Ciência e Agrotecnologia
SCIENCE AND AGROTECHNOLOGY

Ciência e Agrotecnologia, 43:e020619, 2019
http://dx.doi.org/10.1590/1413-7054201943020619

Agricultural Sciences
eISSN 1413-7054

Eucalyptus cloeziana seed count data: a comparative analysis of statistical models

Dados de contagem em sementes de *Eucalyptus cloeziana*: uma análise comparativa entre modelos estatísticos

Thomas Bruno Michelon[1*], Cesar Augusto Taconeli[2], Elisa Serra Negra Vieira[3], Maristela Panobianco[1]

[1]Universidade Federal do Paraná/UFPR, Departamento de Fitotecnia e Fitossanidade, Curitiba, PR, Brasil
[2]Universidade Federal do Paraná/UFPR, Departamento de Estatística, Curitiba, PR, Brasil
[3]Empresa Brasileira de Pesquisa Agropecuária/Embrapa, Embrapa Florestas, Colombo, PR, Brasil
[*]Corresponding author: thomasbrunomichelon@gmail.com

Received in August 28, 2019 and approved in November 22, 2019

ABSTRACT

Generalized linear models (GLMs) are an extension of the linear model and include the normal, Poisson, and negative binomial distributions. Although GLMs were introduced in 1972, most seed technology studies, especially those involving count data, such as germination tests of seeds from the genus *Eucalyptus*, still using the analysis of variance, without analysis of the fit of other models. Thus, this study aimed to evaluate the most appropriate model in the GLM class for seed count data of *Eucalyptus cloeziana*. Data were obtained from a germination test using seeds from three lots of *E. cloeziana*. Each lot was separated by sieving into three material fractions based on size: small (<0.84 mm), medium (from 1.18 to 1.00 mm), and large (>1.18 mm). The data analysis was based on the use of GLMs adjusted to normal, Poisson, and negative binomial distributions, and the models were evaluated by the Akaike and Bayesian Schwartz criteria and Cook's distance and half-normal diagnostic graphs. Compared to other adjustments, the normal distribution adjustment differed in the configuration of means submitted to the Tukey test, and although the data met all normality assumptions, the adjustment with the Poisson distribution was the most suitable for the count data from a germination test of *E. cloeziana* seeds.

Index terms: Generalized linear models; Poisson; ANOVA; germination test.

RESUMO

Modelos Lineares Generalizados (GLM) são uma extensão do modelo linear e englobam às distribuições normal, Poisson e binomial negativa. Apesar de terem sido introduzidos em 1972, na maioria dos estudos em tecnologia de sementes, especialmente naqueles que envolvem dados de contagem, como o teste de germinação em sementes do gênero *Eucalyptus*, ainda predomina a Análise de Variância, sem análise do ajuste de outros modelos. Assim, o objetivo desse trabalho foi avaliar o modelo mais adequado, na classe dos GLM, para dados de contagem de sementes de *Eucalyptus cloeziana*. Os dados foram obtidos a partir da realização de teste de germinação, utilizando-se sementes de três lotes de *Eucalyptus cloeziana*, sendo que cada lote foi separado por meio de peneiras em três frações de material, com base no seu tamanho: pequeno (<0,84 mm), médio (entre 1,18 e 1,00 mm) e grande (>1,18 mm). A análise dos dados foi fundamentada na utilização dos GLM ajustados à distribuição normal, Poisson e binomial negativa, analisando-se ajuste pelos critérios de Akaike e Bayesiano de Schwarz, pelos gráficos de diagnóstico da distância de Cook, Half-normal. O ajuste de distribuição normal diferiu na configuração de médias submetidas ao teste de Tukey em relação aos outros ajustes e, apesar dos dados atenderem todos os pressupostos de normalidade, pode-se concluir que o ajuste com a distribuição de Poisson foi o que mais se adequou aos dados de contagem do teste de germinação de sementes de *E. cloeziana*.

Termos para indexação: Modelos lineares generalizados; Poisson; ANOVA; teste de germinação.

INTRODUCTION

Count data can be conceptualized as representing how often a specific event occurs, resulting in non-negative integer values (Kosma et al., 2019). In the biological sciences, count data are produced in a variety of experiments, such as the analysis of fungal colonies (Pereira et al., 2016), counting of weed species in a given area (Heap; Duke, 2017), and counting of the number of leaves per seedling (Silva et al., 2019).

In seed analysis, the viability of a lot can be determined by a germination test, which according to the evaluated species, can be conducted by repeatedly measuring a fixed number of seeds (expressed as a percentage of normal seedlings) or using a specific seed weight (number of normal seedlings), the latter of which is recommended for *Eucalyptus* seeds, the most cultivated forest genus in Brazil, which accounts for approximately 73% of the total planted forest area (Ibá, 2019).

