UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE CIÊNCIAS AGRÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

FÁBIO JANONI CARVALHO

MODELOS LINEARES GENERALIZADOS NA AGRONOMIA: ANÁLISE DE DADOS BINOMIAIS E DE CONTAGEM, ZEROS INFLACIONADOS E ENFOQUE BAYESIANO

UBERLÂNDIA
MINAS GERAIS – BRASIL
2019

FÁBIO JANONI CARVALHO

MODELOS LINEARES GENERALIZADOS NA AGRONOMIA: ANÁLISE DE DADOS BINOMIAIS E DE CONTAGEM, ZEROS INFLACIONADOS E ENFOQUE BAYESIANO

Tese apresentada à Universidade Federal de Uberlândia, como parte das exigências do Programa de Pós-graduação em Agronomia – Doutorado, área de concentração em Fitotecnia, para obtenção do título de "Doutor".

Orientadora

Prof. Dra. Denise Garcia de Santana

UBERLÂNDIA
MINAS GERAIS – BRASIL
2019

Dados Internacionais de Catalogação na Publicação
(CIP) Sistema de Bibliotecas da UFU, MG,
Brasil.

C331m
2019

Carvalho, Fábio Janoni, 1991
Modelos lineares generalizados na agronomia [recurso eletrônico] : análise de dados binomiais e de contagem, zeros inflacionados e enfoque bayesiano / Fábio Janoni Carvalho. - 2019.

Orientadora: Denise Garcia de Santana.
Tese (Doutorado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Agronomia.
Modo de acesso: Internet.
Disponível em: http://dx.doi.org/10.14393/ufu.te.2019.1244
Inclui bibliografia.
Inclui ilustrações.

1. Agronomia. 2. Modelos matemáticos. 3. Agronomia - Modelos matemáticos. 4. Agronomia - Métodos estatísticos. I. Santana, Denise Garcia de, 1967, (Orient.) II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Agronomia. III. Título.

CDU: 631

Angela Aparecida Vicentini Tzi Tziboy – CRB-6/947

FÁBIO JANONI CARVALHO

MODELOS LINEARES GENERALIZADOS NA AGRONOMIA: ANÁLISE DE DADOS BINOMIAIS E DE CONTAGEM, ZEROS INFLACIONADOS E ENFOQUE BAYESIANO

Tese apresentada à Universidade Federal de Uberlândia, como parte das exigências do Programa de Pós-graduação em Agronomia – Doutorado, área de concentração em Fitotecnia, para obtenção do título de "Doutor".

APROVADO em 14 de maio de 2019.

Prof. Dr. Marcus Vinicius Sampaio	docente/UFU
Prof. Dr. Lúcio Borges de Araújo	docente/UFU
Prof. Dr. Reinaldo Silva Oliveira Canuto	docente/IFTM
Prof. Dr. Édimo Fernando Alves Moreira	docente/IFTM

Prof. Dra. Denise Garcia de Santana
ICIAG-UFU
(Orientadora)

UBERLÂNDIA
MINAS GERAIS – BRASIL
2019

AGRADECIMENTOS

Agradeço a Deus pela dádiva da vida, proteção, força e saúde durante toda esta trajetória. Aos meus pais Antonio Manoel Carvalho e Maria Edna Janoni Carvalho e irmão Rodrigo Janoni Carvalho por todo o carinho, apoio e atenção dado e por sempre apoiarem e incentivarem meus estudos.

Aos meus amigos e familiares por facilitarem e amenizarem esta árdua jornada com momentos de alegria e descontração.

À Denise Garcia de Santana pela orientação, os incontáveis conselhos, o carinho e a amizade, que mesmo diante das nossas dificuldades do dia a dia, conseguimos estabelecer uma excelente relação orientando-orientador ao longo de tantos anos.

A todos servidores da Universidade Federal de Uberlândia (UFU) e do Instituto de Ciências Agrárias (ICIAG) pelo carinho e atenção que sempre foi oferecido. Aos docentes, agradeço imensamente a todos ensinamentos e experiências compartilhadas.

Aos colegas de trabalho, tanto do ICIAG – Campus Monte Carmelo como do IFTM – Campus Uberaba, que sempre me apoiaram e deram suporte para a minha qualificação.

Aos membros da banca por aceitarem o convite e contribuírem para a melhoria deste trabalho.

