Acta Scientiarum
http://periodicos.uem.br/ojs
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v46i1.62579

Huanglongbing vector insect counting (HLB) by GAMLSS

Mateus Silva Pedroso, Terezinha Aparecida Guedes, Willian Luís de Oliveira, Wérica Bruna da Silva Valim, William Mário de Carvalho Nunes and Vanderly Janeiro*

Universidade Estadual de Maringá, Avenida Colombo, 5790, 87020-900, Maringá, Paraná, Brazil. *Author for correspondence. E-mail: vjaneiro@uem.br

ABSTRACT. Citriculture is one of the most important agricultural activities globally, with Brazil being one of the leading world producers. Thus, such activity is essential for the country's economy and the producers who depend on it. In this sense, the fight against Huanglongbing, one of the most devastating citrus diseases caused by vector insects, is essential to guarantee the quality of the fruit and avoid economic losses. The present work analyzed the counting of insect vectors in a commercial orange orchard in an observational study carried out in the municipality of Paranavaí, state of Paraná, Brazil, using the methodologies of generalized linear mixed models (GLMM) and generalized additive models for location, scale, and form (GAMLSS), with Negative Binomial probability distribution. Data were obtained by counting insects trapped in sticky traps at twelve fixed points in the orchard at three different heights and collected over seven fortnights. The results indicated that the GAMLSS model presented better results by including the linear predictor for modeling the scale parameter associated with the study factors based on the AIC criterion and diagnostic analysis tools.

Keywords: regression models; longitudinal data; repeated measures; random effects.

Received on February 19, 2022.
Accepted on January 9, 2024.

Introduction

Agriculture is one of the main bases of the Brazilian economy, and agribusiness is responsible for boosting the country's GDP and holding economic activity in times of recession since Brazil is one of the largest exporters of agricultural products globally. In this scenario, citrus farming stands out as one of the most relevant activities in the sector, given that Brazil is the world leader in the production of sweet oranges and concentrated orange juice. Thus, several types of research have been produced in search of answers about improvements in production and losses caused by diseases and pests, providing data that must be analyzed with the support of statistics to subsidize the researcher’s interests.

In this sense, the data generated by these surveys are generally discrete, representing counts of organisms in a plant and the presence or absence of a particular disease, for example. In addition, in some cases, there is an interest in following the behavior of the response over time, characterizing longitudinal data and repeated measures. In this way, choosing the appropriate methodology for constructing a model must respect the nature of the data and the experimental design.

In the literature, the use of generalized linear mixed models (GLMM) is highlighted, which allows the inclusion of correlation between responses and the definition of a probability distribution for the response that is not necessarily the normal distribution, belonging to the exponential family of distributions. More recently, Rigby and Stasinopoulos (2001; 2005) introduced generalized additive models for location, scale, and shape (GAMLSS). This class of models was proposed to make the assumptions of the previously existing classes of models even more flexible and allow the adjustment of responses whose distribution belongs to a more generic family of distributions. Furthermore, it allows the modeling of other parameters, in addition to the average, by the linear predictor.

Citrus greening, also called HLB (Huanglongbing), is a disease caused by bacteria *Candidatus* Liberibacter africanus (CLaf), *Candidatus* Liberibacter asiaticus (CLas), and *Candidatus* Liberibacter americanus (CLam), causing yellowing of leaves and premature fruit drop. The psyllid *Diaphorina Citri* is the vector insect responsible for the transmission and propagation of the disease. Therefore, understanding more about the distribution and behavior of HLB vector insects concerning the citrus production environment is essential for devising disease control and management strategies since it is known that the insect is capable of dispersing

over short and long distances, making it difficult to control. Therefore, the present study consisted of monitoring the dispersion height by a flight of insects of the superfamily Psylloidea in commercial orange orchards with symptomatic plants.

This work aimed to fit a negative binomial regression model to the observed insect count, considering possible explanatory factors and random effects that may influence insect distribution behavior. In addition, we compared the methodologies of generalized linear mixed models (GLMM) and generalized additive models for location, scale, and shape (GAMLSS) in estimating model parameters.

Material and methods

Data set

The observational study was carried out in the municipality of Paranavaí, state of Paraná, Brazil, in a commercial orange orchard with the incidence of Huanglongbing (HLB), by the student Wérica Bruna da Silva Valim in her master's studies in Agronomy from the State University of Maringá, 2021. Three plots were selected, and four bamboos were fixed at randomly chosen points. In each bamboo, sticky traps were set at three height levels, namely, 2.5m, 4.5m, and 7m, to capture the insects of the superfamily Psylloidea, disease vectors. The traps were collected every fifteen days for three months, totaling seven collections. In addition, insects of the superfamily Psylloidea were counted by trap/collection, and the value was recorded (response variable), giving rise to a longitudinal study with 252 observations by the combination of factors.

Generalized linear mixed models

The generalized linear mixed models (GLMM) were proposed by Breslow and Clayton (1993) as an extension to the linear mixed models (LMM) introduced by Laird and Ware (1982) and generalized linear models (GLM) (Nelder & Wedderburn, 1972) for modeling non-normally distributed responses as well overdispersion and correlation by including random effects. The formulation is given by (i) distribution for the response variable that is a member of the exponential family of distributions and (ii) a linear predictor that relates to the expected value of the response variable by (iii) an appropriate link function $g(\cdot)$.

i. Let a random sample of size N and n_i observations for each individual i, given a vector b_i of random effects, dimension $q \times 1$, be the conditional distribution of y_{ij} for all $i = 1,2,\dots,N$ and $j = 1,2,\dots,n_i$, belongs to the exponential family of distributions with density function:

$$f(y_{ij}, \theta_{ij}, \phi \mid b_i) = exp\{\phi^{-1}[y_{ij}\theta_{ij} - b(\theta_{ij})] + c(y_{ij}, \phi)\}. \quad (1)$$

where θ_{ij} is called the canonical parameter, ϕ is the dispersion parameter and the functions $b(\cdot)$ and $c(\cdot)$ are known.

ii. The linear predictor is given by:

$$\eta_i = X_i\beta + Z_i b_i, \quad (2)$$

with $\eta_i = (\eta_{i1}, \eta_{i2}, \dots, \eta_{in_i})^T$, X_i, dimension $n_i \times p$, is the model matrix associated with the fixed effects and β, dimension $p \times 1$, is the vector of coefficients (fixed effects). Z_i is a matrix of dimension $n_i \times q$, which can be a subset of X_i ($q \leq p$), with the explanatory variables modeled as random in the model. The random terms are modeled by the vector b_i, dimension $q \times 1$, with multivariate normal distribution with zero means vector and covariance matrix G.

iii. The link function is given by:

$$g(\mu_i) = \eta_i = X_i\beta + Z_i b_i \quad (3)$$

where $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{in_i})^T$.