Among *Eucalyptus* species, *Eucalyptus cloeziana* stands out for its strong and durable wood, which presents a greater density than the wood of other species within the genus and is commonly used in construction and high-added value products, such as furniture (Hicks; Clark, 2001; Boland et al., 2006; Li et al., 2017).

In seed technology studies, analysis of variance (ANOVA) is the most commonly used statistical method for data evaluation by researchers. However, there is an orthodoxy in the choice of the model, in which the ANOVA model is preferable because it is more traditional rather than because it presents a better fit (Sileshi, 2012; Santana; Carvalho; Toorop, 2018).

ANOVA requires that the following assumptions be met: homogeneity of variances, normally distributed residuals and independent distributions. However, the data produced by experiments involving count data in the biological sciences often do not meet such assumptions (St-Pierre; Shikon; Schneider, 2018; Kosma et al., 2019), as seen in studies of seeds of genetically impoverished species, which present a high variability rate, such as forest seeds (Carvalho; Santana; Araújo, 2018). Sileshi (2012) showed that in 429 studies with seed viability data published from 2002 to 2012, 70% of researchers used ANOVA to analyse the data, and among those who did, only 20% performed tests to verify the homogeneity of variances and normality of residuals. Such disregard can lead to inflation of the probability of type 1 errors, in which case a real difference between 2 treatments is not detected and the null hypothesis is erroneously accepted (Kikvidze; Moya-Laraño, 2008).

One way around the problem of non-normality is through data transformation. However, a particular transformation is not always available to satisfy all the assumptions of the analysis. In addition, data transformation creates difficulty in interpreting results due to a change in scale (St-Pierre; Shikon; Schneider, 2018). Thus, the reformulation of the statistical model at the expense of data transformation is an advantageous solution.

In this context, generalized linear models (GLMs) appear to be a general alternative to ANOVA in the analysis of data from seed germination experiments. For these cases, in which the response is a percentage, Santana, Carvalho and Toorop (2018), using *Lychnophora ericoides* seeds, and Carvalho, Santana and Araújo (2018), using copaiba seeds, found that a GLM with a binomial distribution provided an appropriate adjustment of the data, even when the ANOVA assumptions were met.

Nelder and Wedderburn (1972) introduced the GLM class as an extension of linear models, applied to data belonging to the exponential family, such as those with normal, Poisson, and negative binomial distributions. A GLM is composed of a random component, referring to as the response variable; a systematic component, including the explanatory variables; and a link function that connects the components. Although the theory of generalized linear models was first presented in 1970, by 2012, only 13% of germination studies used these models. In contrast, the use was approximately 52% for non-normal count data in other areas of biological science (Sileshi, 2012; St-Pierre; Shikon; Schneider, 2018).

In GLM theory, Poisson and negative binomial distributions are usually applied in count data analysis. The Poisson distribution is usually the first option for analysis, but this distribution assumes that $E[Y_i] = Var[Y_i] = \mu_i$, where Y_i (i = 1, 2, …, n) is the dependent variable; i.e., the data variance must be equal to the mean. When the variance is greater than the mean, a phenomenon called overdispersion occurs, making the use of this model unfeasible. In this sense, alternative models can be used, such as the negative binomial model (Hinde; Demétrio, 1998). The better adjusted the model is, the more consistent the inferences made are. Thus, this study aimed to identify the most appropriate model in the GLM class for *Eucalyptus cloeziana* seed count data.

MATERIAL AND METHODS

The activities were carried out at the Laboratories of Seed Analysis of the Federal University of Paraná, Curitiba, and Embrapa Forestry, Colombo, Paraná, Brazil.

The data were obtained from a germination test conducted with three lots of *E. cloeziana* seeds (harvested in 2018) supplied by Embrapa Florestas, originally collected from an experimental plot of 4 ha in the municipality of Antônio João (Mato Grosso do Sul, Brazil).The seeds were classified by size with the aid of sieves and assigned to four treatments: non-processed material (control) and processed material categorized by size class, namely, small (<0.84 mm), medium (from 1.00 to 1.18 mm), and large (>1.18 mm).

A completely randomized design with four replications was used in a 4 × 3 factorial scheme. The first factor was related to processing (T1 – control; T2 – small; T3 – medium; and T4 – large), and the second factor, to lot (L1, L2, and L3), totalling 48 experimental units.

Sowing was carried out with four replications of 0.5 grams of propagation material in transparent plastic boxes (11.0 × 11.0 × 3.5 cm) with sand substrate previously sterilized and moistened with water to reach 50% field capacity. These boxes were placed in a Mangelsdorf germinator at 25 °C under continuous light. The first and last counts of normal seedlings were performed 14 and 21 days after sowing, respectively (Brasil, 2009).

