Acta Scientiarum

http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v47i1.71591

BIOMETRY, MODELLING AND STATISTIC

Alternatives for simulating and modeling simplified insect feeding eletropenetrography discrete data

Guilherme Pereira de Oliveira[1*], **Frederico Antonio Loureiro Soares**[1], **André Cirilo de Sousa Almeida**[2], **Gustavo da Silva Leite**[2], **Timothy Aaron Ebert**[3] and **Anderson Rodrigo da Silva**[4]

[1]Departamento de Ciências Agrárias, Instituto Federal de Goiás, Campus Rio Verde, Rodovia Sul Goiana, Km 01, 75901-970, Rio Verde, Goiás, Brazil. [2]Departamento de Agronomia, Instituto Federal de Goiás, Campus Urutaí, Urutaí, Goiás, Brazil. [3]Department of Entomology and Nematology, University of Florida/IFAS Citrus Research and Education Center, Lake Alfred, Florida, United States. [4]Laboratório de Estatística e Geoprocessamento, Instituto Federal de Goiás, Campus Urutaí, Urutaí, Goiás, Brazil. *Author for correspondence. E-mail: guilherme.pereira.oliveira@hotmail.com

ABSTRACT. The study of insect feeding behavior using electropenetrography (EPG) typically involves analyzing complex data. EPG data comprises a temporal sequence of behaviors summarized using a collection of counts, durations, and sequential variables. These variables can be counts, means, percentages, or linear combinations of behaviors. This results in numerous variables being correlated to a certain degree. Consequently, statistical analysis is rendered complex, particularly in terms of model fitting and selection. This study proposed a statistical approach to simulate overdispersed correlated count data based on a previous comparative experiment to monitor the feeding behavior of untreated *Euschistus heros* versus *E. heros* treated with an entomopathogen. The waveforms included non-feeding (Z), pathway (Eh1), laceration/maceration of endosperm tissue (Eh3a), short ingestion events of lacerated/macerated endosperm tissue (Eh3b), xylem sap ingestion (Eh2), and ingestion from an unknown location (Eh4). Simulated scenarios involved the creation of differences between groups of insects based on the total number of events or the proportion of events of Z. Several statistical models were then fitted to the simulated data and evaluated based on goodness-of-fit, type-I error rate, and power analysis. The multinomial model exhibited the lowest type-I error rate and was more sensitive in detecting higher (>1.35x) differences between groups. Only the multinomial model achieved a power greater than 0.8. Conversely, models such as the Poisson and normal models exhibited limitations such as inflated type-I error rates in the presence of overdispersion. Among the univariate models, the mixed model exhibited the best fit.

Keywords: EPG; multinomial model; double-poisson.

Received on March 11, 2024.
Accepted on July 30, 2024.

Introduction

Several pests affect soybeans, among which the neotropical brown stink bug (*Euschistus heros* (Fabricius), Hemiptera: Pentatomidae) has garnered attention because of its high losses (Sosa-Gómez et al., 2020). The feeding damage caused by sucking insects on pods can result in losses exceeding 30% (Antúnez et al., 2022). There are numerous challenges to understanding the biology of sucking insects, and elucidating this information will enable the development of more efficient management tools.

In sucking insects, the observance of feeding activities with the naked eye is challenging because they occur inside the opaque host tissue. This renders the identification of initial symptoms difficult. Through a technology called electropenetrography (EPG) (Mclean & Kinsey, 1964), detailed insights into the feeding behavior of this group of insects can be obtained by analyzing voltage pattern data from an electrical circuit that includes both insects and plants (Backus et al., 2019).

In experiments with *E. heros* using EPG, we can monitor and characterize EPG waveforms, determine specific feeding sites, and ascertain the biological significance of waveforms based on their electrical characteristics and histological correlations (Lucini & Panizzi, 2018). However, although these feeding events are intrinsically related, they are generally treated separately. This facilitates the exploitation or underutilization of data.

The EPG generates a large amount of data with various types of variables of duration and number of waveform events. It is common for these data to have excess zeros, which renders statistical analysis complex, particularly in terms of model fitting and the selection of the most important variables (Ying et al., 2021; Hu et al., 2020; Lucini & Panizzi, 2017). Several programs are used to read the data generated by EPG, such as Backus 1.0 (Backus et al., 2007) EPG-Calc (Giordanengo, 2014), Sarria Workbook (Sarria et al., 2009), Ebert 1.0 (Ebert et al., 2015), and INFEST (Silva et al., 2022). The data read by these programs can be used to build statistical models to elucidate issues related to pest biology (Ying et al., 2021; Hu et al., 2020; Lucini & Panizzi, 2017).

There are numerous approaches to evaluating this large volume of data, ranging from simpler analyses, such as linear models and their transformations, to complex models, such as generalized linear models (GLM), generalized additive models for location, scale, and shape (GAMLSS), and even multinomial models (Freitas & Duarte, 2023; Schmidt et al., 2022; Rigby & Stasinopoulos, 2005). To assess the accuracy of these models, the Akaike Information Criterion (AIC), power analysis, and type-I error rate must be examined (Sakamoto et al., 1986).

