



## Article

# Modeling Overdispersed Dengue Data via Poisson Inverse Gaussian Regression Model: A Case Study in the City of Campo Grande, MS, Brazil

Erlandson Ferreira Saraiva <sup>1,\*</sup> , Valdemiro Piedade Vigas <sup>1</sup>, Mariana Villela Flesch <sup>2</sup>, Mark Gannon <sup>3</sup> and Carlos Alberto de Bragança Pereira <sup>3</sup> 

- <sup>1</sup> Institute of Mathematics, Federal University of Mato Grosso do Sul, Campo Grande 79070-900, MS, Brazil  
<sup>2</sup> Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande 79070-900, MS, Brazil  
<sup>3</sup> Institute of Mathematics and Statistics, University of São Paulo, São Paulo 05508-090, SP, Brazil  
\* Correspondence: erlandson.saraiva@ufms.br; Tel.: +55-67-3345-7511

**Abstract:** Dengue fever is a tropical disease transmitted mainly by the female *Aedes aegypti* mosquito that affects millions of people every year. As there is still no safe and effective vaccine, currently the best way to prevent the disease is to control the proliferation of the transmitting mosquito. Since the proliferation and life cycle of the mosquito depend on environmental variables such as temperature and water availability, among others, statistical models are needed to understand the existing relationships between environmental variables and the recorded number of dengue cases and predict the number of cases for some future time interval. This prediction is of paramount importance for the establishment of control policies. In general, dengue-fever datasets contain the number of cases recorded periodically (in days, weeks, months or years). Since many dengue-fever datasets tend to be of the overdispersed, long-tail type, some common models like the Poisson regression model or negative binomial regression model are not adequate to model it. For this reason, in this paper we propose modeling a dengue-fever dataset by using a Poisson-inverse-Gaussian regression model. The main advantage of this model is that it adequately models overdispersed long-tailed data because it has a wider skewness range than the negative binomial distribution. We illustrate the application of this model in a real dataset and compare its performance to that of a negative binomial regression model.

**Keywords:** dengue fever; poisson regression model; negative binomial regression model; Poisson inverse Gaussian regression model; maximum likelihood estimation



**Citation:** Saraiva, E.F.; Vigas, V.P.; Flesch, M.V.; Gannon, M.; de Bragança Pereira, C.A. Modeling Overdispersed Dengue Data via Poisson Inverse Gaussian Regression Model: A Case Study in the City of Campo Grande, MS, Brazil. *Entropy* **2022**, *24*, 1256. <https://doi.org/10.3390/e24091256>

Academic Editors: Philip Broadbridge and Udo Von Toussaint

Received: 14 July 2022

Accepted: 4 September 2022

Published: 7 September 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

According to the world health organization (WHO), dengue fever is a mosquito-borne viral infection that is common in warm, tropical climates. The female of the *Aedes aegypti* mosquito is the main transmitter of the disease, which is caused by four serotypes of a flavivirus, called DENV1, DENV2, DENV3 and DENV4, classified on biological and immunological criteria. As there is still no safe and effective vaccine, the most effective ways to prevent outbreaks of the disease are still to avoid mosquito bites and control the mosquito population [1].

Since the proliferation of the mosquito that transmits dengue depends on temperature, water availability, and some other climatic factors to complete its cycle life, it is of interest to understand the relationships between climatic variables and the recorded number of dengue cases. The Poisson regression (PR) model has been used repeatedly for such applications. For example, Leslie [2] studies the climatic factors that affect the spread of dengue in the city of Colombo, Sri Lanka from the period of 2010 to 2018, using as a primary model a Poisson regression model. Sinaga and Sinulingga [3] model the number of dengue

hemorrhagic fever cases in the city of Medan using a Poisson regression model. The authors consider as explanatory variables population density, number of health workers, number of health facilities, area height, and average waste production. Mukhaiyar et al. [4] propose to predict the number of dengue fever cases in Bandung, West Java, Indonesia, in the period 2001–2016, by fitting a Poisson regression model using the temperature and cumulative rainfall as explanatory variables. In these approaches, the observed values for a response variable are taken as having been generated from a Poisson distribution. Using the theory of generalized linear models [5], a log-linear relationship is constructed relating the average value of the response variable to a set of  $p$  explanatory variables.

An essential assumption of the Poisson regression model is that the mean of the response variable is equal to the variance, a property known as equidispersion. However, dengue-fever data, in general, do not have this property. Therefore, the Poisson regression model is not suitable for modeling such data, because the standard errors may be underestimated, leading to misleading inference from the regression.

For the scenario of overdispersed data, i.e., the variance of the response variable being greater than its average, the usual statistical approach consists of considering a negative binomial regression (NBR) model. Under this approach, it is assumed that the response variable values are generated according to a negative binomial distribution. This distribution is a mixture of a Poisson distribution and a Gamma distribution. Analogously to the PR model, this approach also links the response variable's average value to a set of  $p$  explanatory variables by using a log-linear relationship. However, the NBR model is not adequate to model long-tailed datasets, i.e., datasets in which there are some very large integer values far away from the majority [6–8].

Therefore, we propose modeling a dengue-fever dataset by using the Poisson-inverse-Gaussian regression (PIGR) model as a competitor to the NBR model. In this model, response-variable values are assumed to be generated according to a Poisson-inverse-Gaussian distribution. This distribution is a mixture of a Poisson distribution and an inverse-Gaussian distribution. The main advantage of this distribution is that it may properly model overdispersed long-tail data because it has a larger range of skewness than a negative binomial distribution [9–12]. For this model, we also link the expected value of the response variable to a set of  $p$  explanatory variables by using a log-linear relationship.

We illustrate the fitting of the NBR and PIGR models to a real data set  $\mathbb{D}$ , referring to the number of cases of dengue fever recorded in the city of Campo Grande, Mato Grosso do Sul state, Brazil, in the period from January 2008 to December 2019. The dataset  $\mathbb{D}$  is an excel sheet composed of 144 lines and 6 columns. The first column contains the recorded number of dengue-fever cases in each of the 144 months considered in the study. Columns 2 to 6 contain the recorded values for the following explanatory variables: month, the average temperature in the month, the average humidity in the month, the number of rainy days in the month, and rainfall in the month.

To estimate the model parameters, we adopt the maximum-likelihood method. Since the maximum-likelihood estimators do not have explicit mathematical solutions, we obtain the estimates numerically by using the R software [13] and the command `gamlss` of the Generalized Additive Model for Location, Scale and Shape (GAMLSS) package [14]. According to Stasinopoulos et al. [15], “the GAMLSS were introduced by Rigby and Stasinopoulos (2001, 2005) [14,16] and Akantziliotou et al. (2002) [17] as a way of overcoming some of the limitations associated with Generalized Linear Models (GLM) and Generalized Additive Models (GAM)”. The two main advantages of a GAMLSS model are: (i) it assumes that the response (dependent) variable may follow any parametric distribution and not just distributions belonging to an exponential family, and (ii) all the parameters of the probability distribution of the response variable can be modelled as functions of the available explanatory variables. More details on GAMLSS package may be found in its manual available on the website <http://www.gamlss.com/wp-content/uploads/2013/01/gamlss-manual.pdf> (accessed on 15 March 2022).

We also compare the performance of the NBR and PIGR models by using the Akaike Information criterion [18,19], denoted by AIC, and the Bayesian Information criterion [20], denoted by BIC, and the Root Mean Square Error (RMSE). We also fit both models by considering a P-spline term for the month variable since it has cyclical values, and smooth terms for continuous variables. For this, we use the `pb()` and `pbc()` functions inside the `gam1ss` function. Based on the AIC, BIC and RMSE values, the PIGR model was considered the best model. We also present the quantile-quantile normal plot and worm plot for the randomized quantile residuals [21] generated from the NBR and PIGR fitted models. Both graphs also show PIGR performing better than NBR.

The three main advantages of the proposed modeling are: (i) present better performance in relation to the usual approaches, which are based on the fitting of PR and NBR models; (ii) the fitted model shows that every year a peak will occur, and that the only way to avoid this peak is by the implementation of actions to combat the proliferation of the transmitting mosquito; and (iii) the fitted model shows in which the months of the year combat actions must be implemented.

