–89 (2004).

38. Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D. & Ripley, M. B. Package 'mass'. *Cran r* **538,** 113–120 (2013).

39. Santana, R. A., Conceição, K. S., Diniz, C. A. & Andrade, M. G. Type I multivariate zero-inflated COM–Poisson regression model. *Biometrical Journal* **64,** 481–505 (2022).

40. Santana, R. A. *Modelos multivariados para dados de contagem com excesso de zeros* PhD thesis (Universidade Federal de São Carlos, 2019).

41. Schwarz, G. Estimating the dimension of a model. *The annals of statistics,* 461–464 (1978).

42. SciELO. *Scientific Electronic Library Online* Acesso mar. 2022. 2022. https://search.scielo.org/?q=Poisson&lang=pt&count=15&from=1&output=site&sort=&format=summary&page=1&where=&filter\%5Bsubject_area\%5D\%5B\%5D=Agricultural+Sciences.

43. SciELO. *Scientific Electronic Library Online* Acesso mar. 2022. 2022. https://search.scielo.org/?q=Generalized+Poisson+distribution&lang=pt&count=30&from=1&output=site&sort=&format=summary&fb=&page=1&filter\%5Bsubject_area\%5D\%5B\%5D=Health+Sciences&q=*&lang=pt&page=1.

44. Sellers, K., Lotze, T., Raim, A. & Raim, M. A. Package 'COMPoissonReg'. *Package "COMPoissonReg* (2019).

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

46. Sellers, K. F. & Shmueli, G. A flexible regression model for count data. *The Annals of Applied Statistics,* 943–961 (2010).

47. Sellers, K. F. & Shmueli, G. Data dispersion: now you see it... now you don't. *Communications in Statistics-Theory and Methods* **42,** 3134–3147 (2013).

48. Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. & Boatwright, P. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54,** 127–142 (2005).

49. *Tabnet* Disponível em: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?ibge/censo/cnv/rendaal. Acesso em: 24 agosto 2022. 2022.

50. Team, R. C. *et al.* R: A language and environment for statistical computing (2013).

51. Yee, T. VGAM: Vector generalized linear and additive models (2017).