SUMÁRIO

RESUMO GERAL i
GENERAL ABSTRACT ii
CAPÍTULO I - A ESTATÍSTICA É NEGLIGENCIADA PELOS PESQUISADORES DAS CIÊNCIAS AGRÁRIAS? 1
1) INTRODUÇÃO 3
2) METODOLOGIA 4
3) RESULTADOS E DISCUSSÃO 4
3) CONSIDERAÇÕES FINAIS 13
4) REFERÊNCIAS 17
CAPÍTULO II - AFASTANDO-SE DA NORMALIDADE E ANOVA: MODELOS LINEARES GENERALIZADOS PARA DADOS BINOMIAIS E DE CONTAGEM NO AMBIENTE COMPUTACIONAL R. 17
1) INTRODUÇÃO 25
2) MÉTODOLOGIA E DISCUSSÃO 27
2.1 Ambiente R 27
2.2 Distanciando-se da normalidade: Modelos Lineares Generalizados (MLG) 28
2.3 Distribuição binomial 31
2.4 Distribuição de Poisson 36
2.5 Sub e sobredispersão nos modelos binomial e Poisson 40
2.6 Quasi-verossimilhança 42
2.7 Distribuição binomial negativa (BN) 49
2.8 Modelo fatorial com tratamento quantitativo – Aplicação da Regressão Logística 52
2.9 Modelos Zero Inflacionados e Zero Truncados 60
3) CONSIDERAÇÕES FINAIS 72
4) REFERÊNCIAS 73
5) ANEXOS 77
CAPÍTULO III - THE USE OF GENERALIZED ADDITIVE MODELS AND BAYESIAN STATISTICS HELP TO SOLVE THE OVERDISPERSION, AUTO-CORRELATION AND ZERO INFLATION PROBLEMS IN APHID POPULATION STUDY. 81
1) INTRODUCTION 83
2) MATERIAL AND METHODS 86
2.1) Experimental data 86
2.2) Data distribution 88
2.3) Temporal correlation 88
2.4) Zero inflation 89
2.5) Computer modeling 89
2.6) First model: GAM (mgcv package) 89
2.7) Second model: GAM with autocorrelation for time (mgcv package) 90
2.8) Third model: BGAM with zero inflation (brms package) 90
3) RESULTS 92
3.1) First model: GAM 92
3.2) Second model: GAM with autocorrelation for time 94
3.3) Third model: BGAM with zero inflation 96
4) DISCUSSION 101
5) CONCLUSIONS 106
6) REFERENCES 107
7) SUPPLEMENTARY MATERIAL 112

RESUMO GERAL

CARVALHO, Fábio Janoni. **Modelos Lineares Generalizados na Agronomia: análise de dados binomiais e de contagem, zeros inflacionados e enfoque bayesiano**. 2019. 115f. Tese (Doutorado em Agronomia) – Universidade Federal de Uberlândia, Uberlândia[1].

Para a validação de qualquer pesquisa científica é necessário um embasamento estatístico correto. Apesar de reconhecer a estatística como elemento-chave da integridade de uma investigação, o uso inadequado é comum nas Ciências Agrárias. No primeiro capítulo há uma pesquisa bibliográfica feita na revista Ciência e Agrotecnologia, em que se discutiram os métodos estatísticos e os erros encontrados, tendo estimulado abordagens mais apropriadas para dados agronômicos. Diante da negligência em relação à estatística observada, o objetivo desta tese foi disseminar técnicas para análise de dados agronômicos que ainda vêm sendo utilizadas de modo incipiente por pesquisadores. Enquanto isso, o segundo capítulo apresenta técnicas para analisar dados binomiais e de contagem, assim como tratar dados com excesso de zeros a partir de três exemplos agronômicos: germinação de *Peltogyne confertiflora* (dado binomial); controle biológico do pulgão-do-algodoeiro, *Aphis gossypii* Glover (dado de contagem); e controle de plantas infestantes (dado binomial). Em cada exemplo, modelos para corrigir a sobredispersão foram ajustados. Já para a correção do excesso de zeros nos dados, apresentaram-se os modelos de zeros inflacionados para o exemplo do pulgão-do-algodoeiro. A tese é finalizada com o terceiro capítulo, que aplica a técnica dos Modelos Aditivos Generalizados (MAGs) em um experimento que objetiva avaliar o efeito de fatores abióticos na população do pulgão-da-couve, *Brevicoryne brassicae* (L.). Tal capítulo aborda técnicas para o tratamento de zeros inflacionados e o ajuste da autocorrelação de medidas ao longo do tempo. Metodologias frequentistas e a análise bayesiana foram aplicadas por meio da simulação de dados pela Cadeia de Markov Monte Carlo (MCMC). Os resultados demonstram a importância de corrigir a sobredispersão, a partir da família binomial negativa. A autocorrelação foi resolvida com a estrutura ARMA, e a modelagem bayesiana conseguiu construir o modelo proposto, com o único contratempo de que a simulação de dados consome maior tempo de análise em detrimento a outras técnicas. Felizmente, com o avanço dos programas computacionais, os resultados têm sido exibidos em menor tempo: a incidência do pulgão *B. brassicae*, por exemplo, mostrou que fatores abióticos podem ser facilmente modelados e analisados por MAGs.

Palavras-chave: autocorrelação temporal; distribuição binomial negativa; Modelos Aditivos Generalizados; simulação de dados; sobredispersão.