The remainder of the paper is organized as follows. In Section 2, we describe the PR, NBR and PIGR models and present the estimation procedure. Section 3 presents the main results, including the comparison of the NBR and PIGR models and the residual analysis. Section 4 presents the final remarks.

## 2. Statistical Modeling

Let  $\mathbf{y} = (y_1, \dots, y_n)$  be a vector of data composed of the number of dengue-fever cases recorded in a period of  $n$  months in a country, state, or city. Assume that recorded value  $y_t$  is a realization of the random variable  $Y_t$ , for  $Y_t \in \mathcal{Y} = \{0, 1, 2, 3, \dots\}$ .

In addition, assume that measurements of  $p$  explanatory variables are available, denoted by  $X_1, \dots, X_p$ , that can be associated with mosquito reproduction and dengue transmission, and consequently also associated with the number of recorded cases of dengue. Consider  $\mathbf{x}$  to be an  $n \times (p + 1)$  matrix in which the first column contains only values 1 and columns 2 to  $p + 1$  are composed of the recorded measurements for variables  $X_1$  to  $X_p$ , respectively. Denote the  $t^{\text{th}}$  line of  $\mathbf{x}$  by  $\mathbf{x}_t = (1, x_{t1}, \dots, x_{tp})$ , for  $t = 1, \dots, n$ .

### 2.1. Poisson Regression Model

Since random variable  $Y_t$  is a discrete variable that counts the number of cases in a time period of one month  $t$ , it is usual to assume that  $Y_t$  follows a Poisson distribution with parameter  $\mu_t$ , i.e.,

$$Y_t \sim \text{Poisson}(\mu_t),$$

where,  $\mu_t = \mathbb{E}(Y_t)$  is the expected value of  $Y_t$ , with  $\mu_t > 0$ ,  $t = 1, \dots, n$ . Its probability mass function is given by

$$P(Y_t = y_t | \mu_t) = \frac{\mu_t^{y_t} e^{-\mu_t}}{y_t!},$$

for  $y_t \in \mathcal{Y}$  and  $t = 1, \dots, n$ .

Using the theory of generalized linear models [5], we can link the expected value of  $Y_t$  to explanatory variables  $\mathbf{x}$  through the following log-linear relationship:

$$\eta(\mu_t) = \log(\mu_t) = \boldsymbol{\beta} \mathbf{x}_t = \beta_0 + \sum_{j=1}^p \beta_j x_{tj}, \quad (1)$$

where  $\eta(\mu_t)$  is the linear predictor,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$  is the vector of parameters of the model and  $\mathbf{x}_t$  is the  $t$ -th line of the matrix  $\mathbf{x}$ , for  $t = 1, \dots, n$ .

Given  $(\mathbf{y}, \mathbf{x})$  the log-likelihood function for parameters  $\boldsymbol{\beta}$  is given by

$$l(\boldsymbol{\beta} | \mathbf{y}, \mathbf{x}) \propto \sum_{t=1}^n y_t \boldsymbol{\beta} \mathbf{x}_t - \exp\{\boldsymbol{\beta} \mathbf{x}_t\}.$$

In order to get the maximum-likelihood estimates for the parameters  $\beta$ , we first need to determine the first-order partial derivatives of the log-likelihood function, which are given by

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{t=1}^n x_{tj} \left( y_t - \exp \left\{ \beta_0 + \sum_{j=1}^p \beta_j x_{tj} \right\} \right), \quad (2)$$

for  $j = 0, \dots, p$ .

The maximum-likelihood estimates are the solutions of equations in (2) when they are set to 0,  $\frac{\partial l(\beta)}{\partial \beta_j} = 0$ , for  $j = 0, \dots, p$ . However, these equations do not have explicit analytic solutions. Therefore, we apply numerical methods to solve these equations. We can obtain the maximum-likelihood estimates  $\hat{\beta}$  of the parameters  $\beta$  using the R software [13] and the function `glm()` [22].

Although the Poisson distribution is a natural choice for modeling the number of dengue-fever cases recorded in a month, this distribution has the restriction that the expected value is equal to the variance,  $\mathbb{E}(Y_t) = \text{Var}(Y_t)$ , for  $t = 1, \dots, n$ . Thus, before considering a Poisson regression model it is essential to check if recorded data present some evidence for overdispersion or underdispersion.

Hinde and Demétrio [23] propose to check the evidence for overdispersion or underdispersion by using the index

$$\mathbb{IS} = \frac{S_y^2 - \bar{y}}{\bar{y}}, \quad (3)$$

where  $S_y^2$  and  $\bar{y}$  are the sampled variance and mean of the recorded values for  $Y$ , respectively. The decision is based on the following interpretation: If  $\mathbb{IS} = 0$ , the recorded data indicate equidispersion, and the Poisson regression model can be used. On the other hand, if  $\mathbb{IS} < 0$ , the recorded data indicate underdispersion, and if  $\mathbb{IS} > 0$ , the recorded data indicate overdispersion. The Poisson regression model is not appropriate for nonzero  $\mathbb{IS}$ .

Cameron and Trivedi [24] propose to check for evidence of overdispersion using a hypothesis test. To do this, the authors assume that  $\text{Var}(Y) = \mu + \lambda\mu^2$ , and specify the following statistical hypotheses  $H_0 : \lambda = 0$  versus  $H_1 : \lambda > 0$ . The test statistic is calculated according to the following four steps:

- (i) Fitting a Poisson regression model;
- (ii) Calculating the fitted values  $\hat{\mu}_t$ , for  $t = 1, \dots, n$ ;
- (iii) Calculating the auxiliary values

$$Y_t^* = \frac{(y_t - \hat{\mu}_t)^2 - y_t}{\hat{\mu}_t}, \quad \text{for } t = 1, \dots, n;$$

- (iv) Fitting of an auxiliary linear model  $Y_t^* = \lambda \hat{\mu}_t + \varepsilon_t$ , where  $\varepsilon_t$  is a random error, for  $t = 1, \dots, n$ .

According to Cameron and Trivedi [24], the t-statistic for  $\lambda$  is asymptotically normal under the null hypothesis of no overdispersion. The null hypothesis is rejected whenever the p-value associated with the calculated statistic is smaller than a significance level  $\alpha$ , with  $0 < \alpha < 1$ . This overdispersion test may be performed in the R software using the `overdisp()` function of the `overdisp` package [25]. For overdispersed data, an alternative is to consider the negative binomial regression model.

## 2.2. Negative Binomial Regression Model

Assume  $Y_t$  follows the negative binomial distribution with parameters  $\mu_t$  and  $\nu$ ,

$$Y_t \sim \text{NB}(\mu_t, \nu),$$

for  $\mu_t > 0, \nu > 0, Y_t \in \mathcal{Y}$  and  $t = 1, \dots, n$ .

According to [24], the negative binomial distribution that accomodates overdispersion in the data has the following probability mass function:

$$P(Y_t = y_t | \mu_t, \nu) = \frac{\Gamma(y_t + \nu^{-1})}{\Gamma(\nu^{-1})\Gamma(y_t + 1)} \left(\frac{\nu^{-1}}{\nu^{-1} + \mu_t}\right)^{\nu^{-1}} \left(\frac{\mu_t}{\nu^{-1} + \mu_t}\right)^{y_t}$$

where  $\Gamma(\cdot)$  is the gamma function. The expected value and variance of  $Y_t$  are given by  $\mathbb{E}(Y_t) = \mu_t$  and  $\text{Var}(Y_t) = \mu_t + \nu\mu_t^2$ , respectively, for  $t = 1, \dots, n$ .

As it is in the PR model, the expected value for  $Y_t$  in the NBR model is linked to the explanatory variables  $\mathbf{X}$  via a function of the form given in expression (1). The log-likelihood function for the parameters  $(\beta, \nu)$  is

$$l(\beta, \nu | \mathbf{y}, \mathbf{x}) \propto \sum_{t=1}^n \mathbb{A}_t(\nu) - \frac{1}{\nu} \log(1 + \nu \exp\{\beta \mathbf{x}_t\}) + y_t \beta \mathbf{x}_t - y_t \log(1 + \nu \exp\{\beta \mathbf{x}_t\}) + y_t \log(\nu),$$

where  $\mathbb{A}_t(\nu) = \log\left(\Gamma\left(y_t + \frac{1}{\nu}\right)\right) - \log\left(\Gamma\left(\frac{1}{\nu}\right)\right)$ , for  $t = 1, \dots, n$ .