The analysis was based on the use of generalized linear models (GLMs), which include the normal, Poisson, and negative binomial distributions. These distributions belong to the parametric exponential family, with the probability density function (Lamb; Demetrius, 2013) $f(y;\theta,\phi) = \exp\left[\frac{(y(\theta)-b(\theta))}{a(\phi)} + c(y;\phi)\right]$, where θ is the canonical parameter of the distribution, ϕ is the dispersion parameter, and *a*(.), *b*(.), and *c*(.) are the real functions referring to each distribution.

Regarding the normal, Poisson, and negative binomial distributions, the probability density functions are defined by $f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{(y-\mu)^2}{\sigma}\right]$, $f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$, and $f(y;\mu,\phi) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}\left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)$, respectively, where in the context of this study, y is the number of normal seedlings, μ is the mean number of normal seedlings, σ^2 is the variance, and $\pi \approx 3.1416$.

The number of normal seedlings comprises the random component, while the systematic component of the three models is related to the factors processing, lot, and their interaction, with their linear effects combined by $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \sim x_i^T\beta$ or $\eta = X\beta$, where $X = (x_1, x_2, ..., x_n)^T$ is the model matrix composed of the variables indicating sieve type, lot, and their interaction; x_i^T is the *i*-th row of experimental matrix X; $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ is the parameter vector for the model; and $\eta = (\eta_1, \eta_2, ..., \eta_n)^T$ is the linear predictor.

The connection between the random and systematic components is made by the link function and represents how the effects of the experimental factors impact the mean of y, being $\eta_i = g(\mu_i) = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})$, where g(.) is the monotonic differentiable real function.

The identity link function was used for the normal distribution, represented by $g(\mu_i) = \mu_i$. The assumptions of normality of residuals and stability of variance were evaluated for this model by the Shapiro-Wilk and Levene tests, respectively. The logarithmic link function $g(\mu_i) = In\ (\mu_i)$ was used for the Poisson and negative binomial distributions.

The Akaike information content (AIC) and Bayesian (Schwartz) information content (BIC) criteria were considered to identify the model with the best fit among the evaluated distributions (Cordeiro; Demétrio, 2013). These criteria penalize models of greater complexity or poorly adjusted models. Thus, the lower the values of the criteria are, the higher the evidence of adequacy is, which is defined as $AIC = -2log\ L + 2p$ and $BIC = -2log\ L + p\ log\ (n)$, where p is the number of parameters in the model, L is the logarithm of the maximized likelihood under each model, and n is the number of observations (BIC provides a greater penalty than AIC as n increases).

The influence of other variables on the response variable was evaluated by the residual deviance (Cordeiro; Demétrio, 2013) to compare models of different complexities by adding predictors.

In addition to the information criteria, the adjustments provided by the models were analysed graphically with residual and half-normal simulated envelope graphs, and Cook's distance was used to identify possible influential data.

The best-fitting model was submitted to hypothesis testing using the residual deviance value and a χ^2 distribution at a 0.05 probability, in which the null hypothesis was a good fit of the model to the data.

Tukey's test was used to compare the means, considering a significance level of 0.05. All analyses were performed using the software R version 3.5.2.

RESULTS AND DISCUSSION

The AIC and BIC values compared between the three models are shown in Table 1. The Poisson model had the best indices, with the lowest AIC (326.06) and BIC (348.51), followed by the negative binomial model, with an AIC of 328.06 and a BIC of 352.39. In comparison, the normal model had the highest AIC (331.91) and BIC (356.23) indices. Thus, considering that simpler models better explain the data than more complex models (AIC and BIC, 2004; Carvalho; Santana; Araújo, 2018) and the normal distribution was expressed in a less general way, this distribution incurred a higher penalty due to the presence of additional parameters that made the model more complex and, consequently, a worse estimator.

Table 1: Values of the Akaike (AIC) and Bayesian (BIC) information criteria, degrees of freedom, and residual deviance for different normal, Poisson, and negative binomial distribution models.

Model	Degrees of Freedom	*Deviance* Residual	P-value[1]	AIC	BIC
		Poisson			
Null	47	888.10	-	1150.14	1152.01
Treatment	44	334.85	2.20E-16**	602.89	610.37
Treatment + Lot	42	94.05	2.20E-16**	366.09	377.32
Treatment * Lot	36	42.02	1.84E-09**	326.06	348.51
		Negative Binomial			
Null	47	51.243	-	453.62	457.36
Treatment	44	51.225	9.69E-09**	419.43	428.78
Treatment + Lot	42	52.132	4.33E-15**	357.28	370.38
Treatment * Lot	36	42.016	2.62E-07**	328.06	352.39
		Normal			
Null	47	39437	-	462.36	466.1
Treatment	44	13719	2.20E-16**	417.67	427.03
Treatment + Lot	42	3557	3.97E-16**	356.89	369.99
Treatment * Lot	36	1646	5.70E-05**	331.91	356.23

[1]p-value based on a X^2 test for the Poisson and negative binomial distributions and an F-test for the normal distribution.
**Significant at a 0.01 probability.