[1]Orientadora: Denise Garcia de Santana

GENERAL ABSTRACT

CARVALHO, Fábio Janoni. **Generalized Linear Models in Agronomy: binomial and counting data analysis, inflated zeros and Bayesian approach.** 2019. 115f. Thesis (Doctorate's degree in Agronomy) – Federal University of Uberlândia, Uberlândia[1].

The validation of any scientific research requires a correct statistical background. Despite the knowledge that statistics is a key element for the integrity of a investigation, its misuse is common in the Agricultural Sciences. In the first chapter, a bibliographical research was carried out in the journal "Science and Agrotechnology", to discuss the statistical methods and mistakes found, stimulating more appropriate approaches to agronomic data. Because of the negligence with the statistics, the aim of this thesis was to spread techniques for the analysis of agronomic data that are still being used in an incipient way by researchers. Meanwhile, the second chapter presented techniques to analyze binomial and counting data, as well as the adjustment of zero inflation with three agronomic examples: germination of *Peltogyne confertiflora* (binomial data); biological control of cotton aphid *Aphis gossypii* Glover (counting data); and control of weed plants (binomial data). In each example, models to correct the overdispersion were adjusted. For the excess of zeros in the data, zero inflated models were presented for the cotton aphid example. The thesis is finished with the third chapter, applying the Generalized Additive Models (GAMs) in an experiment that aims to evaluate the effect of abiotic factors on the population of the aphid *Brevicoryne brassicae*. This chapter approaches models to control zero inflation and adjusts the autocorrelation of measurements over time. Frequentist methodologies and Bayesian analysis were applied through Monte Carlo Markov Chain (MCMC) simulation. Results demonstrate the importance to correct overdispersion from the negative binomial family. The autocorrelation was solved with ARMA structure, and the Bayesian model was able to construct the proposed model, with the only setback that data simulation consumed longer analysis time in detriment to other techniques. Fortunately, with the advancement of computer software, the results have been exhibited in less time: *B. brassicae* incidence, for example, showed that abiotic factors can be easily modeled and analyzed by GAMs.

Keywords: temporal autocorrelation; negative binomial distribution; Generalized Additive Models; data simulation; overdispersion.

[1]Supervisor: Denise Garcia de Santana

CAPÍTULO I

A ESTATÍSTICA É NEGLIGENCIADA PELOS PESQUISADORES DAS CIÊNCIAS AGRÁRIAS?

Fábio Janoni Carvalho

Resumo: Para a validação de qualquer pesquisa científica é necessário um embasamento estatístico correto. Apesar de a estatística ser um elemento-chave da integridade de uma investigação, o uso inadequado é comum nas Ciências Agrárias. Com o objetivo de avaliar os métodos utilizados nas análises estatísticas dos artigos científicos publicados nas Ciências Agrárias e estabelecer uma relação entre a qualidade e a negligência da estatística, realizou-se uma pesquisa bibliográfica na revista Ciência e Agrotecnologia. Também foram discutidos os métodos estatísticos empregados para estimular abordagens mais apropriadas para dados agronômicos. Foi feita uma busca na base de dados da revista, mais especificamente nos 36 volumes publicados durante os anos de 2012 a 2016, totalizando 115 trabalhos com estatística na área. Cada artigo foi avaliado quanto a delineamento e esquema experimental, número de repetições, tratamento dos dados obtidos, análise estatística e programas computacionais. Aproximadamente 83% dos trabalhos empregaram a Análise de Variância (ANOVA) para testar a significância dos tratamentos. Entretanto, quase 80% dos artigos que a utilizaram não aplicaram (ou citaram a aplicação de) alguma metodologia para avaliar a homogeneidade de variância ou normalidade dos resíduos. A situação se agrava para a aditividade de blocos, pois nenhum dos 45 trabalhos que utilizaram o delineamento em blocos casualizados relatou testes para a pressuposição de aditividade. Em relação ao restante dos trabalhos que aplicaram a ANOVA, aproximadamente 18% testaram as pressuposições de homogeneidade de variâncias e normalidade dos resíduos, sem relatar nenhum valor das estatísticas dos testes. Os resultados comprovam que a aplicação da estatística nas Ciências Agrárias é superficial, escassa e inerte ao tempo. Diversos modelos, estatísticas e testes já foram criados, e os adventos computacionais permitem que eles sejam analisados de maneira rápida, concisa e correta. Uma publicação com resultados apresentados de forma incorreta favorece o uso dos resultados como base de outros estudos, o que aumenta ainda mais as publicações equivocadas. Esta pesquisa também questiona a falta de artigos que incentivem e demonstrem a aplicação de técnicas estatísticas mais avançadas para o tratamento de dados agrários. Apesar de haver métodos estatísticos mais incisivos e confiáveis, a complexidade da modelagem gera repúdio e insegurança para que pesquisadores fora do âmbito estatístico apliquem tais metodologias.