The maximum-likelihood estimates are obtained by determining the first-order partial derivatives of the log-likelihood function, then equating them to zero:

$$\begin{aligned} \frac{\partial l(\beta, \nu)}{\partial \beta_j} &= \sum_{t=1}^n x_{t(j+1)} \left( \frac{y_t - \exp\{\beta \mathbf{x}_t\}}{1 + \nu \exp\{\beta \mathbf{x}_t\}} \right) = 0; \\ \frac{\partial l(\beta, \nu)}{\partial \nu} &= \sum_{t=1}^n \left[ \frac{\partial \mathbb{A}_t(\nu)}{\partial \nu} + \frac{y_t}{\nu} - \left( y_t - \frac{1}{\nu} \right) \frac{\exp\{\beta \mathbf{x}_t\}}{1 + \nu \exp\{\beta \mathbf{x}_t\}} - \frac{1}{\nu^2} \log(1 + \nu \exp\{\beta \mathbf{x}_t\}) \right] = 0. \end{aligned}$$

for  $j = 0, \dots, p$ .

These equations also do not have explicit solutions. Analogously to the case of the Poisson regression model, we obtain the maximum-likelihood estimates  $(\hat{\beta}, \hat{\nu})$  of the parameters  $(\beta, \nu)$  using the R software, but for this case we use the `gamlss()` function from the `gamlss` package [26] with the option `family=NBI`.

### 2.3. Poisson-Inverse-Gaussian Regression Model

As an alternative to the negative binomial model, consider that  $Y_t$  follows the Poisson-inverse Gaussian distribution with parameters  $\mu_t$  and  $\tau$ , i.e.,

$$Y_t \sim \text{PIG}(\mu_t, \tau),$$

for  $t = 1, \dots, n$ . This distribution is a mixture of a Poisson distribution and an inverse Gaussian distribution. Let  $Y_t | V$  follow a Poisson distribution with mean  $\mu_t V$ , where  $V$  follows an Inverse Gaussian distribution with mean equal to 1 and dispersion parameter  $1/\tau$  [8]. The marginal probability mass function for  $Y_t$  is

$$P(Y_t = y_t | \mu_t, \tau) = \frac{\mu_t^{y_t}}{y_t!} \left(\frac{2}{\pi\tau}\right)^{0.5} \exp\{1/\tau\} (1 + 2\tau\mu_t)^{-S_t/2} \mathbb{K}_{S_t}(\Psi_t),$$

where  $S_t = y_t - \frac{1}{2}$ ,  $\Psi_t = \frac{\sqrt{1+2\tau\mu_t}}{\tau}$  and  $\mathbb{K}_{S_t}(\Psi_t)$  is the modified Bessel function of second kind [11], for  $t = 1, \dots, n$ .

Considering the link function given in Equation (1), the log-likelihood function for parameters  $(\beta, \tau)$  is given by

$$l(\beta, \tau | \mathbf{y}, \mathbf{x}) \propto \sum_{t=1}^n y_t \beta \mathbf{x}_t - \frac{\log(\tau)}{2} + \frac{1}{\tau} - \frac{S}{2} \log(1 + 2\tau \exp\{\beta \mathbf{x}_t\}) + \log(\mathbb{K}_{S_t}(\Psi_t)).$$

Setting the first-order partial derivatives of the log-likelihood function equal to zero, we obtain

$$\frac{\partial l(\boldsymbol{\beta}, \tau | \mathbf{y}, \mathbf{x})}{\partial \beta_j} = \sum_{t=1}^n \left[ x_{tj} \left( y_t - \frac{\tau S_t \exp\{\boldsymbol{\beta} \mathbf{x}_t\}}{1 + 2\tau \exp\{\boldsymbol{\beta} \mathbf{x}_t\}} \right) + \frac{\partial \log(\mathbb{K}(\Psi_t))}{\partial \beta_j} \right] = 0;$$

$$\frac{\partial l(\boldsymbol{\beta}, \tau | \mathbf{y}, \mathbf{x})}{\partial \tau} = -\frac{n}{\tau^2} - \frac{n}{2\tau} - \sum_{t=1}^n \left( \frac{S_t \exp\{\boldsymbol{\beta} \mathbf{x}_t\}}{1 + 2\tau \exp\{\boldsymbol{\beta} \mathbf{x}_t\}} + \frac{\partial \log(\mathbb{K}(\Psi_t))}{\partial \tau} \right) = 0,$$

for  $j = 0, \dots, p$ .

Since the maximum-log-likelihood equations are nonlinear, they cannot be solved analytically. Therefore, we obtain the maximum-likelihood estimates  $(\hat{\boldsymbol{\beta}}, \hat{\tau})$  of the parameters  $(\boldsymbol{\beta}, \tau)$  using the R software and the `gamlss` package's `gamlss()` function, with the option `family = PIG`.

### Simulation Study for PIGR Model

Since the main focus of this article is to describe the performance of the PIGR model, in this section we present a simulation study that illustrates the performance of this model. For this purpose, we generated values  $Y_t$  from a PIG distribution with parameters  $\mu_t = \exp(\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t})$  and  $\tau = 1$ , for  $t = 1, \dots, n$ . The sample sizes considered were  $n = \{50, 100, 150, 200\}$ . We set  $\beta_0 = 1.5$ ,  $\beta_1 = 1.5$  and  $\beta_2 = -1$  and generate values for covariates  $X_1$  and  $X_2$  from the following normal distributions,  $X_{1t} \sim \mathcal{N}(0, 1)$  and  $X_{2t} \sim \mathcal{N}(4, 1)$ , for  $t = 1, \dots, n$ .

In order to verify the frequentist properties of the maximum-likelihood estimator (MLE)  $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\tau})$  for the parameters of the PIGR model, we generate  $B = 1000$  different artificial datasets for each sample size  $n$  and summarize the results in terms of the average of estimates, bias, and mean square error (MSE). Table 1 shows these values for each of the parameters. As one can note, as sample size increases, there is a reduction in the bias and MSE values. These results show us empirically that there is no reason for doubting that the ML estimator  $\hat{\theta}$  satisfies the asymptotic properties of MLEs [27], i.e.,  $\hat{\theta}$  is asymptotically consistent, unbiased, and is approximately a normal random variable.

**Table 1.** Average of estimates, bias and mean square error (MSE) values.

Parameter	$n = 50$			Parameter	$n = 100$		
	Average	BIAS	MSE		Average	BIAS	MSE
$\beta_0$	1.8198	0.3198	14.1991	$\beta_0$	1.4642	-0.0357	0.2001
$\beta_1$	-1.9400	-0.4400	7.9899	$\beta_1$	-1.5418	-0.0418	0.2140
$\beta_2$	-0.6580	-0.1580	1.3534	$\beta_2$	-0.5136	-0.0136	0.0192
$\tau$	1.2681	0.2681	0.4752	$\tau$	1.1292	0.1292	0.2048
Parameter	$n = 150$			Parameter	$n = 200$		
	Average	BIAS	MSE		Average	BIAS	MSE
$\beta_0$	1.4623	-0.0376	0.0896	$\beta_0$	1.4541	-0.0458	0.0579
$\beta_1$	-1.5777	-0.0177	0.0978	$\beta_1$	-1.4972	0.0027	0.0559
$\beta_2$	-0.5036	-0.0036	0.0082	$\beta_2$	-0.4982	0.0017	0.0055
$\tau$	1.0774	0.0774	0.0843	$\tau$	1.641	0.0641	0.0594

### 3. Application

In this section, we apply the PR, NBR and PIGR models to a real data set containing the number of dengue-fever cases recorded in the city of Campo Grande, state of Mato Grosso do Sul, Brazil, in the period from January 2008 ( $t = 1$ ) to December 2019 ( $t = 144$ ).