Maximum likelihood estimation

The estimation of generalized linear mixed models is based on the maximum likelihood method, although its complexity is higher than the models presented above. Specifying the joint probability distribution of $\boldsymbol{Y_i}$ and $\boldsymbol{b}_i$ as:

$$f(Y_i, b_i) = f(Y_i \mid b_i)f(b_i), \quad (4)$$

the distribution of $f(\boldsymbol{Y}_i \mid \boldsymbol{b}_i)$ belongs to the exponential family of distributions, and $f(\boldsymbol{b}_i)$ has a normal distribution. The marginal likelihood function, in turn, can be determined by integrating over the distribution of unobservable effects, $\boldsymbol{b_i}$, given by:

$$L(\theta, \phi, y) = \prod_{i=1}^{n} \int f(Y_i \mid b_i) f(b_i) db_i, \quad (5)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ and $\boldsymbol{\alpha}$ represents the vector of the model's variance parameters, known as variance components.

Thus, the likelihood function maximization in Equation 5 about the parameters will lead to the estimates of $\boldsymbol{\theta}$ and ϕ. The complexity, in this case, is because such maximization has no analytical solution, and therefore, numerical integration methods must be used for its resolution, such as the Gaussian quadrature method and Laplace approximation, among others. Some functions implemented in the R environment can be used to estimate the parameters in this case. For instance, from the glmmTMB package (Brooks et al., 2017), the function of the same name; however, in its documentation, users are warned that the REML option "can also be useful for some non-Gaussian response variables if used with caution." Another option is to use the gamlss function from the gamlss package (Rigby & Stasinopoulos, 2005).

Generalized additive models for location, scale, and shape (GAMLSS)

The generalized additive models for location, scale, and shape were proposed by Rigby and Stasinopoulos (2005) and allow flexibilization of the assumption associated with the exponential family of distributions by a more general family of distributions. In addition, it enables modeling all population parameters of the distribution associated with factors and covariates by including a linear predictor. Furthermore, it is possible to include nonparametric smoothing functions and random effects in the linear predictor, and, therefore, they are known to be semi-parametric.

Given a random sample of size N, let y_i be conditionally independent observations on θ^i, for $i = 1,2, \ldots, N$, with probability function $f(y_i \mid \theta^i)$ and $\theta^{iT} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})$ is a vector of p parameters associated with explanatory variables and random effects. Taking $y^T = (y_1, y_2, \ldots, y_N)$ as a vector of observations of the response variable whose parameters θ_k, with $k = 1,2, \ldots, p$, will be linked to the explanatory variables and random effects through a link function $g_k(\cdot)$, we have the following relationship:

$$g_k(\theta_k) = \eta_k = X_k \beta_k + \sum_{j=1}^{J_k} Z_{jk} \gamma_{jk}, \quad (6)$$

where $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k$ are vectors of size N, $\boldsymbol{\beta}_k$ is the vector of regression parameters of size $J_k{}'$, $\boldsymbol{X}_k$ and $\boldsymbol{Z}_{jk}$ are planning matrices of dimensions $N \times J_k{}'$ and $N \times q_{jk}$. γ_{jk} is a q_{jk}-dimensional random variable generally with $\gamma_{jk} \sim N_{jk}(0, \boldsymbol{G}^{-1}{}_{jk})$, where $\boldsymbol{G}^{-1}{}_{jk}$ is the generalized inverse of $\boldsymbol{G}_{jk}(\lambda_{jk})$, and λ_{jk} is a vector of hyperparameters, naming the so-called model GAMLSS. The gamlss function of the *gamlss* package of the R software is used to adjust the model.

Estimation

For the estimation of the GAMLSS models, the maximum likelihood and penalized maximum likelihood methods are used in the case of parametric and nonparametric models, respectively. The log-likelihood function, given an observed sample of size N, is given by:

$$l = \sum_{i=1}^{N} log\, f(y_i \mid \theta^i) \quad (7)$$

while the penalized log-likelihood is given by:

$$l_p = l - \frac{1}{2} \sum_{k=1}^{p} \sum_{j=1}^{J_k} \lambda_{jk} \gamma_{jk}{}^T G_{jk} \gamma_{jk}. \quad (8)$$

In the simplest case of parametric GAMLSS models, estimation of only $\boldsymbol{\beta}$ is required by maximizing the function given in Equation 7. For nonparametric GAMLSS models, in addition to $\boldsymbol{\beta}$, it is necessary to estimate the parameters γ given a vector of hyperparameters $\boldsymbol{\lambda}$ fixed from the maximization of Equation 8. The assumption that $\boldsymbol{\lambda}$ is constant is guaranteed from the setting of these parameters or their estimation, which is done automatically within the estimation algorithms discussed later.

RS and CG algorithms

According to Rigby and Stasinopoulos (2005), the algorithms used to maximize the (penalized) log-likelihood function are generalizations of the CG and RS algorithms, based on Cole and Green (1992) and Rigby and Stasinopoulos (1996), respectively. The CG algorithm uses the expected or observed value of the second derivative and cross-derivatives of the log-likelihood function concerning the distribution parameters. However, when the parameters are orthogonal, that is, when the cross derivatives are equal to zero, estimation by this method will not be possible. In this way, the RS algorithm can be used, considered more straightforward, since it does not use cross derivatives but can be computationally slower. It is still possible to use a combination of the two algorithms, alternating between them during iterations until convergence is reached. The definition of the algorithm to be used in the gamlss is given by method argument, which accepts the RS()inputs, by *default*, CG() and mixed().

Model selection

The model selection process comes from choosing a probability distribution (D), link functions (G), and hyperparameters $\boldsymbol{\lambda}$. In addition, the variables/factors to compose the model must be selected, such as its shape, using tools that allow analyzing its quality and measures and test statistics. In this way, a complementary analysis provides filtering and selecting a model considered adequate, given the assumptions.