Palavras-chave: Análise de Variância; Modelos Lineares Generalizados; pressuposições; p-valor.

IS THE STATISTICS NEGLIGENCED BY AGRICULTURAL RESEARCHERS?

Fábio Janoni Carvalho

Abstract: he validation of any scientific research requires a correct statistical background. Although statistics is a key element for the integrity of a research, its misuse is common in the Agricultural Sciences. In order to evaluate the statistical analysis of scientific articles published in the Agricultural Sciences field and to establish a relation between quality and negligence of the statistics, a bibliographical research was carried out in the journal "Science and Agrotechnology". The statistical methods used in the articles to stimulate more appropriate approaches to agronomic data were also discussed. A search was performed in the journal's database, specifically in the 36 volumes published from 2012 to 2016, totalizing 115 articles with statistics in this area. Each article was evaluated by the experimental design and scheme, number of repetitions, statistical treatment and data analysis, and computational programs. Approximately 83% of the studies used the Analysis of Variance (ANOVA) to test the significance of the treatments. However, almost 80% of the articles that used ANOVA did not apply (or cited the application of) some methodology to evaluate the assumptions of homoscedasticity or residuals normality. The situation is worse for block additivity assumption because none of the 45 papers that used the randomized block design reported any tests for this assumption. From the remaining studies that applied ANOVA, 18% tested the variance homogeneity and residuals normality assumptions, without reporting any of the test statistic values. These results prove that the application of statistics in Agricultural Sciences is superficial, scarce and inert to time. An expressive number of new models, statistics and tests have already been created, and computational advents allow models to be analyzed quickly, concisely and correctly. A publication with incorrectly results favors the use of these results as the basis of other researches, which could increase the number of misleading publications. This research also questions the lack of articles that encourage and demonstrate the application of more advanced statistical tools for the treatment of agronomic data. Although statistical methods are increasingly incisive and reliable, the complexity of modeling generates repudiation and insecurity for agronomic researchers to apply these methodologies.

Keywords: Analysis of Variance; Generalized Linear Models; assumptions; p-value.

1) INTRODUÇÃO

Para a validação de qualquer pesquisa científica é necessário um embasamento estatístico correto. Apesar de a estatística ser um elemento-chave da integridade de uma investigação, inadequações são comuns em diversas áreas que a utilizam como suporte, inclusive nas Ciências Agrárias. O mau uso pode ser proveniente da falta de conduta, negligência ou incapacidade do analista em questão (BAILAR, 1986).

Métodos, teorias, técnicas e modelos estatísticos desempenham papel importante em diversos estágios da pesquisa científica. A estatística é essencial para o arranjo experimental adequado, a análise e a interpretação corretas dos dados – sem essas garantias, as conclusões são passíveis de serem refutadas. Diante disso, Nelson e Rawlings (1983) ressaltam dez erros mais comuns na estatística da pesquisa agronômica: não realizar o planejamento estatístico antes da implantação do experimento; usar incorretamente o delineamento experimental; falhar na aleatoriedade do experimento, levando à dependência entre os resíduos; dimensionar, de maneira errada, as parcelas; apresentar número reduzido de repetições; usar, de modo impróprio, as técnicas experimentais como os blocos, por vezes posicionados inadequadamente; no momento da interpretação e análise dos dados construir, de maneira incorreta, os graus de liberdade e somar os quadrados do resíduo; falhar na observação de padrões de variância nos dados; ter dependência exclusiva de apenas uma classe de análise estatística; e, ao relatar os resultados experimentais, se confundir na escolha do teste de comparação entre médias e falhar, na descrição de material e métodos, em relação ao delineamento experimental e aos procedimentos estatísticos utilizados de fato.

Há diferentes abordagens estatísticas que vão desde as ferramentas mais simples, como um teste de *t* ou a regressão linear, até os métodos mais complexos, como os Modelos Lineares Generalizados Mistos e a abordagem bayesiana. Para a correta escolha do método estatístico, o pesquisador deve conhecer a natureza da variável em análise (contínua ou discreta), as distribuições associadas a elas (como binomial, normal e Poisson), as informações sobre o processo de amostragem (tamanho da amostra, independência, aleatoriedade, representatividade, entre outras), a teoria e os pressupostos do modelo (GARDENIER; RESNIK, 2002). Se o investigador não utilizar um método coerente, as conclusões poderão ser super ou subestimadas.

Ademais, se os pesquisadores são descuidados ou enganados com a estatística utilizada, há um prejuízo aplicado diretamente para a comunidade, visto que a estatística

pobre conduz a uma ciência com característica igual. Nesse caso, o registro do estudo pode ser corrupto ou poluído, levando à perda de tempo por parte de outros pesquisadores (DeMETS, 1999).