The city of Campo Grande is located in the transition zone between a humid mesothermal climate without drought and a humid tropical climate, with a rainy season in the summer and a dry season in the winter. The city has its climate controlled by three characteristic air masses: the Atlantic Polar Mass, coming from the south, the Continental



Equatorial Mass, coming from the north and the Continental Tropical Mass, which forms in the lower Chaco region. The rainy season runs from October to March. The average total annual precipitation is 1225 mm. The relative humidity of the air presents values close to 80% from December to February. From March onwards, relative humidity shows a gradual decline, reaching its minimum value of approximately 60% in August. From August onwards, the relative humidity of the air rises again. The average maximum temperature is around 25 °C in the period from October to March [28].

Due to the favorable climate for the proliferation of the dengue-transmitting mosquito, especially, between October and March, the city has a large number of dengue cases recorded every year. A dengue-control strategy implemented by the city government is based on the availability of health agents in city neighborhoods to provide information on dengue and how to eliminate the transmitter mosquito. Additionally, the city government has a program for cleaning neighborhoods to eliminate possible breeding sites of the dengue-transmitting mosquito.

Thus, in order to contribute to the dengue surveillance system in the city of Campo Grande—MS, this article proposes the fitting of a statistical model to identify the climatic variables that can influence the number of dengue cases. Once the variables are identified, the fitted model allows projections to and simulation of different scenarios of evolution of the number of cases of the disease. Therefore, it can help in decision-making regarding the implementation of measures to combat and/or control the vector that transmits the disease.

### Results

Consider  $\mathbf{y} = (y_1, \dots, y_n)$  to be the number of dengue-fever cases recorded in the city of Campo Grande, MS state, Brazil, in the period from January 2008 ( $t = 1$ ) to December 2019 ( $t = 164$ ). These measures are freely available on the website <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinanet/cnv/denguebbr.def> (accessed on 10 November 2020) and also can be obtained by emailing the authors of the present article.

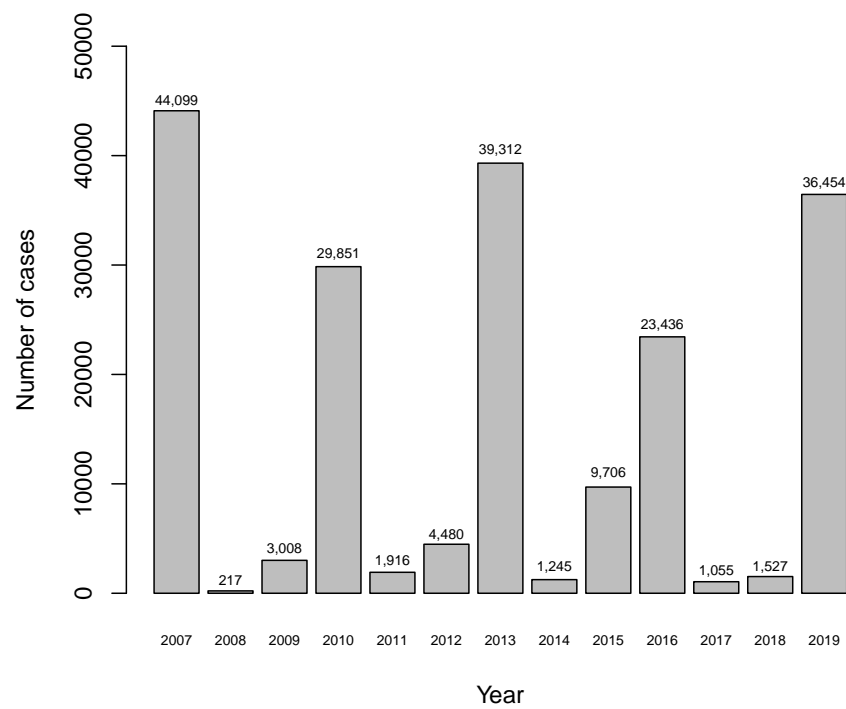
Let  $\mathbf{x}$  be a matrix of dimension  $n \times 5$  composed of the recorded measures of the variables

- $X_1$  : Month of the year, coded from 1 to 12;
  - $X_2$  : average temperature in the month;
  - $X_3$  : average humidity in the month;
  - $X_4$  : number of rainy days in the month;
  - $X_5$  : rainfall in the month.
- (4)

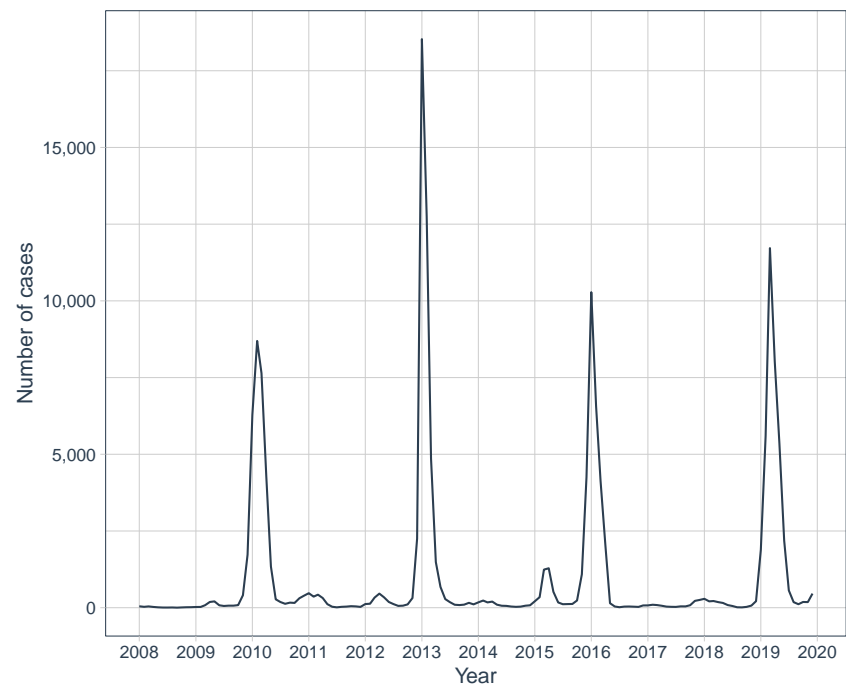
The recorded measures for variables  $X_2$  to  $X_5$  are freely available at <https://www.cemtec.ms.gov.br> (accessed on 8 December 2020). Denote this dataset by  $\mathbb{D} = (\mathbf{y}, \mathbf{x})$ , which is a matrix of dimension  $n \times 6$ . The first column contains the recorded number of dengue-fever cases in each of the 144 months considered in the study. Columns 2 to 6 contain the recorded values of the explanatory variables  $X_1$  to  $X_5$ .

Figure 1 shows the number of recorded dengue-fever cases from 2007 to 2019. The figure includes the number of cases recorded in 2007 just to show that every three years the city of Campo Grande presents a larger outbreak of dengue-fever cases. However, the recorded number of dengue-fever cases in 2007 was not considered to fit the models because the website <https://www.cemtec.ms.gov.br> (accessed on 8 December 2020) does not include values for the explanatory variables  $X_2$  to  $X_5$  in 2007.

Figure 2 shows the evolution of the number of dengue-fever cases by month. Due mainly to the climate of the city, characterized by high heat and heavy rains from October to March, this period contains most of the recorded dengue-fever cases in the city. This fact shows the importance of having a model for projection for the number of dengue cases from environmental variables, to support actions to combat the proliferation of the mosquito and consequently the reduction of the number of cases.



**Figure 1.** Number of recorded dengue-fever cases by year from 2007 to 2019.



**Figure 2.** Evolution of the number of dengue-fever cases by month in the period considered (January 2008 to December 2019).

Table 2 shows the descriptive statistics of the recorded  $y$  values in the period from January of 2008 to December of 2019. The smallest recorded value was 2 cases in August of 2008. The highest recorded value was 18,530 cases in January of 2013. On average, 1057 cases were recorded per month in the period considered.



**Table 2.** Descriptive statistics of the recorded numbers of dengue-fever cases.

Minimum	1st Quartile	Median	Average	3rd Quartile	Maximum	Variance
2	45.5	124.5	1057	340.2	18,530	7,269,590

Table 3 shows the correlations for each pair of variables. As one can note, the highest correlation is between variables  $X_3$  and  $X_5$ . However, since it is not a strong correlation ( $>0.75$ ), we opt to maintain both variables for the fitting of the models.