One of the main criteria used in the literature is the Akaike's Information Criterion (AIC), proposed by Akaike (1974), which is a particular case of its generalized form, the GAIC (Generalized Akaike's Information Criterion), given by:

$$GAIC(k) = GD + k \cdot df, \tag{9}$$

where $GD = -2l(\widehat{\theta})$ is the Global Deviation and k is a penalty associated with the model's degrees of freedom, df. We have the AIC criterion taking $k = 2$ and other measures, such as the Bayesian Information Criterion (BIC), when $k = ln(n)$ can be calculated, and the model is chosen whose $GAIC$ has the smaller value. Since the information criteria have a similar form and aim to conclude the same thing, it was decided to calculate only the AIC in comparing the models using the GAIC() function of gamlss.

For the case of nested models, considering two models M_0 and M_1 with df_0 and df_1 degrees of freedom, respectively, the test statistic is obtained:

$$\Lambda = DG_0 - DG_1, \tag{10}$$

which has an asymptotic distribution χ^2 with $d = df_0 - df_1$ degrees of freedom, calling the so-called generalized likelihood ratio test. To do this, the LR.test() function of the gamlss package can be used.

Diagnostic analysis

Another critical and indispensable step for selecting a final model is a diagnostic analysis, which consists of checking the adequacy of the evaluation model. Graphic resources are mainly used for the behavior of the residuals, which, in this case, are the randomized quantile residuals; for more details see Dunn and Smyth (1996). The worm plot is one of the main diagnostic analysis tools, as it allows to check that the fit has been adjusted and identify patterns that can lead to a model correction. For the case-randomization software, Stasinopoulos et al. (2017) recommend using the rqres.plot() function from the R software, plotting the simulated residuals, and using multiple worm plots.

Results and discussion

For modeling discrete data, in particular, count data, the choice of probability distributions with this characteristic is limited, such as Poisson and Negative Binomial, in the case of the exponential family of distributions. In addition, the link function must also have the necessary support since, for the average, the count must be greater than or equal to zero.

Considering the insect count of the superfamily Psylloidea as a response variable, the Negative Binomial probability distribution was defined to adjust the data to compare the models by the GLMM and GAMLSS methodologies. There are stand, height, and the collection as fixed effect factors. The collection points, represented by bamboos, are considered a random effect factor in the modeling since they represent only a sample of all possible locations within the stand.

Thus, the interest is in the difference between the levels of field factors and height, the effect of time on the insect count distribution, and variability within the stand, represented by the point's random effect. In addition, the height factor can also be included as a random effect factor since, within the same point, there are responses at different heights, and possibly there is a correlation between the measures. Since the gamlss function of the R software allows adjusting any model by specifying its parameters correctly, all models were adjusted by this function, and the analysis tools of the GAMLSS models could be used.

A very common graph in statistical analysis is the boxplot, which allows to graphically visualize the main characteristics of data distribution, such as quartiles, dispersion, and especially, the visualization of outliers. In addition, one of the research interests is to verify the behavior of insect distribution over time. Thus, Figures 1 and 2 show the insect count boxplots in the collection order by field and height.

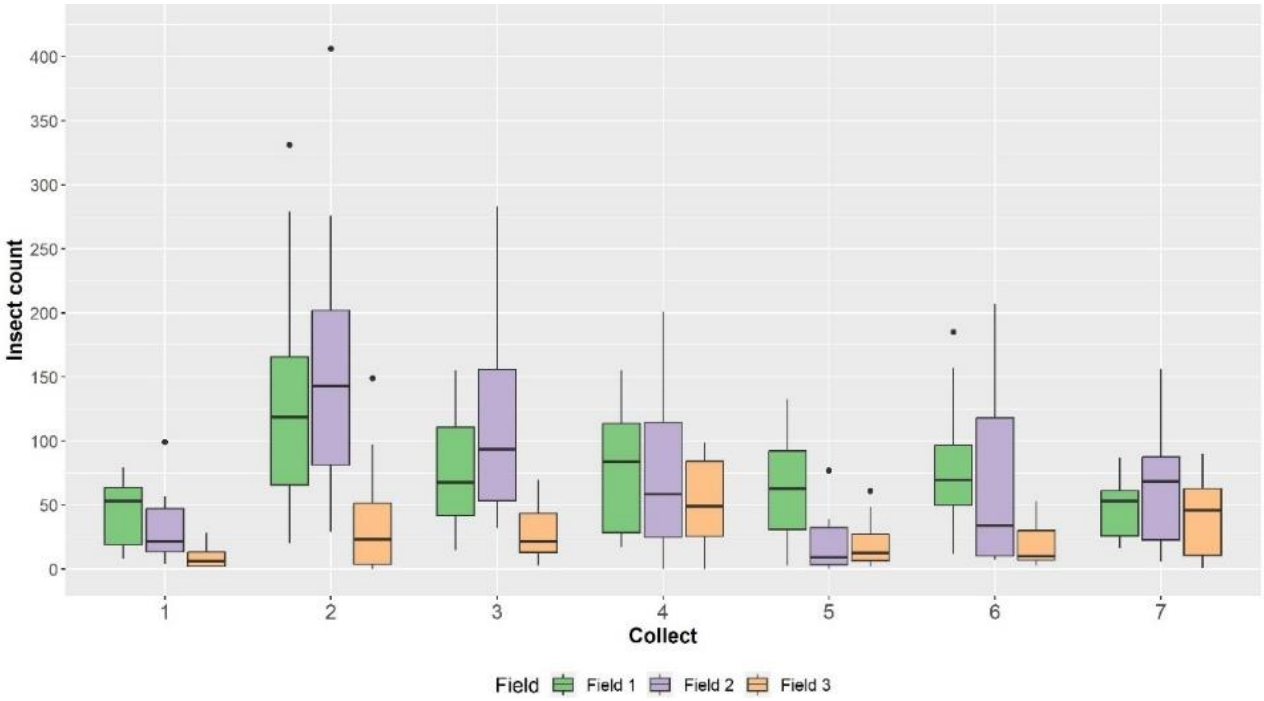


Figure 1. Boxplots of Psylloidea counts per stand, throughout the collection.