One method to evaluate the model performance and importance of variables involves data simulation based on sets of rules or probabilistic distributions observed in previous experiments (Tyralis & Papacharalampous, 2024). However, studies involving simulations of EPG data are scarce. This may be because of the nature of the raw data generated by the equipment, which comprises recordings with voltage information (Backus & Shih, 2020). Consequently, the identification and classification of voltage patterns in specific types of waves are time-consuming, and the process requires a highly trained and skilled user. Moreover, variables generated by EPG generally exhibit a correlation, which must be considered in the simulation process.

Thus, this study aimed to simulate and model discrete feeding EPG data for *Euschistus heros* and compare these models based on goodness-of-fit, type-I error rate, and power analysis.

Material and methods

Simulation strategy and scenarios

The simulation scenarios were based on an EPG study (Rodrigues, 2023) to monitor the behavior of *Euschistus heros* (Fabricius) (Hemiptera: Heteroptera) feeding on soybean pods in two groups of 16 insects each (control and treatment: application of the entomopathogenic fungus *Metarhizium anisopliae*). The recordings were completed after 72h. The following waveform events were recorded: Z (non-feeding), Eh1 (pathway), Eh3a (laceration/maceration of the endosperm), Eh3b (short ingestion event of lacerated/macerated endosperm tissue), Eh2 (xylem sap ingestion), and Eh4 (ingestion from an unknown location). From the experimental data on the number of waveform events by insects, 10,000 data simulations were performed.

Let $y = [y_1\ y_2 \dots y_k]$ represent the k-dimensional vector of the number of events per waveform by an insect (NWEi), which is mutually exclusive. Let $N = \sum_{k=1}^{K} y_k$ be the total number of insect events during the entire recording period. As previously reported (Terza & Wilson, 1990; Schmidt et al., 2022), we considered the conditional probability distribution of y given N as

$y|N \sim Multinomial(N,\pi)$

where: $y_k = 0,\ 1,\ 2,\dots,\infty$; and is $\pi = [\pi_1\ \pi_2 \dots\ \pi_k]$ the k-dimensional vector of parameters representing the multinomial probabilities, with $\sum_{k=1}^{K} \pi_k = 1$.

Because N is not expected to be homogeneous across insects, we considered it a random variable that could be modelled by a discrete probability model, say, $h(.)$,

$N \sim h(\mu,\sigma)$

where: μ represents the mean and σ the dispersion parameter; $N = 0,\ 1,\ 2,\dots,\infty$. Here, a natural choice for $h(.)$ is *Poisson*$(\mu,\sigma=1)$. Because N can be affected by factors that may vary among insects, such as recording time and treatment, it may be overdispersed. Thus, its observed variability is greater than that captured by parametric models such as *Poisson*. Alternatively, examples of models that can accommodate overdispersion are *Double-Poisson*(μ,σ) and *Gamma-Poisson*(μ,σ), because both have $\sigma > 0$. We computed the estimates of the maximum likelihood (EMV) of μ and σ and used Akaike's information criterion (AIC) to select the best-fitting model.

From the joint distribution of y and N, we obtain

$Pr(y,N \mid \mu,\sigma,\pi) = Pr(y \mid N,\pi)\ Pr(N \mid \mu,\sigma)$

We simulated 10,000 data matrices *Y* with dimensions 32 x 16 (*n* insects × *K* waveforms) under the null hypothesis of no difference between experimental groups, that is, H_0: $\theta_1 = \theta_2$; where θ: $\{\mu,\sigma,\pi\}$. This was done to evaluate the type-I error rate of statistical tests as one criterion to evaluate and select regression models. The next section describes this calculation and that of the power analyses.

Regression models, type-I error rate and power analysis

After simulating *Y*, the univariate regression models were fitted to the selected response variable y_1, corresponding to waveform Z (no feeding activity), as presented in Table 1.

Table 1. Regression models.

Model	Type*	Mean/Linear Predictor**	Dispersion***
Normal	Linear	$\mu_i=\beta_0+\tau_i$	σ
Normal with sqrt transformation	Linear	$\sqrt{\mu_i} =\beta_0+\tau_i$	σ
Normal with log transformation	Linear	$\log(\mu_i)=\beta_0+\tau_i$	σ
Poisson	GLM	$\log(\mu_i)=\beta_0+\tau_i$	--
Negative Binomial type II	GLM	$\log(\mu_i)=\beta_0+\tau_i$	σ
Gamma	GLM	$\mu_i^{-1}=\beta_0+\tau_i$	σ
Gamma-Poisson	GAMLSS	$\log(\mu_i)=\beta_0+\tau_i$	$\log(\sigma_i) = \alpha_0$
Poisson-inverse Gaussian	GAMLSS	$\log(\mu_i)=\beta_0+\tau_i$	$\log(\sigma_i) = \alpha_0$
Mixed effects	Linear	$\mu_i=\beta_0+\tau_i$	$\sigma_\tau + \sigma$
Mixed effects with heteroscedasticity	Linear	$\mu_i=\beta_0+\tau_i$	σ_1,σ_2,σ

*GLM: generalized linear model; GAMLSS: generalized additive model for location, scale, and shape. **μ_i: expected mean of Group *i* (*i* = 1, 2) for the response y_1, β_0: intercept; τ_i: effect of Group *i*. ***α_0: intercept for the dispersion parameter σ.