**Table 3.** Correlations.

Variables	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
$X_1$	1.00	0.09	−0.11	0.17	−0.39
$X_2$	0.09	1.00	0.33	0.09	0.20
$X_3$	−0.11	0.33	1.00	−0.14	0.59
$X_4$	0.17	0.09	−0.14	1.00	−0.19
$X_5$	−0.39	0.20	0.59	−0.19	1.00

In addition, we also verify if there is multicollinearity among explanatory variables by means of variance inflation factor (VIF) values for the PR and NBR models [29]. At this point, we remind the reader that multicollinearity occurs when two or more explanatory variables are highly correlated with one another in a regression model. That is, one explanatory variable can be predicted from another explanatory variable. A VIF value equal to 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. In general, values smaller than 5 indicate weak correlation, values between 5 and 10 indicate moderate correlation, and values equal to or greater than 10 indicate high correlation.

In order to calculate the VIF values, we first fit the PR and NBR models using the R software and the `glm` function. We then obtain the VIF values by applying the `vif` function of the `car` package. Listing 1 shows the R code used. The VIF values are presented in Table 4. As one can see, all values are less than five, which indicates weak multicollinearity. Therefore, all five explanatory variables are used to fit the models.

```

1 ##### Package car
2 library(car)
3 ##### Dataset
4 D <- data.frame(X1, X2, X3, X4, X5, Y)
5 ##### VIF for PR model
6 PR.model <- glm(YD ~ 1 + X1 + X2 + X3 + X4 + X5, data=D, family=poisson)
7   vif(PR.model)
8 ##### VIF from the NBR model
9 NBR.model <- glm.nb(YD ~ 1 + X1 + X2 + X3 + X4 + X5, data=D)
10  vif(NBR.model)
    
```

**Listing 1.** R code.

**Table 4.** VIF values.

Model	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
PR	1.4549	1.1254	2.2633	1.5898	2.5500
NBR	1.1998	1.5677	3.2224	2.5796	4.4537

Using the sample average and sample variance presented in Table 2, the overdispersion index given in Expression (3) is  $\mathbb{I}\mathbb{S} = \frac{7269590-1057}{1057} = 6876.614$ . That is, the recorded values are overdispersed. Additionally, we also apply the overdispersion test of Cameron and Trivedi [24] (CT test), using the `overdisp()` function of the R software. Figure 3 shows

the output of the test in the R software. As one can note, the null hypothesis is rejected for the usual significance levels  $\alpha = \{0.10, 0.05, 0.01\}$ , meaning that there is evidence for overdispersion.

#### Overdispersion Test - Cameron & Trivedi (1990)

```
data: DA
Lambda t test score: = 2.9975, p-value = 0.003211
alternative hypothesis: overdispersion if lambda p-value is less
than or equal to the stipulated significance level
```

**Figure 3.** Outputs of the CT test for overdispersion using the `overdisp()` function.

Both results described above indicate that the PR model is not appropriate for this dataset. Due to this, hereafter we fit the NBR and PIGR models to the dataset and compare these two models according to the AIC and BIC model-selection criteria. The best model is the one that has the smallest AIC and BIC values.

We fit NBR and PIGR models using the `gamlss()` function of the `gamlss` package of the R software. Since the month variable has cyclical values, we fit both models by considering a cyclical P-spline term for this variable. For this, we use the `pb()` function inside the `gamlss` function. In addition, we fit both models by considering smooth terms for continuous variables  $X_2$ ,  $X_3$  and  $X_5$ . For this case, we use the `pb()` function. We call the models fitted with `pb()` function of NBR-S and PIGR-S, respectively. Listing 2 shows the R code used for fitting the models.

```
1 ##### Dataset
2     D <- data.frame(X1, X2, X3, X4, X5, Y)
3
4 ##### Fit of the NBR model
5 NBR <- gamlss(Y ~ pbc(X1) + X2 + X3 + X4 + X5, data=D, family=NBI)
6     summary(NBR.model)
7     AIC(NBR.model)
8     BIC(NBR.model)
9
10 NBR.S <- gamlss(Y ~ pbc(X1) + pb(X2) + pb(X3) + X4 + pb(X5), data=D, family=
11     NBI)
12     summary(NBR.model.S)
13     AIC(NBR.model.S)
14     BIC(NBR.model.S)
15
16 ##### Fit of the PIGR model
17 PIGR <- gamlss(Y ~ pbc(X1) + X2 + X3 + X4 + X5, data=D, family=PIG)
18     summary(PIGR.model)
19     AIC(PIGR.model)
20     BIC(PIGR.model)
21
22 PIGR.S <- gamlss(Y ~ pbc(X1) + pb(X2) + pb(X3) + X4 + pb(X5), data=D, family=
23     PIG)
24     summary(PIGR.model.S)
25     AIC(PIGR.model.S)
26     BIC(PIGR.model.S)
```

**Listing 2.** R code.

To significance level  $\alpha = 0.10$ , none of the variables was significant for the NBR model ( $p$ -value  $> \alpha$ ). For NBR-S and PIGR models, variables  $X_4$  and  $X_5$  were not significant ( $p$ -values  $> \alpha$ ). For the PIGR-S model,  $\beta_0$  and the variables  $X_4$  and  $X_5$  were not significant ( $p$ -values  $> \alpha$ ). Due to this, we discard the NBR model and refit NBR-S, PIGR, and PIGR-S models without the non-significant variables.

Table 5 shows model-comparison criteria for the three fitted models. The smallest values are highlighted in boldface. Since the AIC and BIC values for the PIGR and PIGR-S models are very similar and the RMSE values are equal, we opt to maintain the PIGR as

the best model because the smooth terms have not led to a significant improvement in the model.

**Table 5.** Model-comparison criteria.

Model	AIC	BIC	RMSE
NBR-S	2044	2059	2885
PIGR	2001	2016	<b>2560</b>
PIGR-S	<b>1999</b>	<b>2011</b>	<b>2560</b>

With the models fitted, it is important to perform a residuals analysis in order to identify the discrepancies between the models and the data, and to assess the overall model goodness-of-fit. In a normal linear regression scenario, the Pearson and deviance residuals are usually considered. However, these residuals are not suitable for problems in which the response variable is discrete because they are not normally distributed, and according to Feng et al. [30], “have nearly parallel curves according to the distinct discrete response values, imposing great challenges for visual inspection”. To circumvent this issue, Dunn and Smyth [21] propose the use of randomized quantile residuals (RQR). According to the authors, this kind of residuals is particularly ideal for visualizing the goodness-of-fit of count regression models.

In order to calculate the RQR, we first need to obtain the cumulative distribution function,  $F(y_t|\hat{\mu}_t, \hat{\tau})$  of the model considered, for  $t = 1, \dots, n$ . For the continuous case,  $F(\cdot)$  values are uniformly distributed on interval  $(0,1)$ , and the RQR is defined as  $r_t = \Phi^{-1}(F(y_t|\hat{\mu}_t, \hat{\tau}))$ , where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution. However, since the cumulative distribution function  $F(\cdot)$  for the models considered (NBR and PIGR) is not strictly continuous, but a step function, a randomization is introduced to produce continuous normal residuals. Thus, in order to get the RQR, Dunn and Smyth [21] propose the following strategy. For  $t = 1, \dots, n$ :

- Determine a point  $a_t = \lim_{y \uparrow y_t} F(y_t|\hat{\mu}_t, \hat{\tau})$ , i.e.,  $a_t$  is the value of  $F(\cdot)$  when approaching  $y_t$  from the left;
- Determine  $b_t = F(y_t|\hat{\mu}_t, \hat{\tau})$ , i.e., the value of  $F(\cdot)$  at the point  $y_t$ ;
- Generate a value  $u_t$  from a uniform distribution on interval  $(a_t, b_t]$ ;
- Calculate the RQR  $\hat{r}_t = \Phi^{-1}(u_t)$ .

We obtained the RQR values for NBR and PIGR models using the `residuals` function of the R software.