Com o objetivo de constatar as análises estatísticas de artigos científicos publicados na área das Ciências Agrárias e estabelecer uma relação entre a qualidade e a negligência da estatística, realizou-se uma pesquisa bibliográfica. Foram também discutidos os métodos estatísticos, com o objetivo de estimular abordagens mais apropriadas para dados agronômicos.

2) METODOLOGIA

Uma pesquisa bibliográfica foi realizada na base de dados da revista Ciência e Agrotecnologia, especificamente na seção Agricultural Sciences. Tal periódico foi selecionado em virtude do grande número de publicações em diversas áreas e pela qualidade científica (Qualis A2 – Classificação Quadriênio 2013-2016). Os artigos foram retirados dos 36 volumes publicados na referida seção, de 2012 a 2016, totalizando 171 investigações, das quais 56 não apresentaram estatística (nem mesmo descritiva) ou demonstraram alguma modelagem específica e, por isso, foram descartadas do levantamento. Portanto, a amostra do estudo consistiu em 115 trabalhos científicos da área de Ciências Agrárias.

Cada artigo foi avaliado quanto aos métodos estatísticos empregados para a validação das hipóteses, ao delineamento e esquema experimental, ao número de repetições, aos testes de comparação de médias, aos procedimentos para a análise estatística e aos programas computacionais. Os resultados foram apresentados em percentuais, no que tange à quantidade total de trabalhos verificados.

3) RESULTADOS E DISCUSSÃO

Cinco trabalhos aplicaram o teste de *t* para verificar a significância dos efeitos, o que também constitui uma técnica paramétrica. O teste de *t* avalia se as médias da população de duas amostras se diferem uma da outra, quando o desvio-padrão não é conhecido e a amostra é pequena; logo, o teste define se as duas amostras são ou não da mesma população. Ademais, essa ferramenta assume que a variável é normalmente distribuída, a média é conhecida e a variância da população é calculada conforme a

amostra (SOKAL; ROHLF, 1995). Entretanto, se for utilizado para múltiplas comparações binárias entre os tratamentos, há perda de poder do teste, sendo recomendado teste de *t* com ajuste de Bonferroni, que protege a taxa de erro da família dos testes (PIMENTEL-GOMES, 2009).

Aproximadamente 83% dos artigos utilizaram os modelos de análise de variância (ANOVA) para testar a significância dos tratamentos. Essa técnica paramétrica foi desenvolvida e introduzida por Ronald A. Fisher em 1925 e se refere a um dos testes de hipótese mais utilizados, que possui fácil interpretação e pode ser encontrado em qualquer programa estatístico básico. Se atendidas as pressuposições e repetições com um número adequado, a ANOVA fornece um dos mais poderosos testes de hipótese (McGUINNESS, 2002). O poder de um teste se refere à capacidade de evitar o erro tipo II, em que a hipótese de nulidade não é rejeitada, o que deveria acontecer de fato. Tal erro é intrínseco ao teste estatístico e ao banco de dados, o que impossibilita fixar a taxa de erro aceitável, assim como ocorre em α, com o erro do tipo I. Convém salientar que o erro tipo I concerne à rejeição da hipótese nula, quando ela é verdadeira (ZAR, 1998).

Constata-se que a ANOVA é regida por pressuposições que precisam ser testadas e atendidas. Os resíduos necessitam seguir distribuição normal e ser independentes; as variâncias precisam ser homogêneas; e os blocos, quando existirem, devem possuir efeito aditivo com os tratamentos (PIMENTEL-GOMES, 2009). Entretanto, quase 80% dos artigos que utilizaram a ANOVA não aplicaram (ou citaram a aplicação de) algum teste para avaliar a homogeneidade de variância e a normalidade dos resíduos. A situação se agrava para a aditividade de blocos, pois nenhum dos 45 trabalhos que empregaram o delineamento em blocos casualizados relataram algum teste para a pressuposição de aditividade.

No que tange ao restante dos trabalhos que aplicaram a ANOVA, aproximadamente 18% testaram as pressuposições de homogeneidade de variâncias e normalidade dos resíduos, mas sem relatar nenhum dos valores das estatísticas dos testes. Em um caso, houve equívoco nos testes utilizados para cada pressuposto, ao passo que os outros três artigos testaram apenas uma das pressuposições.

Testes de Shapiro-Wilk e Kolmogorov-Smirnov foram utilizados para checar a normalidade. Na literatura, há mais de 40 testes de normalidade (DUFOUR et al., 1998), como os de Shapiro-Wilk (SHAPIRO; WILK, 1965), Kolmogorov-Smirnov (KOLMOGOROV, 1933), Anderson-Darling (ANDERSON; DARLING, 1954) e Lilliefors (LILLIEFORS, 1967). A análise gráfica também pode inferir a normalidade

dos resíduos, atrelando-se aos valores de curtose e simetria (KAO; GREEN, 2006). Os gráficos quantil-quantil (Q-Q Plot) são os mais empregados para checar a normalidade, mas há também os histogramas, *box-plots* e diagramas caule e folha (RAZALI; WAH, 2011). No entanto, os métodos gráficos não permitem uma aferição segura sobre a pressuposição de normalidade dos dados, pois são subjetivos e possuem uma aplicação auxiliar para outras metodologias. Cumpre dizer que nenhuma inferência gráfica para a normalidade foi utilizada nos artigos avaliados.