The normal model produced the worst fit when compared to the other models, even with the residuals presenting a normal distribution based on the Shapiro-Wilk test (W = 0.961; P = 0.115) and homoscedastic variances based on Levene's test (F = 1.502; P = 0.174). This result demonstrates that even if the data meet the assumptions of normality, the normal distribution is not always the distribution that best represents them, which is corroborated by a study conducted by Carvalho, Santana, and Araújo (2018) using copaiba seed germination data, in which a GLM with a binomial distribution fit best, even when the assumptions of linear models were met. Similarly, Sileshi (2012) used non-normal rapeseed data from Piepho (2003) and reported better performance with a GLM than with the arcsine transformation $\left(\sqrt{y/100}\right)$ of the data.

Although the AIC and BIC criteria are efficient in selecting models, they are not able to discriminate data overdispersion effects, which makes the Poisson distribution unfeasible. Consequently, although the Poisson distribution was better adjusted, it was essential to check this possibility.

Diagnosis by graphical analysis of residuals versus adjusted values (Figures 1A, 2A, and 3A) should take into account the random scattering of points around zero and the absence of extreme values (Kozak; Piepho 2017). Both the Poisson and negative binomial models, as well as the normal model, presented ungrouped points, i.e., points not tending to fall within a specific area, and no outliers, thus reflecting a good fit of the data to the respective distributions. Residual analysis is also widely used to verify the relationship between the variance and mean of the distribution, and its use is particularly advantageous in identifying overdispersion in the data (McCullagh; Nelder, 1989; Stroup, 2015). Figure 2A shows no evidence of overdispersion in the Poisson model.

Half-normal graphs (Figures 1B, 2B, and 3B) provide a visual analysis of how residuals are distributed, allowing an evaluation of their adherence to a normal distribution and identification of potential outliers (Kozak; Piepho, 2017). This graphical analysis was efficiently used to select different statistical models by Santana, Carvalho and Toorop (2018), who used data of *Lychnophora ericoides* seed germination.

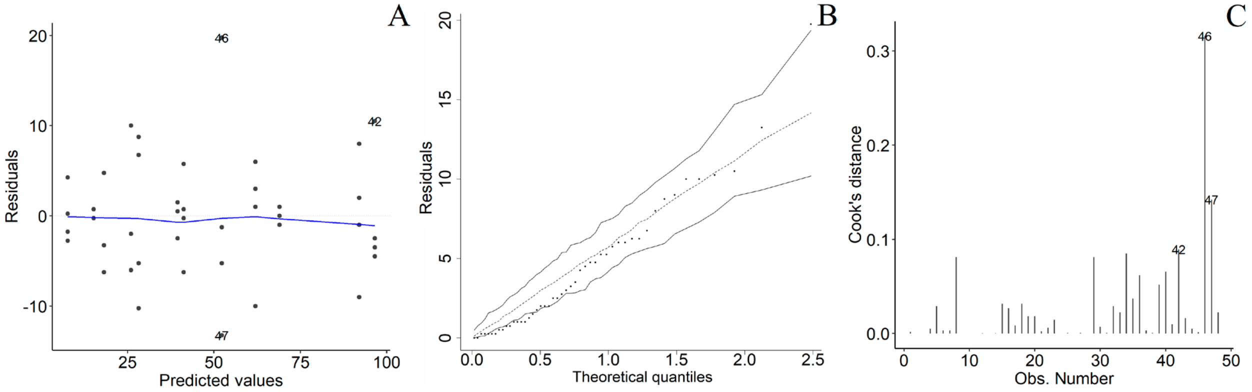

Figure 1: Diagnostic graphs of the normal model: residuals versus adjusted values (A); half-normal plot (B); and Cook's distance (C).