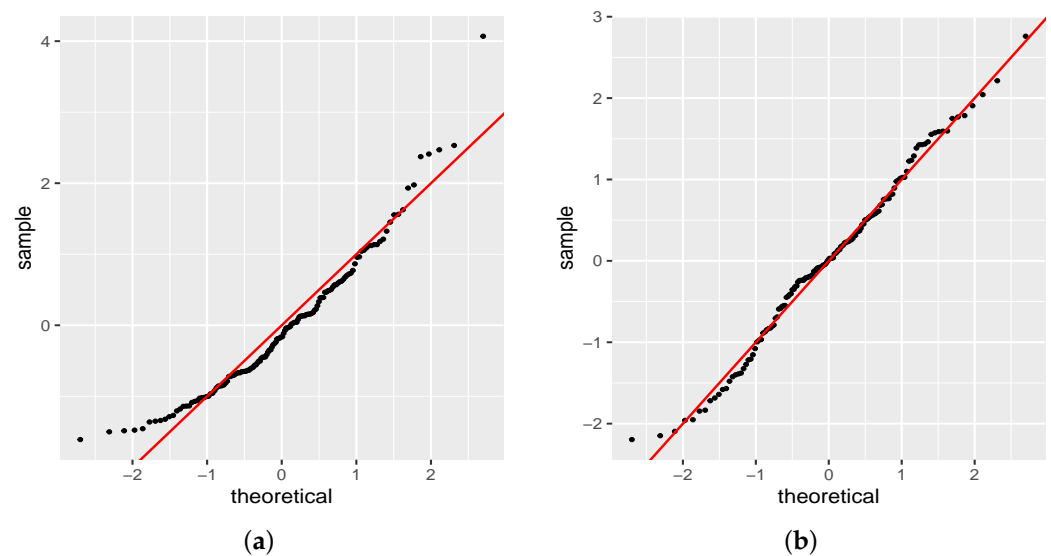
Figure 4 shows the normal quantile-quantile plot (q-q plot) for the randomized quantile residuals of the NBR-S and PIGR fitted models. The q-q plot is a scatterplot created by plotting the empirical quantiles of the residuals against the theoretical quantiles of the normal distribution. If residuals are normally distributed then they should form an approximately straight line. Figure 5 shows the worm plot. This graph was proposed by van Buuren and Fredriks [31] to identify regions (intervals) of the explanatory variable within which the model does not fit the data adequately [15]. In this graph, the upwardsline of the q-q plot is rotated to the horizontal in order to remove the trend and the Y axis contains the difference between its location in the theoretical and empirical distributions. If the residuals follow a normal distribution then the Y values are near the horizontal line and consequently inside the confidence band. The R function `wp()` provides the worm plot for a `gamLSS` fitted model. As one can note, both figures indicate the PIGR model performs better than the NBR model. In addition, the graphs of the residuals from the PIGR model indicate that there is no reason to worry about the inadequacy of the fit. Table 6 shows the estimates for the parameters of the PIGR model.

Figure 6 shows estimated relationships between the response variable and explanatory variables. As expected, the relationship with  $X_1$  (month) presents a cyclical behavior,

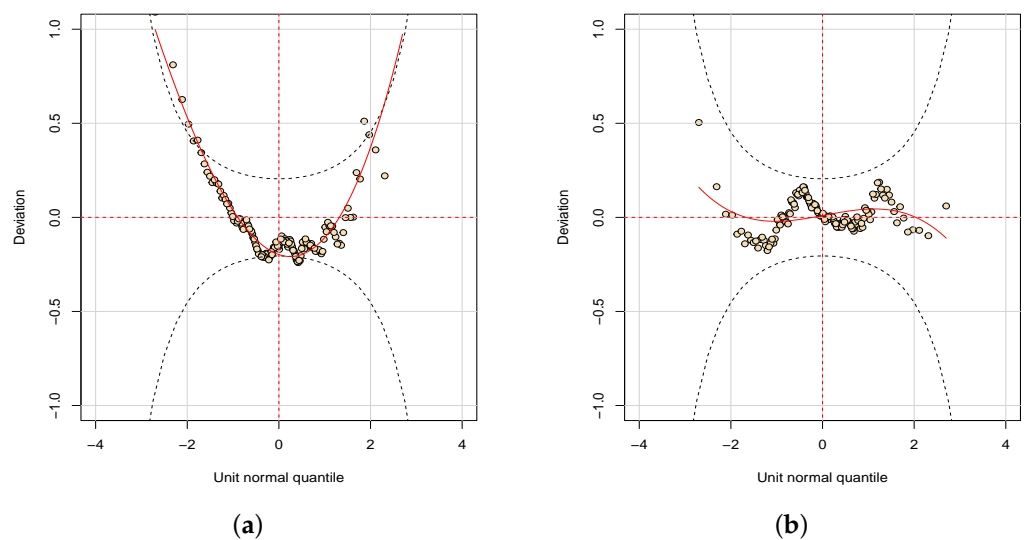
and the relationship with  $X_2$  and  $X_3$  is linear. These graphs were constructed using the `term.plot` function of the R software.

**Table 6.** Estimates for parameters of PIGR model.

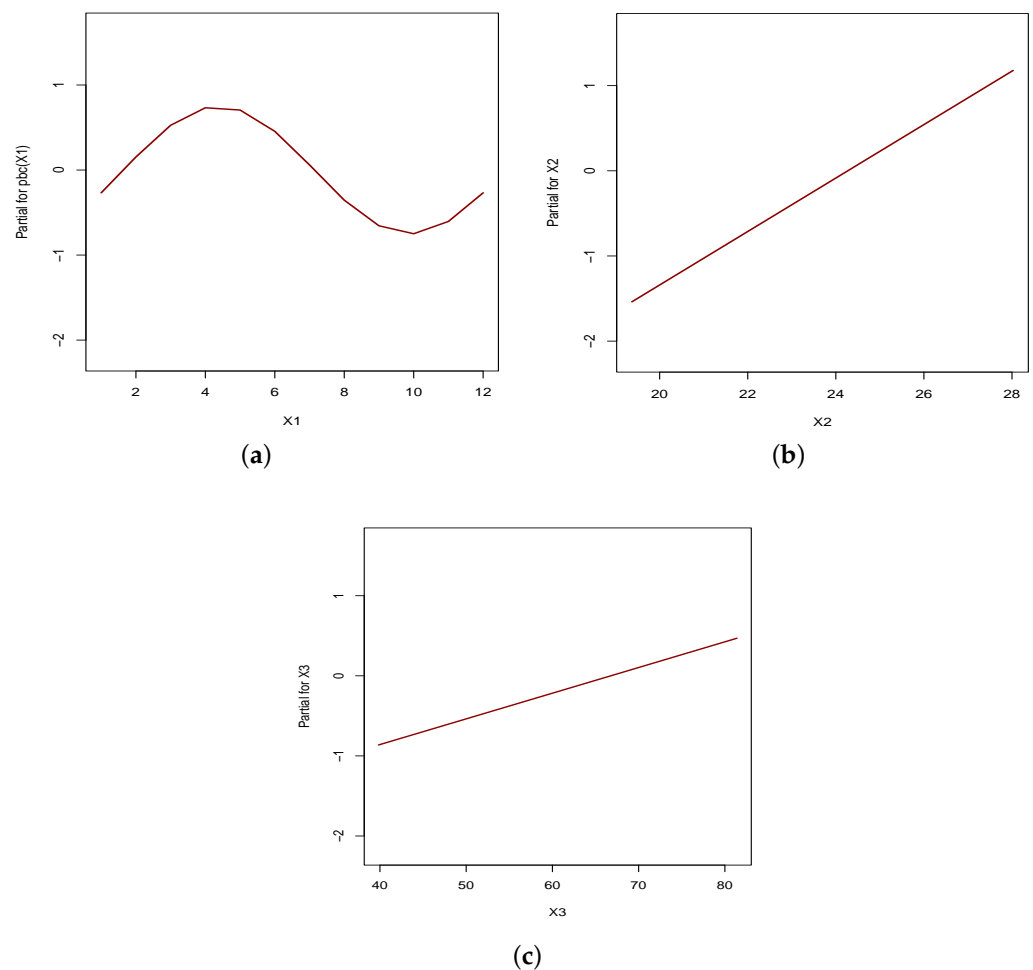
Parameter	Estimate	Str. Dev.	p-Value
$\beta_0$	-3.3276	1.5920	0.0384
$\beta_2$	0.313 8	0.0596	<0.0001
$\beta_3$	0.0321	0.0119	0.0080
$\tau$	2.0484	0.2649	<0.0001