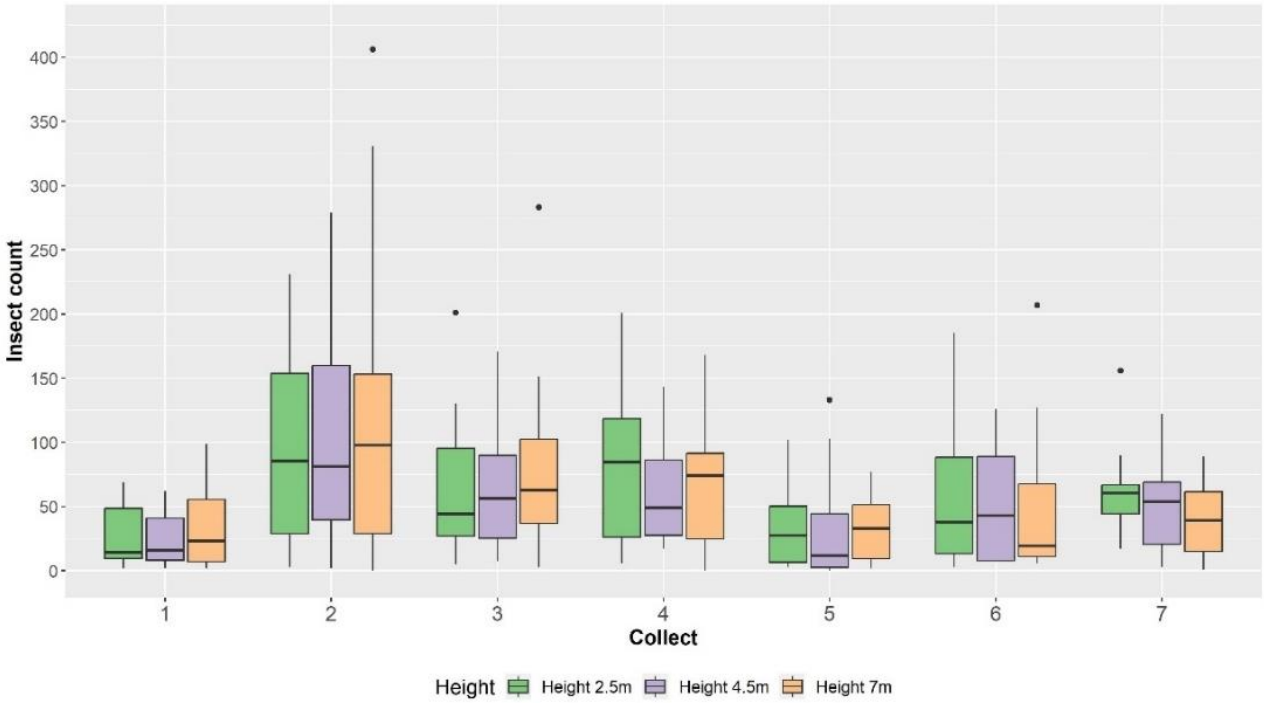


Figure 2. Boxplots of Psylloidea counts by height, throughout the collection.

In all these graphs, it is possible to observe that time, represented by the collection factor, influences the average insect count, oscillating throughout the collections. Regarding the field factor, there seems to be no difference in the insect count distribution over time in field 1, with the median oscillating in the range of approximately 50 to 80, except for collection 2, which had a median above 100 well for dispersion. Concerning stands 2 and 3, this difference in the count along the collections is more evident, both in the average and dispersion, with growth and decrease patterns. Comparing the stands to each other, only stand 3 significantly differed from the others, with observations primarily concentrated in the range from 0 to 50.

In Figure 2, there seems to be no effect of height since the distribution of insect counts throughout the collection was very similar at the three height levels. It is possible to observe the presence of several outliers, reaching the maximum value of the observations in the range of 400. Table 1 lists the descriptive measures associated with the study factors.

Table 1. Summary measures of the Psylloidea count variable.

		Minimum	Maximum	Median	Mean	Std. Deviation
Stand	1	3.00	331.00	64.00	74.11	58.43
	2	0.00	406.00	48.00	75.68	76.81
	3	0.00	149.00	19.50	29.39	29.76
Heigth	2.5	2.00	231.00	46.00	61.08	57.05
	4.5	0.00	279.00	42.50	56.87	57.02
	7.0	0.00	406.00	45.00	61.23	71.27
Collection	1	2.00	99.00	18.00	28.50	26.34
	2	0.00	406.00	94.50	110.47	100.91
	3	3.00	283.00	52.50	71.75	62.22
	4	0.00	201.00	71.00	68.08	50.71
	5	0.00	133.00	27.00	34.03	35.25
	6	3.00	207.00	31.00	54.75	55.53
	7	1.00	156.00	53.00	50.50	34.50

Finally, to complement the analysis, profile graphs allow viewing the behavior of the distribution over time by plotting the points, as in a scatter plot, and joining them by line segments. In Figure 3, stand, height, and point are identified, while in Figure 4, there are profile graphs for each subject. The subjects, identified from 1 to 36, refer to observational units classified according to the study factors combinations, as shown in Table 2.