After fitting the regression models, type-I error rate was calculated as the proportion of the p-values of the F-test or likelihood ratio test (LRT) (depending on the regression model) for the group factor that was lower than the nominal significance level, $\alpha = 0.05$.

A power analysis was performed by simulating the data matrices *Y* under the alternative hypothesis H_a: $\theta \neq \theta_2$, employing 2 methods.

1) A multiplicative effect size δ was applied to the mean of Group 2, $\mu2 = \delta\hat{\mu}$, where $\hat{\mu}$ is the maximum likelihood estimate (MLE) of the general mean of the total number of events by insect; y_1 (waveform Z) was selected as response variable. The following values of effect size were used: $\delta = \{0.2, 0.3, 0.5, 0.7, 0.8, 0.9, 1.1, 1.2, 1.3, 1.5, \text{and } 1.8\}$. The proportion of p-values of the F-test or LRT lower than $\alpha=0.05$ was considered as the power.

2) A multiplicative effect size δ was applied to the probability of y_1 (waveform Z) of Group 2, that is, $\pi_{12} = \delta\hat{\pi}_1$; where $\hat{\pi}_1$ is the MLE of the general probability of y_1. To maintain the constraint $\sum_{k=1}^{K} \pi_k = 1$ of the multinomial model, we subtracted the value $(\pi_{12} - \hat{\pi}_1)/(K-1)$ from all the other probabilities *(k≠1)* uniformly. The following values of effect size were used: $\delta = \{1.1, 1.25, 1.5, 1.75, \text{and } 2.0\}$. The proportion of p-values of the F-test or LRT lower than $\alpha = 0.05$ was considered as the power.

In addition to the regression models described in Table 1, the multinomial regression model was also used. Consequently, the type-I error rate and power analysis (multiplicative effect size δ to the probability of (waveform Z) of Group 2) were calculated. The multinomial regression model is a generalization of the logistic y_1 regression model wherein the mean is modelled by the linear predictor using the following equation:

$$logit(\mu_i)=\beta_0+\tau_i$$

where: μ_i represents the vector of proportions of waveform events for the Group i.

Computing

EPG recording files were processed using INFEST® (Silva et al., 2022). Statistical analysis and computing were performed in R (R Core Team, 2023), using the packages MASS, nnet (Venables & Ripley, 2002), gamlss (Rigby & Stasinopoulos, 2005), nlme (Pinheiro et al., 2021), and extraDistr (Wolodzko, 2020).

Results

Table 2 lists the total number of events and waveform proportions from the experimental data used in the simulations. The experimental groups showed large numerical differences in the number of waveform events, such as Eh3b (74 vs. 218) and Eh2 (16 vs. 95). Moreover, a low frequency was observed in Eh4 for Group 1, which corresponded to one or fewer events per insect. This resulted in the occurrence of zeros in the data to be subjected to statistical analysis. In contrast, waveforms such as Eh3a exhibited a considerably higher number of events.

Table 2. Total number of waveform events (N) and proportions from the data used for simulations.

			Waveforms					
Group	#Insects		Z	Eh1	Eh3a	Eh3b	Eh2	Eh4
1	16	N	201	206	239	74	16	15
		Proportion	0.2676	0.2743	0.3182	0.0985	0.0213	0.0200
2	16	N	255	306	424	218	95	44
		Proportion	0.1900	0.2280	0.3159	0.1624	0.0708	0.0328

Z = non-feeding, Eh1 = stylet penetration, Eh3a = seed disruption, Eh3b = ingestion from seeds, Eh2 = xylem sap ingestion, and Eh4 = ingestion from unknown location.

The total number of activities per insect (N) was overdispersed, and distribution models with a dispersion parameter greater than one presented a better fit (Figure 1). The three models correctly estimated a general mean of approximately 65.5 events, however, the Poisson model underestimated the dispersion of the data. The double-Poisson estimate for the dispersion parameter was 29.67, with a slightly better fit (lower AIC) than that of the Gamma–Poisson, whose dispersion estimate was 47.31. Thus, a double-Poisson model was used to simulate the data using the proposed approach based on a multinomial distribution.