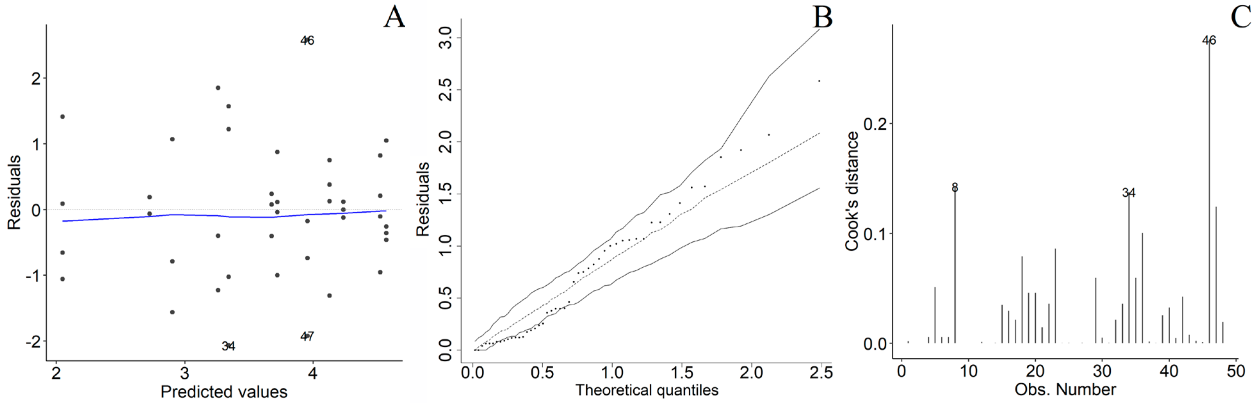

Figure 2: Diagnostic graphs of the Poisson model: residuals versus adjusted values (A); half-normal plot (B); and Cook's distance (C).

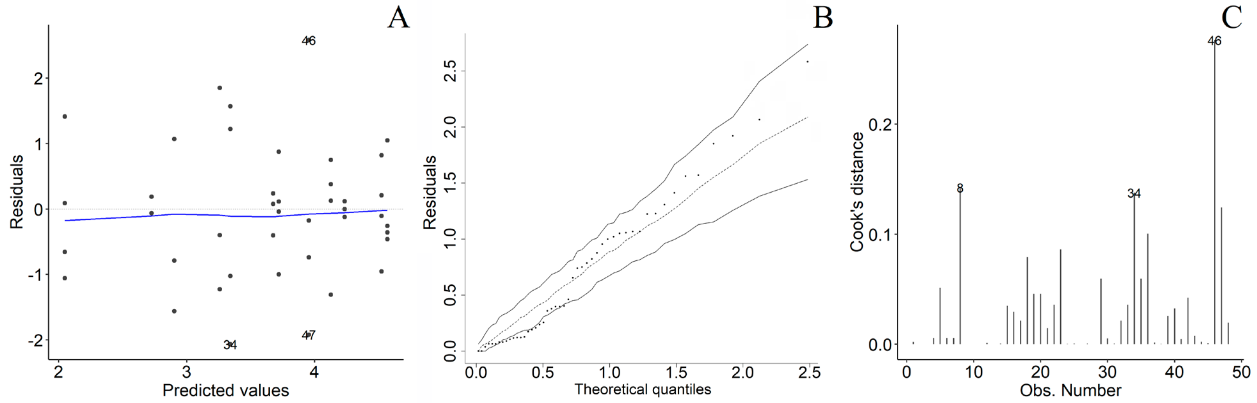

Figure 3: Diagnostic graphs of the negative binomial model: residuals versus adjusted values (A); half-normal plot (B); and Cook's distance (C).

As in the residual analysis, the three models presented adequate fits of the data to the distribution based on their half-normal graphs, with a few incongruent points, which may be characterized as outliers (discrepant points) but are not necessarily due to the lack of adjustment, which can be further analysed by graphical evaluation of Cook's distance. The similarities between these three distributions, as evidenced by the graphical analysis, can be explained by the Poisson and negative binomial distributions tending to approximate a normal distribution as the count increases (Stroup, 2015), as in the present study.

As with the half-normal graphs, Cook's distance (Figures 1C, 2C, and 3C) should be considered when evaluating the distribution of data in relation to the model. The larger the number of influential points a model has, the poorer the fit is (Altman; Krzywinski, 2016; Carvalho; Santana; Araújo, 2018). In addition to the distribution, the presence of extreme points can also be evaluated. Specifically, Cook and Weisberg (1982) consider points with values above 1 on the y-axis to be extreme.

Although there were no points above 1, point 46 (related to the mean of the large material in lot 3) was separated from the others in the three models and could still be considered an influential point (Bollen; Jackman, 1985). This may be due to an effect of some uncontrolled experimental factor. Similarly, other authors have relied on such analysis to select predictive models and remove inconsistent points (Jagadeeswari; Harini; Kumar, 2013; Mihalovits et al., 2019).

In seed analysis, the Tukey test is traditionally applied following analysis of variance to identify differences between pairs of treatment means (Sileshi, 2012). However, the same data may assume different configurations for each distribution since the calculation is based on the standard error of the means.