**Figure 4.** Normal quantile-quantile plot for the residuals. (a) NBR model. (b) PIGR model.



**Figure 5.** Worm plot. (a) NBR model. (b) PIGR model.

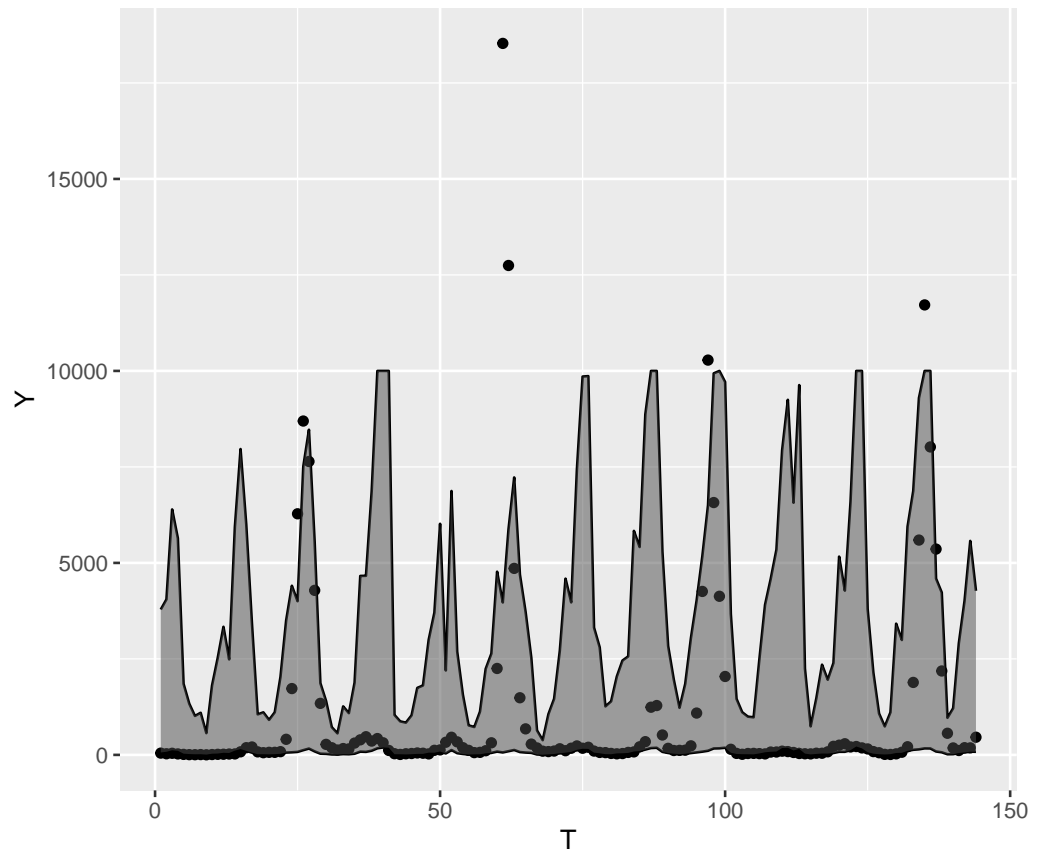


**Figure 6.** Estimated relationship between response variables and explanatory variables. (a) Relationship with  $X_1$ . (b) Relationship with  $X_2$ . (c) Relationship with  $X_3$ .

Figure 7 shows the number of registered dengue cases (symbol  $\bullet$ ) and a confidence band of 95% generated from the fitted PIGR model. In order to construct the confidence band we use a parametric bootstrap. That is, from estimated value  $\hat{\mu}_t$  and  $\hat{\tau}$ , we generate  $L = 1000$  values from a PIG distribution using the `rPIG` function of the R software. Then we set the lower and upper limits as being the percentiles 2.5% and 97.5% of the generated values. As one can note, the fitted model indicates that every year a peak will occur. How high or low the recorded number of dengue cases will be in relation to the expected peak (given by the fitted model) is controlled by action taken to combat the proliferation of the mosquito. If such action is effective, there is no occurrence of a peak, as in years 2008, 2009, 2011, 2012, 2014 and 2017. Otherwise, the peak may be higher than expected, i.e., there may be a larger outbreak, as in the years 2010, 2013, 2016 and 2019. That is, human behavior has a great influence on the number of cases that will be recorded. However, since this behavior is very difficult to quantify and is not present in the proposed model, this also has an influence on the predictive performance of the fitted model.

For example, in the next year after the years with peaks of cases (2007, 2010, 2013 and 2016), there was a significant reduction of recorded cases due to the implementation of actions to combat the proliferation of the disease vector and awareness campaigns reminding the population what happened the previous year. However, with the expected reduction in the number of recorded dengue cases obtained, the combat actions and awareness campaigns were not maintained, leading to an increase in the number of cases in the following two years. This has been occurring cyclically over the last 13 years.

Thus, although the proposed model does not present a satisfactory predictive performance, especially due to our inability to quantify and insert into the model the actions taken to combat the transmitting mosquito, it has at least three advantages: (i) better performance in comparison to the usual approaches, which are based on the fitting of PR and NBR models; (ii) the fitted model shows that a peak will occur every year and that the only way to avoid this peak is via the implementation of actions to combat the proliferation of the transmitting mosquito; and (iii) the fitted model shows which are the months of the year in which combat actions must be implemented.



**Figure 7.** Recorded values and confidence band (95%) generated from fitted model.

#### 4. Final Remarks

Dengue is a disease that affects millions of people every year, especially in tropical nations, causing a great impact on public health systems. Due to this, there is an interest in the development of statistical models that can forecast the number of dengue-fever cases and also identify which environmental variables may be related to the number of recorded cases.

In this article, we present statistical modeling for an overdispersed, long-tailed dengue-fever dataset. The proposed modeling is based on the assumption that the recorded number of dengue-fever cases in a month is generated according to a Poisson-inverse-Gaussian distribution. According to Zhu and Joe [7], this distribution may be used for modeling overdispersed, long-tailed datasets and presents a larger range of skewness than negative binomial distribution.

We model the expected number of dengue-fever cases as being linked to a set of explanatory variables through a log-linear function. This approach is called a Poisson-inverse-Gaussian regression (PIGR) model. In order to estimate the parameters of interest, we adopt the maximum-likelihood method. Since the estimators do not have known analytic solutions, we obtain estimates numerically by using the `gamLSS()` function of the `gamLSS` package of the R software.



We compare the proposed modeling to the usual approach based on the use of a negative-binomial regression (NBR) model. The two models were compared by using the AIC, BIC and RMSE criteria. Additionally, we compare the two models in terms of randomized quantile residuals. Three model-selection criteria indicate the PIGR model as better than the NBR model for this application. The randomized quantile residuals also indicate that PIGR performs better than NBR. That is, it has a quantile-quantile normal plot with residuals near the line  $y = x$ , and a worm plot with residuals near the horizontal line.

The fitted PIGR model indicates that variables  $X_1$  (month),  $X_2$  (temperature) and  $X_3$  (humidity) are related to the recorded number of dengue-fever cases. Variables  $X_2$  and  $X_3$  are positively related to the number of dengue cases. That is, an increase of the temperature and/or the humidity in the air is expected to lead to an increase in the recorded number of dengue cases. This makes intuitive sense because these two variables are directly related to favorable conditions for the development of the mosquito that transmits dengue fever. According to Silva et al. [32] "the female mosquito, infected and subjected to temperatures of approximately 32 °C, has 2.64 times more chance of completing the incubation period than those subjected to mild temperatures".

As a final result, the fitted model expects a peak in cases every year (see Figure 7). Based on this result, we conjecture the only way to avoid the peak is through an intervention by humans which avoids the proliferation of the transmitting mosquito. All analyses were performed using the R software and source code can be obtained by emailing the authors.