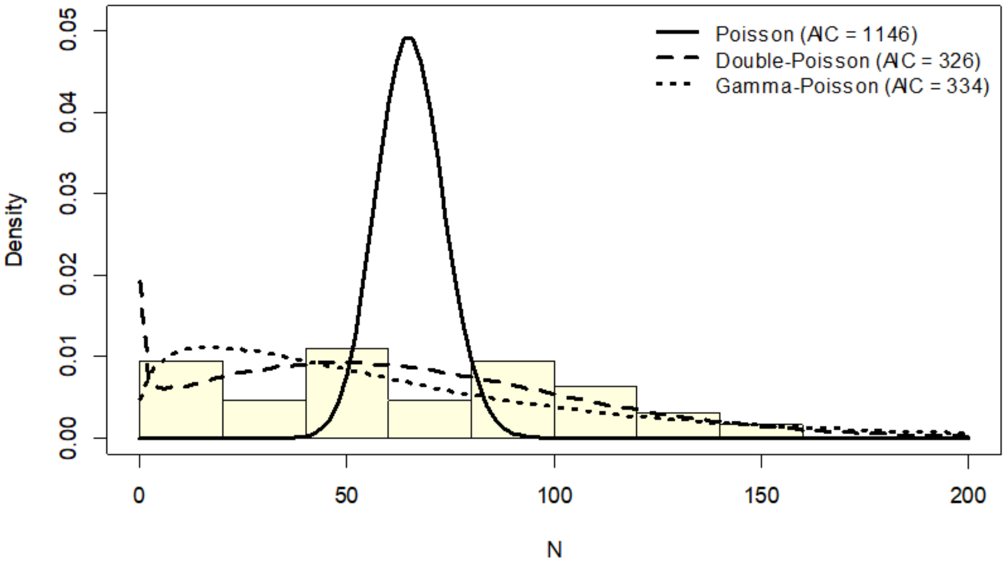

Figure 1. Goodness-of-fit of the probability distribution models to the total number of activities per insect (N).

To check whether the simulation strategy could maintain the correlation structure among the waveforms, we calculated a coincidence index between each resulting correlation matrix from the simulated data and that from the original dataset. The index is based on the average of the absolute differences between the correlation values. We obtained index values ranging as 0.65–0.81, with a mean coincidence of 0.73.

The type-I error rate for 10,000 simulations is shown in the level plot (Figure 2). It was calculated for all fitted regression models, including the multinomial model, which considered all waveform events simultaneously.

The type-I error rate must be measured because it is undesirable for a test to reject a true hypothesis. The multinomial model had the lowest type-I error rate. In general, the other models identified non-existent differences that did not exist, primarily for the waveforms Eh1, Eh3a, and Eh2. The Poisson model exhibited the highest type-I error rate for all waveforms (above 0.4). Except for waveform Z, the Normal model exhibited a type-I error exceeding 0.2.

The power of tests and the root-mean-squared-error (RMSE) are shown in Figure 3, which is based on the effects size of μ and the mean number of total activities by insect. Figure 4 shows the power based on the effect sizes of the probability of the waveform Z.

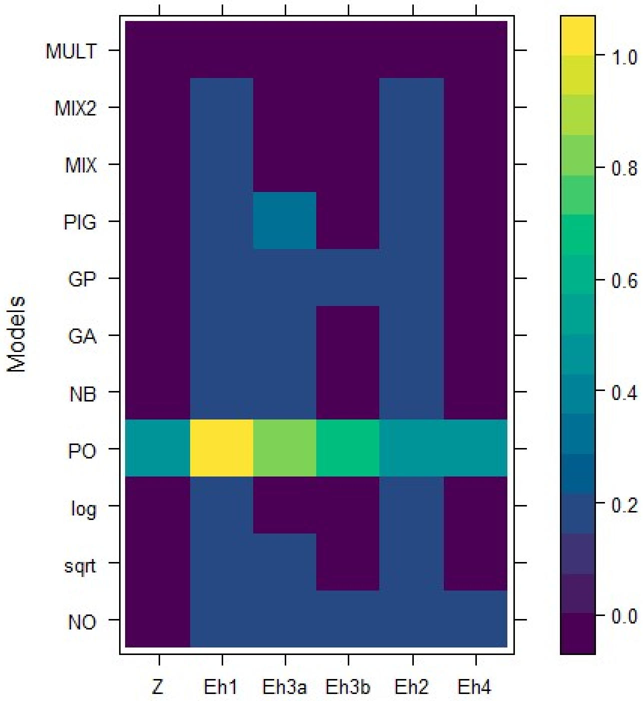

Figure 2. Type-I error rate of likelihood ratio tests (α = 0.05) performed on regression models fitted to simulated data on the number of events per waveform by insect. NO = normal, sqtr = normal with sqrt transformation, log = normal with log transformation, PO = Poisson, NB = negative binomial type II, GA = Gamma, GP = Gamma–Poisson, PIG = Poisson-inverse gaussian, MIX = mixed and MIX2 = Mixed with heteroscedasticity. Z = non-feeding, Eh1 = stylet penetration, Eh3a = seed disruption, Eh3b = ingestion from seeds, Eh2 = xylem sap ingestion, and Eh4 = ingestion from unknown location.