O teste de Shapiro-Wilk se limita para tamanhos de amostras menores que 50 e é utilizado com frequência pelos pesquisadores graças ao grande poder estatístico, como observado em diversos estudos (MENDES; PALA, 2003; KESKIN, 2006; RAZALI; WAH, 2011). Para amostras maiores que 50, recomenda-se o teste de Kolmogorov-Smirnov, com correção de Lilliefors (GHASEMI; ZAHEDIAS, 2012).

Nesse contexto, o teste de Kolmogorov-Smirnov não deve ser aplicado sem o ajuste de Lilliefors, em virtude do baixo poder estatístico quando os parâmetros são estimados em relação aos dados, ao invés da amostra, o que usualmente ocorre (STEINSKOG et al., 2007; GHASEMI; ZAHEDIAS, 2012). Alguns programas utilizam o ajuste de Lilliefors no teste de Kolmogorov, mantendo-se o nome original do teste; portanto, é de suma importância verificar se o programa utilizado o faz com a correção.

Apesar de haver diversos relatos na literatura sobre a robustez da ANOVA no tocante a uma pequena violação na normalidade dos resíduos (COCHRAN, 1947; BOX, 1953; GLASS et al., 1972; WINER et al., 1991; UNDERWOOD, 1997, KAO; GREEN, 2006), mostrando que os erros dos tipos I e II são pouco afetados, é difícil quantificar um "pequeno desvio" na normalidade. Pergunta-se ao leitor: O quanto pode ser considerado como pequeno desvio? Como um teste identificaria que não há normalidade no banco de dados testados, mas ao mesmo tempo ressalta que a amostra se aproximou a essa distribuição? Nesses casos, a transformação de dados provavelmente resolveria o problema, mas, se isso não ocorrer, os testes não paramétricos, como Kruskal-Wallis, devem ser utilizados. Cabe lembrar que, enquanto um teste rejeita uma pressuposição, outro pode aceitá-la, e, se há dúvidas sobre o não atendimento de pressuposições por parte de um banco de dados, mantêm-se a ideia do uso de testes não paramétricos (OEHLERT, 2000).

Para a homogeneidade de variâncias, os artigos utilizaram os testes de Brown-Forsythe, Cochran, Bartlett e Levene – este último possui maior poder, se comparado aos demais testes, porém é mais conservativo (LIM; LOH, 1996). Os testes de Bartlett e de

Cochran só devem ser utilizados se os resíduos apresentarem distribuição normal (CONOVER et al.,1981; NETER et al., 1985; ZAR, 1998), pois conseguem ser mais sensíveis à não normalidade que a ANOVA. Estudos de Vorapongsathorn et al. (2004) indicaram que, se os dados possuem distribuição normal, o teste de Bartlett é uma escolha adequada, pois não é afetado pelo tamanho das amostras. Quando os dados não atendem à pressuposição de normalidade, o teste de Levene pode ser utilizado para amostras balanceadas menores.

A pressuposição mais crítica e possivelmente a menos estudada, apesar de ser a mais fácil de ser atendida, se refere ao fato de os erros serem independentes, atributo facilmente garantido com a correta casualização do experimento, algo rotineiro na elaboração de experimentos. Há testes na literatura para essa pressuposição, como o de Durbin-Watson (DURBIN; WATSON, 1950), ou inferências gráficas que, na prática, são ineficazes, pois os resíduos precisam ser alinhados em função da ordem e do arranjo estrutural do experimento. A dependência entre resíduos afeta seriamente o teste F, proporcionando altas taxas de erros dos tipos I e II (COCHRAN, 1947; GLASS et al., 1972).

De fato, a transformação dos dados é uma tentativa para que as pressuposições da ANOVA sejam atendidas e a transformação seja informada na metodologia do artigo. São descritos diversos métodos de transformação na literatura que visam diminuir as discrepâncias entre os dados e atender aos pressupostos da ANOVA. Com relação às transformações, quatro trabalhos aplicaram a angular $arcoseno\sqrt{x/100}$ para percentuais, e cinco, a transformação $\sqrt{x+0{,}5}$. Foram utilizadas também as transformações $\sqrt{\frac{(x+2)}{log(x)}}$, $ln(x+1)$, $log(x+1)$ e o método Yeo-Johnson. Outras quatro pesquisas relataram a transformação, mas não justificam a utilização e nem aplicam os testes para pressupostos. Salienta-se que todas as pressuposições precisam ser novamente testadas para os dados modificados, pois uma pressuposição atendida nos dados originais pode ser violada na escala transformada.