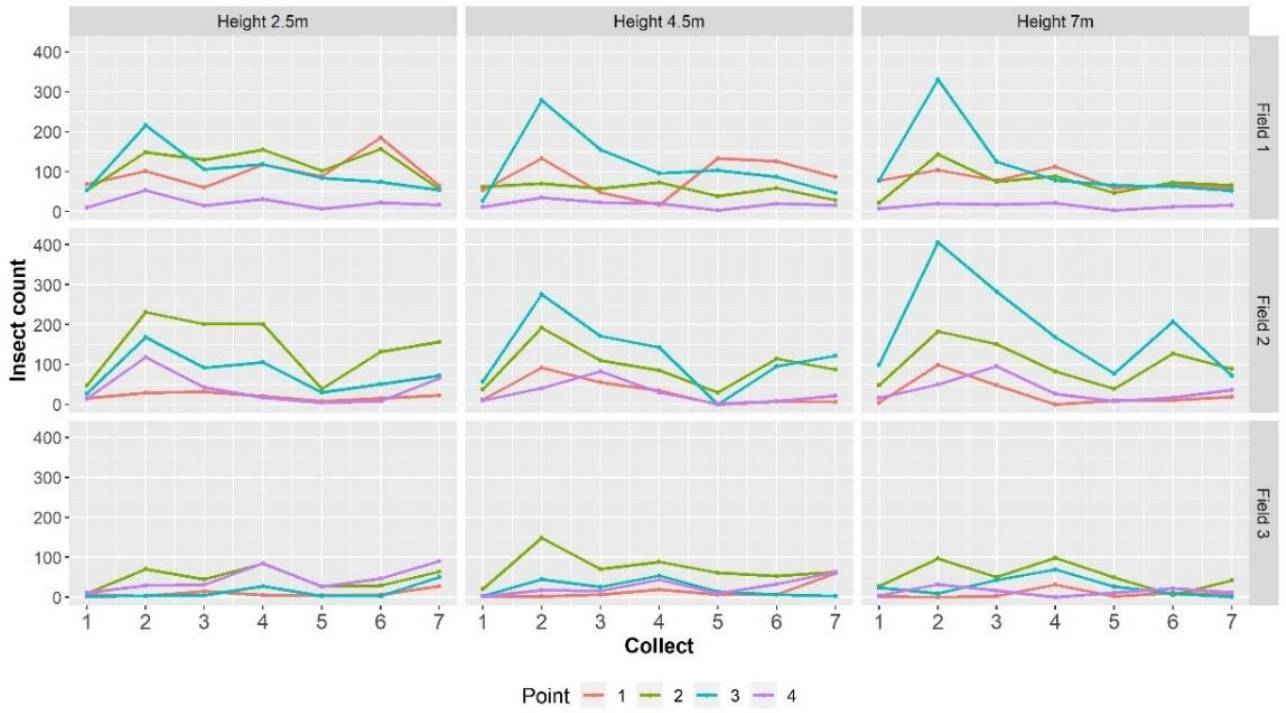


Figure 3. Profile plots of Psylloidea counts over time, identified by field, height, and point.

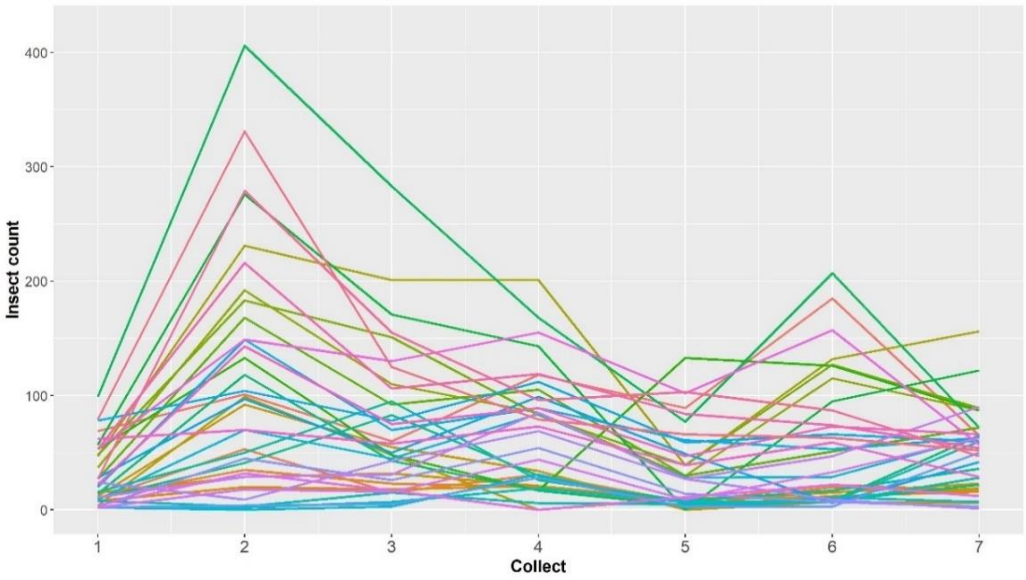


Figure 4. Individual profile plot.

Table 2. Identification of experimental units.

		Height (m)					Height (m)		
Stand	Point	2.5	4.5	7	Stand	Point	2.5	4.5	7
1	1	S1	S2	S3	1	3	S19	S20	S21
2	2	S4	S5	S6	2	4	S22	S23	S24
3	3	S7	S8	S9	3	1	S25	S26	S27
1	4	S10	S11	S12	1	2	S28	S29	S30
2	1	S13	S14	S15	2	3	S31	S32	S33
3	2	S16	S17	S18	3	4	S34	S35	S36

GLMM methodology

Initially, a generalized linear mixed model modeled only the mean via linear predictor, considering field, height, and collection as fixed effects factors and tested point and height and only point as random effects factors, as shown in Table 3. Therefore, we have the link function:

$$\begin{aligned} \text{Model M1: } g_1(\mu_{ijkt}) = \eta_{ijkt} &= \beta_0 + \alpha_i + \theta_j + \gamma_t + b_{j|k} + b_k \\ &= (\beta_0 + b_{j|k} + b_k) + \alpha_i + \theta_j + \gamma_t \\ \text{Model M2: } g_1(\mu_{ijkt}) = \eta_{ijkt} &= \beta_0 + \alpha_i + \theta_j + \gamma_t + b_k \\ &= (\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t \end{aligned} \quad (11)$$

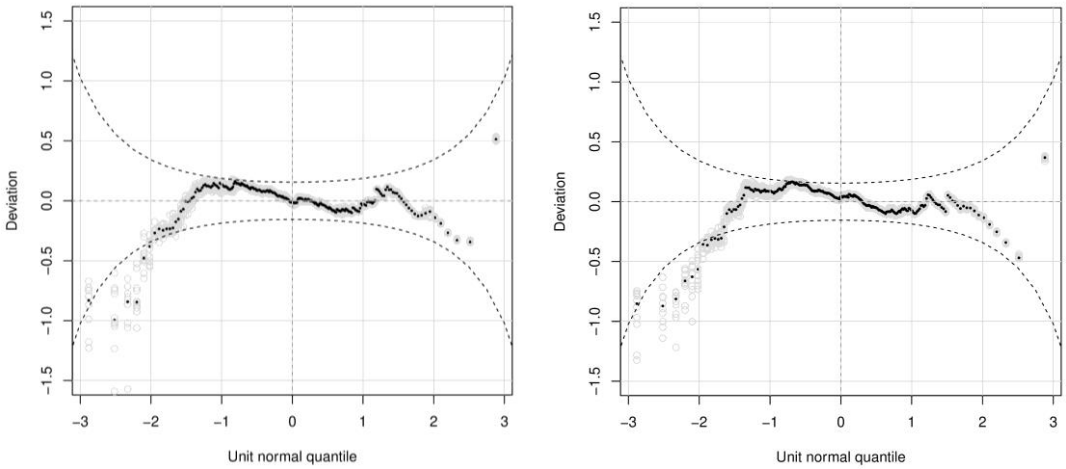
β_0 is the model intercept; α_i is the fixed effect associated with the i-th field, $i = 1,2,3$; θ_j is the fixed effect associated with the j-th height, $j = 1,2,3$, and γ_t is the fixed effect associated with the t-th collection, $t = 1,2,\ldots,7$. b_k and $b_{j|k}$ are the random effects associated with the k-th point and j-th height within the k-th point, with $k = 1,2,\ldots,12$, respectively. Assuming that b_k and $b_{j|k}$ follow Gaussian distribution with zero mean and constant variance, i.e., $b_k \sim N(0,\ \sigma_A^2)$ and $b_{j|k} \sim N(0,\ \sigma_B^2)$.