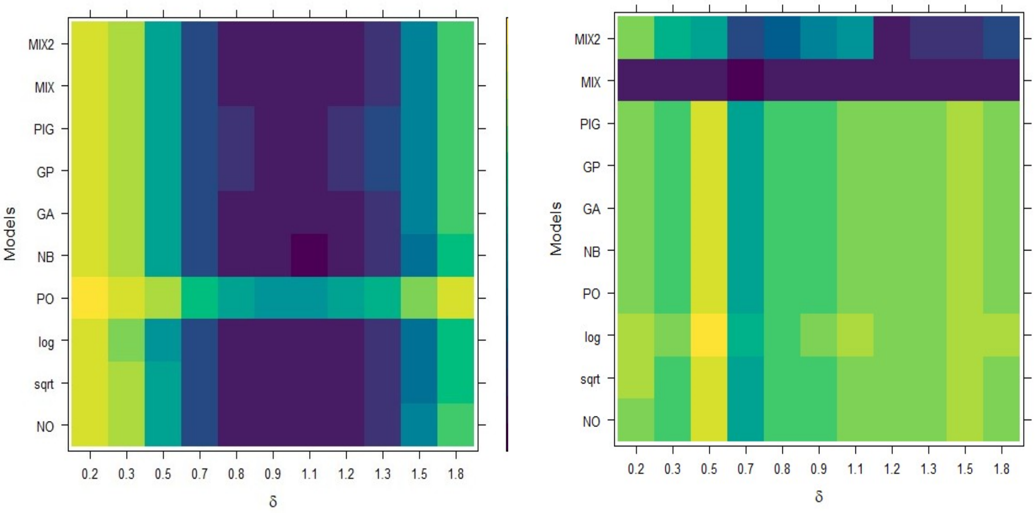

Figure 3. Power analysis (left side) and RMSE (right) based on the effect size of μ and the general mean of the total number of events by insect. NO = normal, sqtr = normal with sqrt transformation, log = normal with log transformation, PO = Poisson, NB = negative binomial type II, GA = Gamma, GP = Gamma–Poisson, PIG = Poisson-inverse gaussian, MIX = mixed, MIX2 = Mixed with heteroscedasticity, and RMSE = root mean-square error.

Power analysis can quantify the extent to which the model can detect statistical differences, and RMSE measures the accuracy of the model. Through the application of a multiplicative effect size δ to the mean of Group 2, the Poisson model could detect differences between treatments with the highest power. The Gamma–Poisson and Poisson-inverse Gaussian models exhibited higher power for small differences (30 and 20%, respectively). In general, the models identified differences with at least 80% power from a 70% difference or greater. Between the two transformations, the log transformation performed slightly better.

The mixed (MIX) model was the most accurate, with RMSE below 1.5, regardless of the effect size of μ. This was followed by the mixed model with heteroscedasticity (MIX2). The other models exhibited similar values.

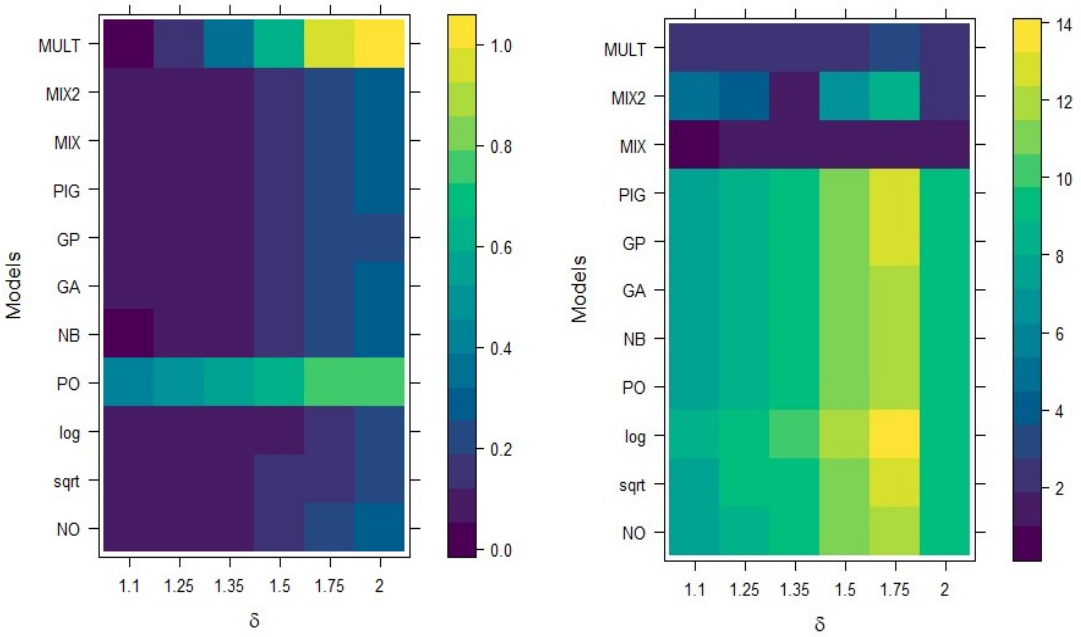

Figure 4. Power analysis (left side) and RMSE (right) based on the effect size δ of the probability of the waveform Z. NO = normal, sqtr = normal with sqrt transformation, log = normal with log transformation, PO = Poisson, NB = negative binomial type II, GA = Gamma, GP = Gamma–Poisson, PIG = Poisson-inverse gaussian, MIX = mixed, MIX2 = Mixed with heteroscedasticity and MULT = multinomial, and RMSE = root mean-square error.

Through the application of a multiplicative effect size δ to the probability of waveform Z for Group 2 in simulations, the Poisson model was observed to be the most sensitive in terms of detecting small differences between Groups. In contrast, the multinomial model was more sensitive in detecting higher (>1.35 ×) differences between the groups. Only the multinomial model achieved 80% power. The multinomial and the mixed models (MIX) exhibited the lowest RMSE regardless of the difference in effect size δ applied to the probability of waveform Z.

Discussion

Overdispersion, as observed in Figure 1, is common in count data, such as EPG experiments (Coly et al., 2016). This is attributable to several factors, such as excess of zeros (Avci et al., 2015). Excess zeros may be caused by a treatment effect, such as pesticide application, or may be a result of insufficient recording time. In the *E. heros* data, Group 1 presented only 15 events of waveform Eh4 throughout the recording (Table 2), which indicated the presence of zeros; although how they were distributed in the data was not indicated.