The interaction effect was significant ($p<0.05$) in the three models, while a difference in the arrangement of means between the Poisson and negative binomial models in relation to the normal distribution model was observed after the Tukey test was applied (Table 2), in which the normal distribution revealed statistically equal means between the control and treatment 2 (for lot 3) but differences between the other treatment pairs. Similar behaviours occurred for lots 2 and 3 within the control and treatment 2. This result shows that the Tukey test had lower sensitivity in differentiating means for the normal model, which had the lowest indices of fit to the data among the models.

Warton et al. (2016) considered some obstacles in the selection of models for count data related to the effective capture of data characteristics and type 1 error control. In this case, although GLMs with Poisson and negative binomial distributions should be prioritized over the linear model, the final model specification is based on the selection and diagnosis stages of the adjustment.

Thus, a flowchart (Figure 4) is proposed to illustrate the main steps in the selection and evaluation process, which can be extrapolated to several experimental situations involving counts. The quality of the adjustment provided by the Poisson model can be confirmed at the end of the process by a hypothesis test based on the distribution for the residual deviance (42.019) with 36 degrees of freedom (Table 1), in which case the hypothesis that the model fits the data well is not rejected at a 5% probability (p-value = 0.22).

In this sense, the largest propagation material generated a larger number of normal seedlings of *E. cloeziana* (Figure 5), whereas seeds smaller than 0.84 mm did not differ statistically from the control.

Studies involving seed counts, such as those of *Eucalyptus* germination, are diverse in the literature (Sousa et al., 2018; Sá-Martins et al., 2019; Nega; Gudeta, 2019), and despite the evolution of research in these areas, statistical analysis of the data mostly follows the traditional pattern, and few studies address the adequacy of different statistical models in the experimental situation. Thus, the present study provides a critical view of the evaluation of experimental data, demonstrating alternative forms of analysis, possibly more suitable for seed analysis since there is not only one method that fits all situations.

Table 2: Partitioning of the effect of the lot × treatment interaction for models adjusted to each distribution.

Factor 1	Factor 2	Estimate of Marginal Means	Standard Error			Tukey[1]		
			Poisson	Negative binomial	Normal	Poisson	Negative binomial	Normal
L1	T4	92	4.80	4.80	3.38	a	a	a
	T3	69	4.15	4.15	3.38	b	b	b
	T2	41	3.21	3.21	3.38	c	c	c
	T1	40	3.14	3.14	3.38	c	c	c
L2	T4	97	4.91	4.91	3.38	a	a	a
	T3	62	3.94	3.94	3.38	b	b	b
	T1	26	2.55	2.55	3.38	c	c	c
	T2	18	2.14	2.14	3.38	c	c	c
L3	T4	52	3.61	3.61	3.38	a	a	a
	T3	28	2.66	2.66	3.38	b	b	b
	T1	15	1.95	1.95	3.38	c	c	c
	T2	8	1.39	1.39	3.38	d	d	c
T1	L1	40	3.14	3.14	3.38	a	a	a
	L2	26	2.55	2.55	3.38	b	b	b
	L3	15	1.95	1.95	3.38	c	c	b
T2	L1	41	3.21	3.21	3.38	a	a	a
	L2	18	2.14	2.14	3.38	b	b	b
	L3	8	1.39	1.39	3.38	c	c	b
T3	L1	69	4.15	4.15	3.38	a	a	a
	L2	62	3.94	3.94	3.38	a	a	a
	L3	28	2.66	2.66	3.38	b	b	b
T4	L2	97	4.80	4.80	3.38	a	a	a
	L1	92	4.91	4.91	3.38	a	a	a
	L3	52	3.61	3.61	3.38	b	b	b

[1] Significant at a 0.05 probability.