Table 3. Models fitted with Negative Binomial distribution, with two different random structures in the linear predictor and three different link functions.

Model	Link function	$g_1(\mu)$	$g_2(\sigma)$	AIC
M1	log	$(\beta_0 + b_{j\|k} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	2342.77
	inverse		-	2382.31
	sqrt		-	2343.78
M2	log	$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	-	2348.15
	inverse		-	2382.32
	sqrt		-	2349.89

According to the AIC values in Table 3, the link function "log" should be preferred for both models M1 and M2. Therefore, in the following analyses, the log link function will be considered.

From the worm plots of the models in Figure 5, the adjustments were not adequate since there was a sharp curvature in the tails of the graphs with several points outside the confidence bands. However, it was possible to verify the presence of an outlier impairing diagnostic analysis, and thus, the observation was removed for the readjustment of the model. Also, since the AIC criterion of model M1, with point and height as factors of random effects, has a lower value, such a random structure was chosen, corroborated by the conclusion of the likelihood ratio test presented in Table 4. In this way, Figure 6 shows the main residual graphs of the selected model, readjusted without the outlier.



(a) Model M1 (b) Model M2

Figure 5. Worm plot of fitted models, with linear predictor for μ.

Table 4. Likelihood Ratio Test (LRT).

Likelihood Ratio Test (LRT) for nested GAMLSS models		
Model M2 (Null):	Deviance = 2303.134	with 22.5055 degrees of freedom
Model M1 (Alternative):	Deviance = 2277.628	with 32.5727 degrees of freedom
LRT = 25.5068 with 10.0671 degrees of freedom, p-value = 0.00464		

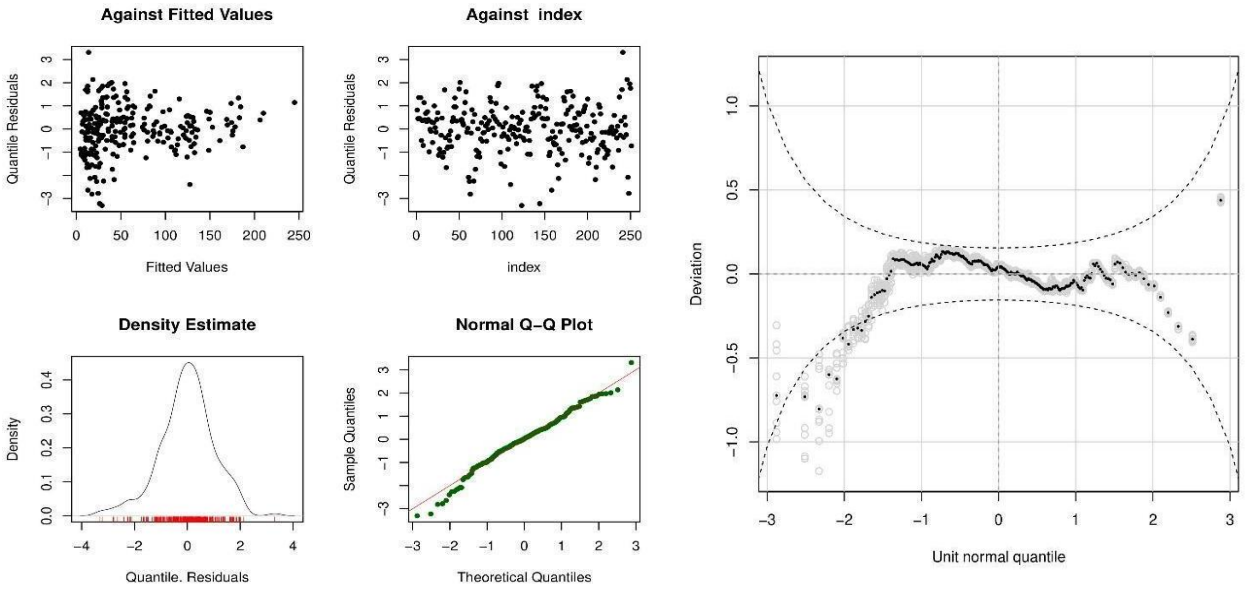


Figure 6. Residual plots of the fitted model, with linear predictor for μ.

In the residual graphs, even after removing the outlier, the model showed a deviation from normality, and the residuals showed a funneling pattern with greater dispersion in the initial range of the adjusted

values. The worm plot maintains the previous pattern with curvature and many points outside the confidence bands.

From the previous analysis, it can be concluded that the generalized linear mixed model was not sufficient to adequately model the data considering that the variability also seems to fluctuate by the levels of the factors, as observed in the descriptive analysis.

GAMLSS methodology

Here the GAMLSS model will be used with the inclusion of the linear predictor for the distribution dispersion parameter not supported in GLMM models. This is an alternative to solve problems observed in Figure 6. So, consider the link functions:

$$\begin{aligned} g_1(\mu_{ijkt}) = log(\mu_{ijkt}) = \eta_{ijkt}{}^{\mu} &= \beta_0 + \alpha_i + \theta_j + \gamma_t + b_{j|k} + b_k \\ &= (\beta_0 + b_{j|k} + b_k) + \alpha_i + \theta_j + \gamma_t \\ g_2(\sigma_{ijkt}) = log(\sigma_{ijkt}) = \eta_{ijkt}{}^{\sigma} &= \beta_0 + \alpha_i + \theta_j + \gamma_t. \end{aligned} \quad (12)$$

According to Figure 7, the inclusion of the linear predictor for σ was sufficient to considerably improve the fit of the model, appearing to have a good fit. The deviation from normality was corrected, and the variability of the residuals, this time, behaved randomly and without outliers. From the model with Negative Binomial probability distribution, point and height as random effects factors, and linear predictor for μ and σ, a more parsimonious model is sought after removing some factors, particularly those that did not show significance according to Table 5. The models tested are listed in Table 7 and were compared based on the AIC criterion and visual inspection of the worm plots.