There are many specific models for overdispersion that can be divided into two classes: (i) models that assume more general forms for the variance function, possibly including additional parameters, such as the binomial model, and (ii) models wherein the response model parameter itself has certain distribution, such as the negative binomial model (Hinde & Demétrio, 1998).

Several statistical models were fitted and evaluated based on goodness-of-fit, type-I error rate, and power analysis. Models used for continuous data were also applied, such as normal data, because there may be asymptotic normal approximations from the count data. Moreover, transformations were evaluated. This is because data not satisfying the assumption of normality is common and prevents the use of classical regression models (Silva et al., 2019).

The multinomial model exhibits the lowest type-I error rate. Furthermore, it facilitates the analysis of complex and interrelated relationships between waveforms (El-habil, 2012; Schmidt et al., 2022). For the power analysis based on the effect size δ of the probability of the waveform Z, the model was the best for detecting differences between groups greater than 1.35x. Nonetheless, no published studies were found in the literature involving the use of the multinomial model to analyze EPG data.

Poisson regression is commonly used to count data from the EPG, as performed by Almeida et al. (2025), to analyze the feeding behavior of *Euschistus heros* treated with *Metarhizium anisopliae*. However, in our study, the Poisson model presented high type-I error rates (above 0.5) for most response variables. This is probably owing to the model's assumption of equidispersion when the variance is equal to the mean, which is a strong limitation (Freitas & Duarte, 2023) to EPG data. In modeling, inappropriate assumptions may result in invalid hypothesis tests (Gourieroux et al., 1984). Thus, the higher power observed with Poisson's law was likely associated with a higher type-I error rate. For data that present overdispersion, models such as negative binomial or Gamma–Poisson are typically a better choice (Hausman et al., 1984).

The Gamma–Poisson model presented type-I error rates of approximately 0.2 for most variables, which is smaller than the Poisson model. The combination of Gamma and Poisson distributions facilitated greater flexibility as it did not assume equidispersion, resulting in a lower type-I error rate for overdispersed data (Freitas & Duarte, 2023). Greenwood and Yule (1920) considered that the number of events of the response variable followed a Poisson distribution, with parameter λ that varied according to a Gamma distribution with parameters α and β. Thus, it is considered that the conditional $Y \mid \lambda \sim \text{Poisson}(\lambda)$, and the parameter λ itself followed a Gamma distribution $\lambda \sim \text{Gamma}(\alpha, \beta)$.

The PIG model is also a derivative of the Poisson distribution proposed by Holla (1967) as an alternative to the Poisson distribution for cases with overdispersion; therefore, lower type-I error rates compared to the Poisson model were expected. Furthermore, the PIG distribution has been considered a better alternative than the Gamma–Poisson distribution to model data with long-tail overdispersion (Putri et al., 2020). In terms of general performance, that is, considering the type-I error, power of tests, and goodness-of-fit, PIG and Gamma–Poisson exhibited similar results, representing good alternatives to model the EPG count data.

The classical normal model is most commonly used to analyze EPG data. However, in our study, the model exhibited high type-I error rates (approximately 0.2). This type of distribution is often used in works involving continuous data and counts (Krithikadatta, 2014), including EPG data (Wayadande et al., 2020; Guedes et al., 2018). For EPG data, particularly from experiments with pesticides, in case of early stops in insect feeding activities, or in cases of different recording times, certain response variables are likely to be heteroscedastic in terms of the experimental factor levels. Thus, the assumption of homoscedasticity is strong and the unique estimate of variance used to test for significant differences may be underestimated, thereby increasing the type-I error rate. In certain situations, the insecticide effect of a treatment can prevent the insects from performing certain feeding activities that the untreated insects usually do. This causes certain response variables to be zero-inflated or, more generically, overdispersed. The classical normal model cannot incorporate this, thereby capturing only part of the observed variability. This renders the related statistical tests more susceptible to type-I error.

In certain studies the count data was assumed to follow a normal distribution, and the t-test was applied to compare treatments (Ebert et al., 2018; Tariq et al., 2017). In our study, data transformation, particularly the logarithm, reduced the type-I error rate close to the nominal level (0.05) for the variables Eh3a, Eh3b, and Eh4.

Thus, probably because of their flexibility in capturing different data structures (Harrison, 2014; Dixon, 2016; Giesselmann & Schmidt-Catran, 2020), mixed models exhibited the lowest RMSE values and relatively low type-I error rates.

Conclusion

This study proposed a comprehensive approach for simulating correlated overdispersed count EPG data and conducted a comparative examination of statistical models. The multinomial model emerged as a robust choice, presenting low values of root-mean-square error, excelling in controlling the type-I error rate, and exhibiting the highest power for the detection of simulated differences between the means. Conversely, the Poisson model and classical normal distribution exhibited inflated type-I error rates in the presence of overdispersion, leading to erroneous conclusions. Among univariate models, the mixed model exhibited the best fit.