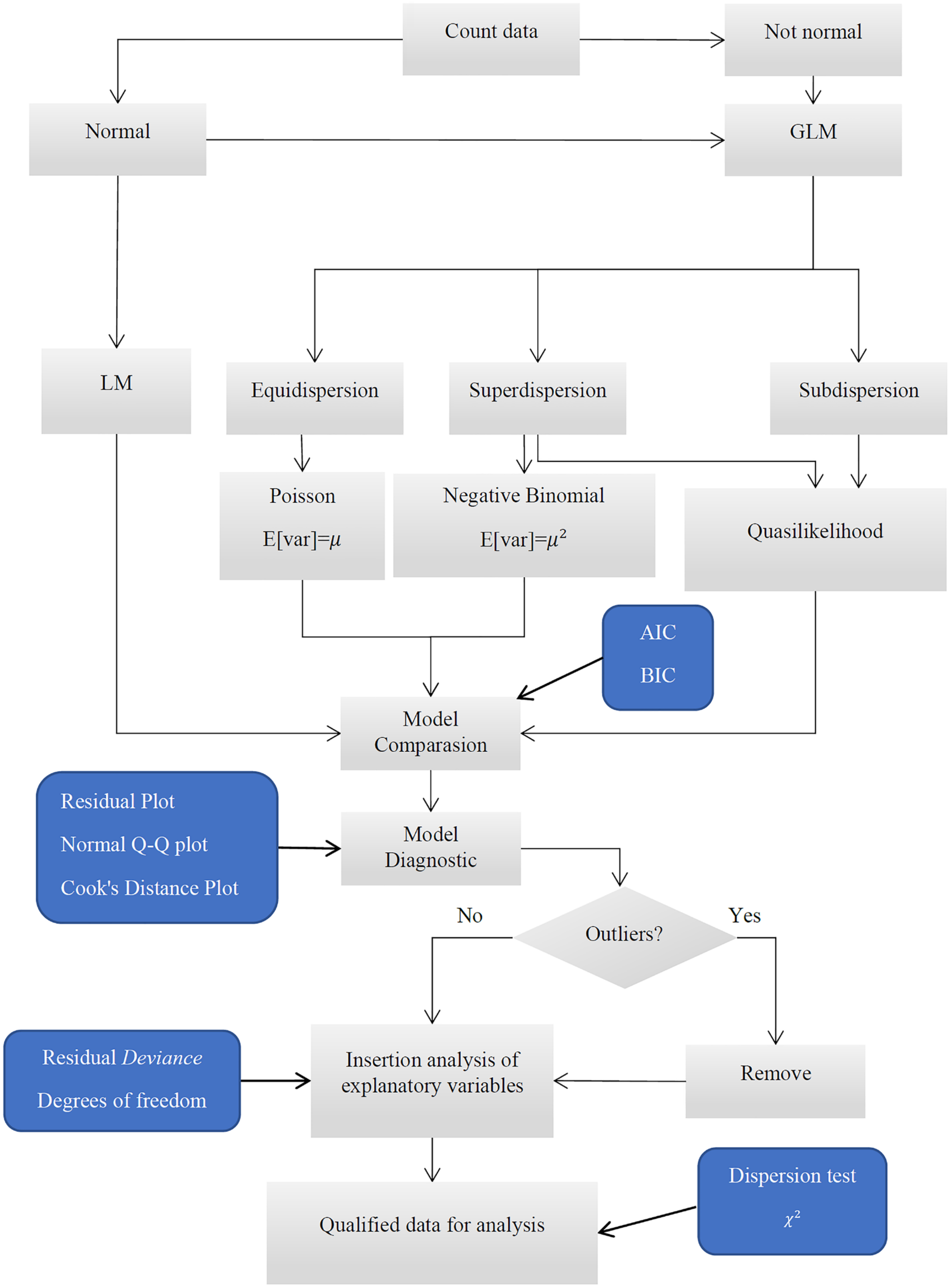

Figure 4: Flowchart developed for statistical analysis of seed count data of *Eucalyptus cloeziana*.

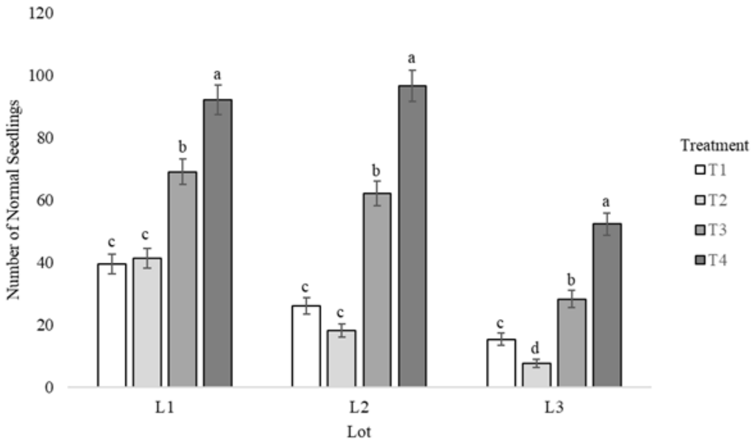

Figure 5: Comparison of means by the Tukey test at a 0.05 probability as a function of the generalized linear model with the Poisson distribution.

CONCLUSION

The generalized linear model with a Poisson distribution provides the best fit to seed count data of *E. cloeziana*.

REFERENCES

AIC and BIC: Comparision of assumptions and performance. **Sociological Methods & Research**, 33(2):188-229, 2004.

ALTMAN, N.; KRZYWINSKI, M. Analyzing outliers: Influential or nuisance? **Nature Methods**, 13(4):281-282, 2016.

BRASIL. **Regras para Análise de Sementes**. Brasília: Ministério da Agricultura, Pecuária e Abastecimento, 2009. 399p.