To end the paper we highlight the following three points: (i) Although the proposed modeling has presented better performance than usual approaches, it shares the basic assumptions of the Poisson and negative binomial regression models, which are: log-linearity in model parameters and independence of individual observations; (ii) as pointed out before, the main advantage is to be able to model overdispersed long-tail datasets; (iii) However, since dengue fever data are recorded longitudinally, they may present some kind of temporal correlation. Thus, the development of a modeling approach that incorporates correlation among recorded values for the answer variable can be viewed as future work. An approach we are currently studying is the development of a PIGR mixed-effect model.

**Author Contributions:** Conceptualization, V.P.V. and C.A.d.B.P.; Data curation, V.P.V. and M.V.F.; Formal analysis, E.F.S.; Methodology, E.F.S. and M.G.; Project administration, E.F.S. and C.A.d.B.P.; Software, V.P.V. and M.V.F.; Supervision, E.F.S. and C.A.d.B.P.; Writing–review & editing, M.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding

**Institutional Review Board Statement:** Not applicable

**Data Availability Statement:** The real dataset is freely available on the websites cited in the article. It also can be obtained by emailing the authors.

**Acknowledgments:** The authors thank the Universidade Federal de Mato Grosso do Sul and CNPq. The authors are grateful to the editor and referees for helpful comments and suggestions which have led to an improvement of this article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. World Health Organization. Dengue and Severe Dengue Fact Sheet. Retrieved from World Health Organization. Available online: <http://www.who.int/mediacentre/factsheets/fs117/en/> (accessed on 29 January 2021).
2. Leslie, C. Statistical Analysis of Climate Factors Influencing Dengue Incidences in Colombo, Sri Lanka: Poisson and Negative Binomial Regression Approach. *Int. J. Sci. Res. Publ.* **2019**, *9*, 133–144.
3. Sinaga, J.P.; Sinulingga, U. Poisson Regression Modeling Case Study Dengue Fever in Medan City in 2019. *J. Math. Technol. Educ.* **2019**, *1*, 94–102.
4. Mukhaiyar, U.; Huda, N.M.; Andirasdini, I.G.; Pasaribu, U.S. Forecasting of dengue hemorrhages fever cases with autoregression distributed lag model using Poisson regression approach. *Biostat. Epidemiol.* **2022**, *1*, 1–12. [[CrossRef](#)]

5. McCullagh, P.; Nelder, J. *Generalized Linear Models*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 1989; ISBN 0-412-31760-5.
6. Zha, L.; Lord, D.; Zou, Y. The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data. *J. Transp. Saf. Secur.* **2016**, *8*, 18–35. [[CrossRef](#)]
7. Zhu, R.; Joe, R. Modelling heavy-tailed count data using a generalised Poisson-inverse Gaussian family. *Stat. Probab. Lett.* **2009**, *79*, 1695–1703. [[CrossRef](#)]
8. Putri, G.N.; Nurrohmah, S.; Fithriani, I. Comparing Poisson-Inverse Gaussian Model and Negative Binomial Model on case study: Horseshoe crabs data. *J. Phys. Conf. Ser.* **2020**, *1442*, 012028. [[CrossRef](#)]
9. Nikoloulopoulos, A.K.; Karlis, D. On modeling count data: A comparison of some well-known discrete distributions. *J. Stat. Comput. Simul.* **2008**, *78*, 437–457. [[CrossRef](#)]
10. Dean, C.; Lawless, J.; Willmot, G. A mixed Poisson-inverse Gaussian regression model. *Can. J. Stat.* **1989**, *17*, 171–181. [[CrossRef](#)]
11. Heller, G.Z.; Couturier, D.L.; Heritier, S. Beyond mean modelling: Bias due to misspecification of dispersion in Poisson-inverse Gaussian regression. *Biom. J.* **2018**, *61*, 333–342. [[CrossRef](#)]
12. Stein, G.Z.; Juritz, J.M. Linear models with an inverse Gaussian Poisson error distribution. *Commun. Stat.-Theory Methods* **1988**, *17*, 557–571. [[CrossRef](#)]
13. R Core Team. *R Foundation for Statistical Computing; R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2020. Available online: <https://cran.r-project.org/> (accessed on 1 November 2020).
14. Rigby, R.A.; Stasinopoulos, D.M. Generalized additive models for location, scale and shape (with discussion). *Appl. Stat.* **2005**, *54*, 507–554. [[CrossRef](#)]
15. Stasinopoulos, M.D.; Rigby, R.A.; Akantziliotou, C. Instructions on How to Use the Gamlss Package in R. Available online: <http://www.gamlss.com/wp-content/uploads/2013/01/gamlss-manual.pdf> (accessed on 15 March 2020).
16. Rigby, R.A.; Stasinopoulos, D.M. The GAMLSS project: A flexible approach to statistical modelling. In *New Trends in Statistical Modelling, Proceedings of the 16th International Workshop on Statistical Modelling, Odense, Denmark, 2–6 July 2001*; Klein, B., Korsholm, L., Eds.; Statistical Modelling Society; Volume 337, 345p. Available online: [http://www.statmod.org/workshops\\_archive\\_proceedings\\_2001.htm](http://www.statmod.org/workshops_archive_proceedings_2001.htm) (accessed on 15 March 2020).
17. Akantziliotou, K.; Rigby, R.A.; Stasinopoulos, D.M. The R implementation of Generalized Additive Models for Location, Scale and Shape. In *Statistical Modelling in Society, Proceedings of the 17th International Workshop on Statistical Modelling, Chania, Greece, 8–12 June 2002*; Stasinopoulos, M., Touloumi, G., Eds.; Statistical Modelling Society; pp. 77–87. Available online: [http://www.statmod.org/workshops\\_archive\\_proceedings\\_2002.htm](http://www.statmod.org/workshops_archive_proceedings_2002.htm) (accessed on 15 March 2020).
18. Akaike, H.A. New look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
19. Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **1987**, *52*, 345–370. [[CrossRef](#)]
20. Schwarz, G.E. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [[CrossRef](#)]
21. Dunn, P.K.; Smyth, G. K. Randomized quantile residuals. *J. Comput. Graph. Stat.* **1996**, *5*, 236–244.
22. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*; Springer: New York, NY, USA, 2002.
23. Hinde, J.; Demétrio, C.G.B. Overdispersion: Models and estimation. *Comput. Stat. Data Anal.* **1998**, *27*, 151–170. [[CrossRef](#)]
24. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*; Cambridge University Press: London, UK, 1998.
25. Freitas, R.F.S.; Fávero, L.P.; Belfiore, P.; Correa, H.L. Package “Overdisp”. 2020. Available online: <https://cran.r-project.org/web/packages/overdisp/overdisp.pdf> (accessed on 10 August 2022).
26. Stasinopoulos, M.D.; Rigby, R.A.; Heller, G.Z.; Voudouris, V.; De Bastiani, F. *Flexible Regression and Smoothing: Using GAMLSS in R*; CRC Press: Boca Raton, FL, USA, 2017.
27. Casella, G.; Berger, R.L. *Statistical Inference*; Duxbury: Pacific Grove, CA, USA, 2002; Volume 2.
28. Instituto Nacional de Pesquisas Espaciais (INPE). The Brazilian National Institute for Space Research. Local Climatology Web Page for Campo Grande. Available online: [http://sonda.ccst.inpe.br/estacoes/campogrande\\_clima.html](http://sonda.ccst.inpe.br/estacoes/campogrande_clima.html) (accessed on 29 January 2021).
29. Fox, J.; Monette, G. Generalized collinearity diagnostics. *JASA* **1992**, *87*, 178–183. [[CrossRef](#)]
30. Feng, C.; Li, L.; Sadeghpour, A. A comparison of residuals diagnosis tools for diagnosing regression models for count data. *MC Med. Res. Methodol.* **2020**, *20*, 175. [[CrossRef](#)] [[PubMed](#)]
31. van Buuren, S.; Fredriks, M. Worm plot: A simple diagnostic device for modelling growth reference curves. *Stat. Med.* **2001**, *20*, 1259–1277. [[CrossRef](#)] [[PubMed](#)]
32. Silva, J.S.; Mariano, Z.F.; Scopel, I. A influência do clima urbano na proliferação do mosquito *Aedes aegypti* em Jataí (GO) na perspectiva da geografia médica. *Hyheia* **2017**, *2*, 33–49.