Acknowledgements

Authors thank the *Instituto Federal Goiano* for the financial support. ARS thanks CNPq for the scholarship (process No. 309733/2021-9) and FAPEG for granting equipment resources (Process No. 202310267001380).

References

Antúnez, C. C. C., Liano, A. T. G., & Parra, M. A. R. (2022). Distribución espacial de *Euschistus heros* (Hemiptera: Pentatomidae) en cultivos de soja (*Glycine max* (L.) Merril) en los departamentos de San

Pedro e Itapúa. *Revista Científica de la UCSA*, *9*(2), 77-85. https://doi.org/10.18004/ucsa/2409-8752/2022.009.02.077

Avci, E., Alturk, S., & Soylu, E. S. (2015). Comparison count regression models for overdispersed alga data. *Libras*, *25*(1), 1-5.

Backus, E. A., Cline, A. R., Ellerseick, M. R., & Serrano, M. S. (2007). *Lygus hesperus* (Hemiptera: Miridae) feeding on cotton: New methods and parameters for analysis of nonsequential electrical penetration graph data. *Annals of the Entomological Society of America*, *100*(2), 296-310. https://doi.org/10.1603/0013-8746(2007)100[296:LHHMFO]2.0.CO;2

Backus, E. A., Cervantes, F. A., Guedes, R. N. C., Li, A. Y., & Wayadande, A. C. (2019). AC-DC electropenetrography for in-depth studies of feeding and oviposition behaviors. *Annals of the Entomological Society of America*, *112*(3), 236-248. https://doi.org/10.1093/aesa/saz009

Backus, E. A., & Shih, H.-T. (2020). Review of the EPG waveforms of sharpshooters and spittlebugs including their biological meanings in relation to transmission of *Xylella fastidiosa* (Xanthomonadales: Xanthomonadaceae). *Journal of Insect Science*, *20*(4), 1-14. https://doi.org/10.1093/jisesa/ieaa055

Coly, S., Yao, A.-F., Abriald, D., & Charras-Garrido, M. (2016). Distributions to model overdispersed count data. *Journal de la Société Française de Statistique*, *157*(2), 39-63.

Dixon, P. (2016). Should blocks be fixed or random? In *Conference on Applied Statistics in Agriculture Proceedings* (pp. 23-39). New Prairie Press. https://doi.org/10.4148/2475-7772.1474

Ebert, T. A., Backus, E. A., & Rogers, M. E. (2018). Handling artificially terminated events in electropenetrography data. *Journal of Economic Entomology*, *111*(4), 1987-1990. https://doi.org/10.1093/jee/toy117

Ebert, T. A., Backus, E. A., Cid, M., Fereres, A., & Rogers, M. E. (2015). A new SAS program for behavioral analysis of electrical penetration graph data. *Computers and Electronics in Agriculture*, *116*, 80-87. https://doi.org/10.1016/j.compag.2015.06.011

El-habil, A. M. (2012). An application on multinomial logistic regression model. *Pakistan Journal of Statistics and Operation Research*, *8*(2), 271-291. https://doi.org/10.18187/pjsor.v8i2.234

Freitas, S. M., & Duarte, C. G. (2023). Uso das distribuições Poisson, Poisson-Gama, Poisson-Inversa Gaussiana e Poisson-Lindley generalizada para dados de contagem. *Sigmae*, *12*(1), 172-189.

Giesselmann, M., & Schmidt-Catran, A. W. (2020). Interactions in fixed effects regression models. *Sociological Methods & Research*, *51*(3), 1000-1127. https://doi.org/ 10.1177/0049124120914934

Giordanengo, P. (2014). EPG-Calc: a PHP-based script to calculate electrical penetration graph (EPG) parameters. *Arthropod-Plant Interactions*, *8*, 163-169. https://doi.org/10.1007/s11829-014-9298-z

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to poisson models. *Econometrica*, *52*(3), 701-720. https://doi.org/ 10.2307/191347

Greenwood, M., & Yule, G. U. (1920). An inquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *Journal of the Royal Statistical Society. A*, *83*(2), 255–279. https://doi.org/10.2307/2341080

Guedes, R. N. C., Cervantes, F. A., Backus, E. A., & Walse, S. S. (2018). Substrate-mediated feeding and egg-laying by spotted wing Drosophila: Waveform recognition and quantification via electropenetrography. *Journal of Pest Science*, *92*, 495-507. https://doi.org/10.1007/s10340-018-1065-y

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, 1-19. https://doi.org/10.7717/peerj.616

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R & D relationship. *Econometrica*, *52*(4), 909-938. https://doi.org/10.2307/1911191

Hinde, J, & Demétrio, C. G. B. (1998). Overdispersion: Models and estimation. *Computational Statistics & Data Analysis*, *27*(2), 151-170. https://doi.org/10.1016/S0167-9473(98)00007-3

Holla, M. (1967). On a Poisson-inverse Gaussian distribution. *Metrika*, *11*, 115-121. https://doi.org/10.1007/BF02613581

Hu, J., Yang, J. J., Liu, B. M., Cui, H. Y., Zhang, Y. J., & Jiao, X. G. (2020). The feeding behavior explains the different effects of cabbage on MEAM1 and MED cryptic species of *Bemisia tabaci*. *Insect Science*, *27*(6), 1276-1284. https://doi.org/10.1111/1744-7917.12739