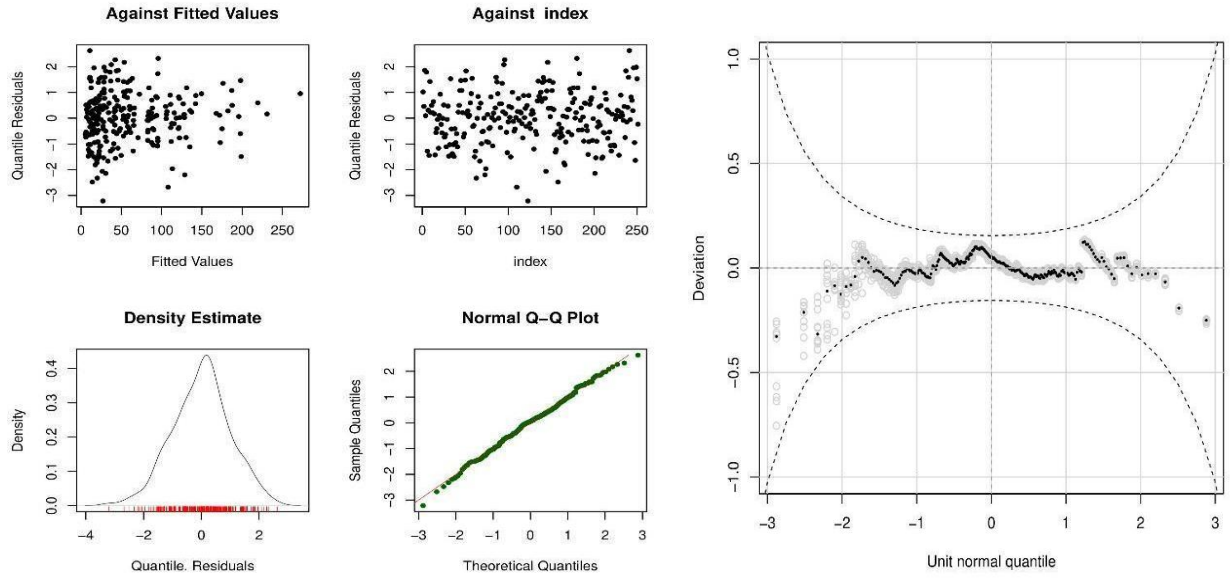


Figure 7. Residual plots of the fitted model, with linear predictors for μ and σ.

Table 5. Estimates of regression coefficients and variance parameters of the Negative Binomial distribution.

	$\widehat{g_1(\mu)}$			$\widehat{g_2(\sigma)}$		
Parameters	Estimative	Std. Error	p-value	Estimative	Std. Error	p-value
β_0	3.47	0.09	< 0.001	-2.66	0.38	< 0.001
α_2	-0.10	0.09	0.25	0.55	0.32	0.09
α_3	-0.99	0.11	0.001	1.50	0.28	< 0.001
θ_2	-0.07	0.08	0.42	0.43	0.25	0.09
θ_3	-0.05	0.08	0.54	0.25	0.26	0.35
γ_2	1.32	0.12	< 0.001	0.66	0.41	0.11
γ_3	0.89	0.12	0.001	0.10	0.45	0.82
γ_4	0.84	0.13	< 0.001	0.74	0.43	0.09
γ_5	0.14	0.13	0.30	0.92	0.42	0.03
γ_6	0.59	0.11	< 0.001	0.39	0.41	0.35
γ_7	0.59	0.14	< 0.001	0.70	0.44	0.11

Table 6. Models tested with the Negative Binomial distribution, with linear predictors for μ and σ, location and scale parameters respectively.

Model	$g_1(\mu)$	$g_2(\sigma)$	AIC
M1.1	$(\beta_0 + b_{j\|k} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	2302.46
M1.2	$(\beta_0 + b_{j\|k} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	**2298.22**
M1.3	$(\beta_0 + b_{j\|k} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i + \gamma_t$	2298.14
M1.4	$(\beta_0 + b_{j\|k} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i + \theta_j$	2294.30
M1.5	$(\beta_0 + b_{j\|k} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i$	2293.53

The models presented AIC values very close to each other, with an amplitude of approximately nine units between the highest and the lowest (Table 6). The model that best fitted the data, chosen as the final model is M1.2, which contains the fixed effects factors field and collection for the predictor of μ, and stand, collection, and height for the predictor of σ, with point and height as factors from random effects for μ. According to the regression coefficients in Table 7, the estimates and conclusions of the tests are in line with the initial model (Table 5).

Table 7. Estimates of regression coefficients and variance parameters of the Negative Binomial distribution – final model.

	$\widehat{g_1(\mu)}$			$\widehat{g_2(\sigma)}$		
Parameters	Estimative	Std. Error	p-value	Estimative	Std. Error	p-value
β_0	3.43	0.08	< 0.001	-2.67	0.38	< 0.001
α_2	-0.10	0.09	0.25	0.53	0.32	0.10
α_3	-1.00	0.11	< 0.001	1.48	0.28	< 0.001
θ_2	-	-	-	0.44	0.25	0.08
θ_3	-	-	-	0.26	0.27	0.33
γ_2	1.33	0.12	< 0.001	0.67	0.41	0.10
γ_3	0.89	0.11	0.001	0.08	0.45	0.85
γ_4	0.85	0.13	< 0.001	0.77	0.43	0.07
γ_5	0.14	0.13	0.30	0.94	0.42	0.03
γ_6	0.60	0.11	< 0.001	0.41	0.41	0.31
γ_7	0.60	0.14	< 0.001	0.75	0.43	0.08

Variance components

Source	$\hat{\sigma}$	95% conf. intervals
Point - b_k	0.6545	0.4272; 1.0027
Height in point - $b_{j\|k}$	0.1929	0.1043; 0.3565
Error	1.0243	0.9325; 1.1251