Enquanto que para alguns autores a transformação dos dados é considerada um instrumento útil para a correção das pressuposições (LITTLE; HILLS, 1978; AHRENS et al., 1990; ZAR, 1998), outros consideram que a mesma não deve ser utilizada e outros modelos estatísticos devem ser aplicados (WILSON; HARDY, 2002; JAEGER, 2008; WARTON; HUI, 2011; SHI et al., 2013). Diante dos adventos computacionais e da facilidade de aplicação de modelos mais parcimoniosos, com o ajuste de outras

distribuições ao banco de dados, recomenda-se que a transformação de dados seja efetuada apenas em última instância e que sejam relatadas a transformação dos dados e a justificativa do uso, como excesso de *outliers* e de zeros, além da alta variabilidade nos dados.

Ao observar uma inúmera lista de possibilidades nos testes paramétricos, caso os pressupostos de cada modelo não sejam atendidos, a última saída é a aplicação de testes não paramétricos. Entretanto, tais testes tendem a perder a informação da amostra, pois os dados numéricos são reduzidos a uma forma qualitativa. Além disso, os não paramétricos não são tão eficientes quanto os paramétricos e, para isso, são necessárias maiores amostras ou diferenças para a rejeição da hipótese nula (McKIGHT; NAJAB, 2010).

Das metodologias não paramétricas, apenas um trabalho aplicou o teste de Kruskal-Wallis, que estuda as diferenças entre três ou mais amostras independentes, quando elas não apresentam normalidade (KRUSKAL; WALLIS, 1952), o que constitui uma extensão do teste para dois grupos de Mann-Whitney U (Wilcoxon Rank). Assim, o teste de Kruskal-Wallis é a forma mais generalizada do teste de Mann-Whitney U e a versão não paramétrica da ANOVA (McKIGHT; NAJAB, 2010).

Mesmo com a avalanche de trabalhos que utilizaram da ANOVA, artigos buscaram outras modelagens para o banco de dados, assim como distribuições mais coerentes para os dados, a exemplo dos Modelos Lineares Generalizados (MLGs), com seis artigos e os *probits*, com um artigo. Criados em 1972 por Nelder e Wedderburn, os MLGs englobam diversas técnicas estatísticas para diferentes distribuições pertencentes à família exponencial (normal, binomial, Poisson, Gamma, entre outras), o que flexibiliza os fatores fixos, as variáveis aleatórias e a função de ligação que une os dois componentes em um único modelo. Além disso, experimentos mais rebuscados, como parcelas subdivididas no tempo ou espaço, além da repetição do experimento em outros locais e anos, podem ser facilmente modelados como componentes aleatórios do modelo, com os Modelos Lineares Generalizados Mistos (MLGMs) (LEE et al., 2006).

Cinco trabalhos aplicaram técnicas de agrupamento, por se tratarem de estudos de diversidade genética com materiais variados. Essas técnicas de agrupamento são pertencentes à análise multivariada dos dados, que se refere a todos os métodos estatísticos que analisam simultaneamente múltiplas medidas em cada indivíduo ou objeto sob investigação.

Diversas modificações na ANOVA foram executadas ao longo dos anos, para que o modelo abrangesse novas estruturas de tratamento aos delineamentos experimentais. Dentre eles, há as parcelas subdivididas, tanto no tempo quanto no espaço, que, com um possível ajuste no aumento de repetições ou número de tratamentos, garante uma quantidade satisfatória de graus de liberdade aos resíduos do modelo. Entretanto, há muitos experimentos que não detectam a necessidade dos delineamentos subdivididos, principalmente quando são feitas coletas no tempo em uma mesma parcela, como colheitas ou aplicações de um produto em diferentes períodos. Como o princípio de casualidade é ferido com esses delineamentos, as medições em um mesmo fator da parcela podem se correlacionar, apontando efeitos significativos para fatores, quando isso não ocorre (KOWALSKI; POTCNER, 2003).

O número de parcelas de um experimento é ponto-chave da análise estatística e da confiabilidade do experimento. No âmbito estatístico, quanto maior a quantidade de repetições, mais confiável será a análise; já no ponto de vista prático, há dificuldade em conseguir um experimento com muitas repetições, seja por falta de espaço, mão de obra ou material. Mesmo assim, para a ANOVA, é necessário um número mínimo de 20 parcelas experimentais e três repetições por tratamento, com vistas a garantir que os graus de liberdade referentes ao resíduo sejam superiores a dez (PIMENTEL-GOMES, 2009). Na prática, isso faz com que o Quadrado Médio do Resíduo (QMR) não seja tão alto, o que pode acarretar em probabilidades para efeitos fixos não significativas, quando de fato o são. Outro ponto importante é que a quantidade maior de repetições permite que a taxa do erro tipo II seja controlada. Sete trabalhos analisados apresentaram número insatisfatório de parcelas em que alguns apresentaram até 12, ao passo que quatro trabalhos não citam o número de repetições ou blocos utilizados.