Krithikadatta, J. (2014). Normal distribution. *Journal of Conservative Dentistry*, *17*(1), 96-97. https://doi.org/10.4103/0972-0707.124171

Lucini, T., & Panizzi, A. (2018). Electropenetrography monitoring of the neotropical brown-sink bug (Hemiptera: Pentatomidae) on soybean pods: An electrical penetration graph-histology analysis. *Journal of Insect Science*, *18*(6), 1-14. https://doi.org/10.1093/jisesa/iey108

Lucini T., & Panizzi, A. (2017). Electropenetrography (EPG): a breakthrough tool unveiling stink bug (Pentatomidae) feeding on plants. *Neotropical Entomology*, *47*(1), 6-18. https://doi.org/10.1007/s13744-017-0574-3

Mclean, D. L., & Kinsey, M. G. (1964). A technique for electronically recording aphid feeding and salivation. *Nature*, *202*, 1358-1359. https://doi.org/10.1038/2021358a0

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2021). *nlme: Linear and nonlinear mixed effects models*. https://CRAN.R-project.org/package=nlme

Putri, G., Nurrohmah, S., & Fithriani, I. (2020). Comparing Poisson-inverse gaussian model and negative binomial model on case study: Horseshoe crabs data. *Journal of Physics: Conference Series*, *1442*, 1-6. https://doi.org/10.1088/1742-6596/1442/1/012028

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project. org

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

Almeida, A. C. d. S., Rodrigues, M. A., Boaventura, H. A., Vieira, A. S., e Silva, J. F. A., de Jesus, F. G., & Quintela, E. D. (2025). Can *Metarhizium anisopliae* reduce the feeding of the Neotropical brown stink bug, *Euschistus heros* (Fabricius, 1798), and its damage to soybean seeds? *Journal of Fungi, 11*(4), 247. https://doi.org/10.3390/jof11040247

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. Springer Netherlands.

Sarria, E., Cid, M., Garzo, E., & Fereres, A. (2009). Excel Workbook for automatic parameter calculation of EPG data. *Computers and Electronics in Agriculture*, *67*(1-2), 35-42. https://doi.org/10.1016/j.compag.2009.02.006

Schmidt, A. M., Freitas, L. P., Cruz, O. G., & Carvalho, M. S. (2022). A Poisson-multinomial spatial model for simultaneous outbreaks with application to arboviral diseases. *Statistical Methods in Medical Research*, *31*(8), 1590-1602. https://doi.org/10.1177/09622802221102628

Silva, A. R., Almeida, A. C. S., Gonçalves de Jesus, F., Barrigossi, J. A. F. (2022). *Infest: Insect feeding behavior statistics*. INPI - Instituto Nacional da Propriedade Industrial [BR512022001098-4]. https://arsilva.shinyapps.io/infest

Silva, E. M., Furtado, T. D. R., Fernandes, J. G., Cirilo, M. A., & Muniz, J. A. (2019). Leaf count overdispersion in coffee seedlings. *Ciência Rural*, *49*(4), 1-7. https://doi.org/10.1590/0103-8478cr20180786

Sosa-Gómez, D. R., Corrêa-Ferreira, B. S., Kraemer, B., Pasini, A., Husch, P. E., Vieira, C. E. D., Martinez, C. B. R., & Lopes, I. O. N. (2020). Prevalence, damage, management and insecticide resistance of stink bug populations (Hemiptera: Pentatomidae) in commodity crops. *Agricultural and Forest Entomology, 22*(2), 99-118. https://doi.org/10.1111/afe.12366

Tariq, K., Noor, M., Backus, E. A., Hussain, A., Ali, A., Peng, W., & Zhang, H. (2017). The toxicity of flonicamid to cotton leafhopper, *Amrasca biguttula* (Ishida) is by disruption of ingestion: An EPG study. *Pest Management Science*, *73*(8), 1661-1669. https://doi.org/10.1002/ps.4508

Terza, J. V., & Wilson, P. W. (1990). Analyzing frequencies of several types of events: A mixed multinomial-Poisson approach. *The Review of Economics and Statistics*, *72*(1), 108-115.

Tyralis, H., & Papacharalampous, G. (2024). A review of predictive uncertainty estimation with machine learning. *Artificial Intelligence Review, 57*, 94. https://doi.org/10.1007/s10462-023-10698-8.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Wayadande, A. C., Backus, E. A., Noden, B. H., & Ebert, T. (2020). Waveforms from stylet probing of the mosquito *Aedes aegypti* (Diptera: Culicidae) measured by AC-DC electropenetrography. *Journal of Medical Entomology*, *57*(2), 353-368. https://doi.org/10.1093/jme/tjz188

Wolodzko, T. (2020). *extraDistr: Additional univariate and multivariate distributions*. R package version 1.9.1. https://cran.r-project.org/web/packages/extraDistr/index.html.

Ying, L., Baiming, L., Hongran, L., Tianbo, D., Yunli, T., & Dong, C. (2021). Effect of *Cardinium* infection on the probing behavior of *Bemisia tabaci* (Hemiptera: Aleyrodidae) MED. *Journal of Insect Science*, *21*(3), 1-6. https://doi.org/10.1093/jisesa/ieab040