The diagnostic graphs and worm plot (Figure 8) indicate the fit is adequate, showing evolution compared to the mixed generalized linear model. Compared to the initial GAMLSS model, with all factors in the predictors, the final model had a similar performance, with the height factor not being significant in explaining the average insect count, as observed in the descriptive analysis. Therefore, the proposed model is suitable for adjusting the study data and can be written as:

$$
\begin{aligned}
log(\hat{\mu}_{ijkt}) = \hat{\eta}_{ijkt}{}^{\mu} = \quad & (3.43 + \hat{b}_k + \hat{b}_{j|k}) - 0.10\ Stand\ 2 - \\
& - 1.00\ Stand\ 3 + 1.33\ Collect\ 2\ + 0.89\ Collect\ 3 + \\
& + 0.85\ Collect\ 4 + 0.14\ Collect\ 5 + 0.60\ Collect\ 6 + \\
& + 0.60\ Collect\ 7 \\
log(\hat{\sigma}_{ijkt}) = \hat{\eta}_{ijkt}{}^{\sigma} = \quad & -2.67 - 0.53\ Stand\ 2 + 1.48\ Stand\ 3 + \\
& + 0.44\ Height\ 4.5 + 0.26\ Height\ 7 + 0.44\ Collect\ 2 + \\
& + 0.26\ Collect\ 3 + 0.77\ Collect\ 4 + 0.94\ Collect\ 5 + \\
& + 0.41\ Collect\ 6 + 0.75\ Collect\ 7.
\end{aligned}
\tag{13}
$$

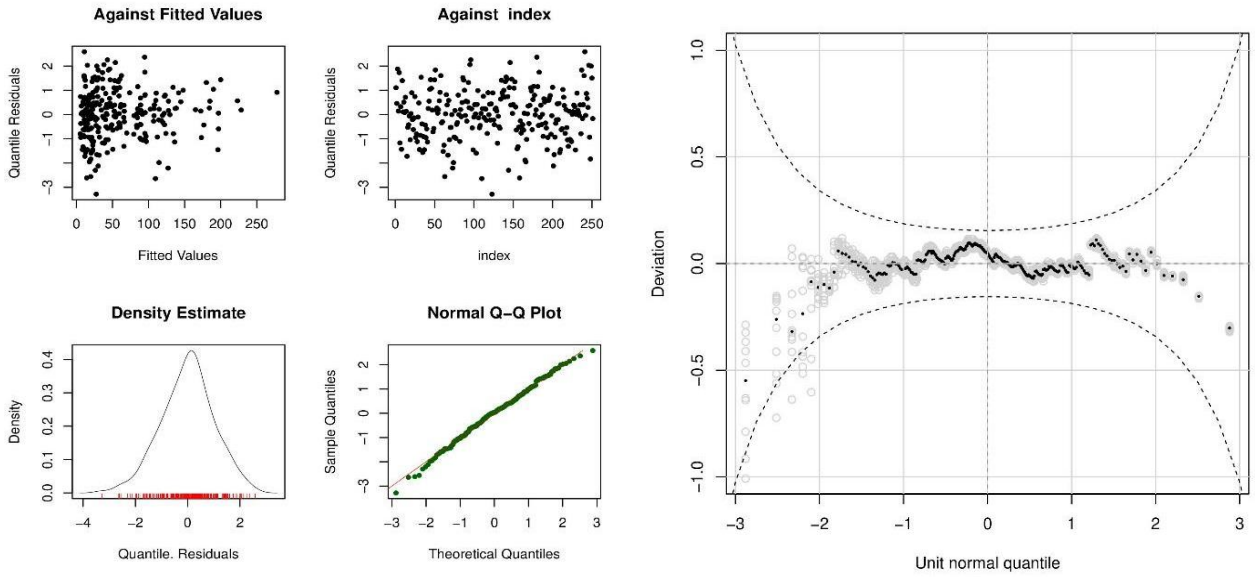


Figure 8. Residual plots of the fitted model, with linear predictors for μ and σ – final model.

Interpreting the regression coefficients of the model from Equation 13, about the baseline plot 1 and collection 1, negative effects were observed in the other plots for the predictor of μ, with the lowest average in plot 3, as previously observed in the exploratory analysis. The estimates of regression coefficients associated with the collection factor were all positive, indicating that the average count is lower in collection 1 and higher in collection 2, which had the highest coefficient.

Regarding the predictor of σ, from the estimates of the regression coefficients associated with the plot, only plot 3 was significant, which presented greater dispersion across the model. Only collection 5 was significant at 5% for height and collection, and all other coefficients were not significant. However, the presence of these factors was necessary for the model's suitability since their removal led to a loss of fit quality.

Conclusion

In this article, an actual application in the agricultural area was presented based on generalized linear mixed models and generalized additive models for location, scale, and shape to model the distribution of counts of insects of superfamily Psylloidea, vectors of the Huanglongbing disease. By using the Negative Binomial probability distribution, the flexibility of the GAMLSS model was essential to find a good fit since the variability differs with the levels of the factors, and the linear predictor for the scale parameter was necessary. Also, with GAMLSS, many other probability distributions can be used that may not belong to the exponential family of distributions. Thus, it is understood that this class of models is powerful and constitutes a viable alternative to classical regression models.

Acknowledgments

The authors are grateful to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for individual financial support to Pedroso, M. S.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. DOI: https://doi.org/10.1007/978-1-4612-1694-0_16

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9-25. DOI: https://doi.org/10.2307/2290687

Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, *11*(10), 1305-1319. DOI: https://doi.org/10.1002/sim.4780111005

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236-244. DOI: https://doi.org/10.2307/1390802

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-974.

Brooks, M. E., Kristensen, K., Benthem, K. J. V., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H.J., Maechler, M & Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378-400. DOI: https://doi.org/10.32614/RJ-2017-066

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *135*(3), 370-384.

Rigby, R. A., & Stasinopoulos, D. M. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, *6*, 57-65. DOI: https://doi.org/10.1007/bf00161574

Rigby, R. A., & Stasinopoulos, D. (2001). The GAMLSS project: A flexible approach to statistical modelling. In *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling* (Vol. 337, p. 345). University of Southern Denmark.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507-554. DOI: https://doi.org/10.1111/j.1467-9876.2005.00510.x

Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. New York, NY: Chapman and Hall. DOI: https://doi.org/10.1201/b21973