BOLLEN, K. A.; JACKMAN, R. W. Regression diagnostics: An expository treatment of outliers and influential cases. **Sociological Methods & Research**, 13(4):510-542, 1985.

CARVALHO, F. J.; SANTANA, D. G. de; ARAÚJO, L. B. de. Why analyze germination experiments using Generalized Linear Models? **Journal of Seed Science**, 40(3):281-287, 2018.

COOK, R. D.; WEISBERG, S. **Residuals and influence in regression**. London, 1982. 229p.

CORDEIRO, M. G.; DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados e Extenções**. São Paulo, 2013. 483p.

HEAP, I.; DUKE, S. O. Overview of glyphosate-resistant weeds worldwide. **Pest Management Science**, 74(5):1040-1049, 2017.

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: Models and estimation. **Computational Statistics & Data Analysis**, 27(2):151-170, 1998.

HICKS, C. C; CLARK, N. B. **Pulpwood Quality of 13 Eucalypt Species with Potential for Farm Forestry.** Australia: RIRDC, 2001. 44p.

INDUSTRIA BRASILEIRA DE ÁRVORES. **Relatório Ibá 2019.** São Paulo, 2019. 80p.

JAGADEESWARI, T.; HARINI, N.; SATYA KUMAR, C. Identification of outliers by cook's distance in agriculture datasets. **International Journal Of Engineering And Computer Science**, 2(6):2045-2049, 2013.

KIKVIDZE, Z.; MOYA-LARAÑO, J. Unexpected failures of recommended tests in basic statistical analyses of ecological data. **Web Ecology**, 8:67-73, 2008.

KOSMA, M. et al. Over dispersed count data in crop and agronomy research. **Journal of Agronomy and Crop Science**, 205(4):414-421, 2019.

KOZAK, M.; PIEPHO, H. P. What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. **Journal Of Agronomy And Crop Science**, 204(1):86-98, 2017.

LI, C. et al. Genetic parameters for growth and wood mechanical properties in *Eucalyptus cloeziana* F. Muell. **New Forests**, 48(1):33-49, 2017.

MCCULLAGH, P.; NELDER J. A. **Generalized linear models**. London, 1989. 506p.

MIHALOVITS, M. et al. Model Building on selectivity of gas antisolvent fractionation method using the solubility parameter. **Periodica Polytechnica Chemical Engineering**, 63(2):294-302, 2019.

NEGA, F.; GUDETA, T. B. Allelopathic effect of *Eucalyptus globulus* Labill. on seed germination and seedling growth of highland Teff [*Eragrostis tef* (Zuccagni) Trotter)] and Barley (*Hordeum vulgare* L.). **Journal of Experimental Agriculture International**, 30(4):1-12, 2019.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society**, 135(3):370-384, 1972.

PEREIRA, J. S. et al. Comparative analysis of fungal communities in colonies of two leaf-cutting ant species with different substratum preferences. **Fungal Ecology**, 21:68-75, 2016.

PIEPHO, H. P. The folded exponential transformation for proportions. **Journal of the Royal Statistical Society**, 52(4):575-589, 2003.

SANTANA, D. G. de; CARVALHO, F. J.; TOOROP, P. How to analyze germination of species with empty seeds using contemporary statistical methods? **Acta Botanica Brasilica**, 32(2):271-278, 2018.

SÁ-MARTINS, R. de. et al. Effect of water and salt stress on seeds germination and vigor of different eucalyptus species. **Journal Of Tropical Forest Science**, 31(1):12-18, 2019.

SILESHI, G. W. A critique of current trends in the statistical analysis of seed germination and viability data. **Seed Science Research**, 22(3):145-159, 2012.

SILVA, E. M. et al. Leaf count overdispersion in coffee seedlings. **Ciência Rural**, 49(9):e20180786, 2019. Available in: <http://www.scielo.br/pdf/cr/v49n4/1678-4596-cr-49-04-e20180786.pdf>. Access in: February, 10, 2019.

SOUSA, M. V. de. et al. Allelopathy of the leaf extract of eucalyptus genetic material on the physiological performance of millet seeds. **American Journal Of Plant Sciences**, 9(1):34-45, 2018.

STROUP, W. W. Rethinking the analysis of Non-Normal Data in plant and soil science. **Agronomy Journal**, 107(2):811-827, 2015.

ST-PIERRE, A. P.; SHIKON, V.; SCHNEIDER, D. C. Count data in biology - Data transformation or model reformation? **Ecology and Evolution**, 8(6):3077-3085, 2018.

WARTON, D. I. et al. Three points to consider when choosing a LM or GLM test for count data. **Methods In Ecology And Evolution**, 7(8):882-890, 2016.