Para os trabalhos que aplicaram testes de comparação entre médias qualitativas, 36 empregaram o teste de Tukey; 31, o agrupamento de Scott-Knott; quatro, o de LSD; cinco o de Dunnett; e cinco, o de Duncan. Esses resultados se equiparam aos encontrados por Ruxton e Beauchamp (2008): em 70 artigos buscados na revista *Behavioral Ecology*, 20 utilizaram o teste de Tukey; 12, o de LSD; um, o de Dunnett; e um, o de Duncan, sendo que nenhuma investigação utilizou o teste de Scott-Knott. As inúmeras opções para a escolha do teste voltado à comparação de médias levam a confusões em relação ao mais liberal, ao mais conservativo ou ao que possui maiores taxas de erro (JONES, 1984).

O teste de Tukey foi criado em 1953 por John Tukey e é conhecido também como teste de Tukey da Diferença Honestamente Significativa (*Honestly Significant Difference*

– HSD). Nesse caso, executa-se a diferença mínima exata entre um conjunto de médias, fazendo com que as comparações par a par sejam avaliadas pela maior diferença encontrada – essa é uma das metodologias mais conservativas e permite que α seja fixado (ABDI; WILLIAMS, 2010). A vantagem é que possui um controle alto das taxas do erro de tipo I, e isso impacta em um custo ao poder do teste que, por sua vez, é reduzido quanto maior for o número de comparações entre médias.

Criado em 1974, o agrupamento de Scott-Knott (SK) leva o sobrenome dois dois autores desse teste que se se consagrou na área das Ciências Agrárias desde 1977 (CHEW, 1977), principalmente na análise de experimentos com um número alto de tratamentos, pois o teste não executa resultados com sobreposição de letras. Diferentemente do teste de Tukey que realiza comparações múltiplas, o de SK é uma análise de agrupamento, impondo uma sucessão dicotômica hierárquica das médias dos grupos.

O método utiliza um algoritmo de agrupamento que se inicia a partir do grupo total das médias observadas, as divide e continua subdividindo os subgrupos formados até que não permaneça nenhuma interseção entre duas médias (JELIHOVSCHI et al., 2014). Outros métodos de agrupamento foram propostos na literatura desde SK, como Jolliffe (1975), Cox e Spjotvoll (1982) e Calinski e Corsten (1985), mas ele é o mais utilizado devido ao simples apelo intuitivo e aos bons resultados gerados.

Willavize et al. (1980) e Carmer e Lin (1983) indicam cautela na aplicação do teste, pois as taxas do erro de tipo I podem ser bem superiores em métodos aglomerativos do que nas comparações múltiplas. As taxas tendem a ser elevadas, principalmente quando poucas médias são comparadas; por isso, o teste é recomendado quando se trabalha com cinco médias ou mais. Nove trabalhos aplicaram o teste SK para quatro médias ou menos, o que pode levar a agrupamentos e divisões errôneas na análise.

Ademais, sete trabalhos aplicaram contrastes, sendo cinco ortogonais e dois não ortogonais. Os contrastes permitem inferências mais específicas do que gerais perante os tratamentos, mas exigem raciocínio no que tange à escolha sobre os melhores contrastes que representarão o comportamento dos dados. Os contrastes serão ortogonais quando forem independentes, podendo ser analisados pelo teste de *t* ou de F. Já os contrastes não ortogonais são avaliados pelo teste de Scheffé que, comparado ao teste de *t*, é mais rigoroso, mas possui baixo poder, não sendo recomendado o uso com menos de duas médias (OEHLERT, 2010). O procedimento de Scheffé consegue ser um dos testes para médias mais conservativos, superando o de Tukey. Se apenas comparações entre pares

são planejadas, o teste de Tukey deve ser utilizado, pois irá resultar em intervalos de confiança mais estreitos (KAO; GREEN, 2008).

O teste de Dunnett é recomendado em situações nas quais um grupo específico é comparado com cada um dos demais tratamentos (DUNNETT, 1955). Normalmente, esse grupo é o controle/testemunha, em que não há aplicação de nenhum tratamento ou, se houver, ela se refere a um tratamento padrão. É preferível a aplicação desse teste em razão do poder estatístico em comparação a uma série de contrastes par a par, pois os contrastes seriam não ortogonais (RUXTON; BEAUCHAMP, 2008). Torna-se um teste indispensável em ensaios fatorais que apresentam tratamento adicional "fora dos cruzamentos entre os fatores".

Dois trabalhos não aplicaram a ANOVA, mas utilizaram testes para comparação de médias, os quais devem ser empregados se a variação dos tratamentos for significativa – estes perdem a razão se forem aplicados sem alguma análise prévia, como a ANOVA. Quando a análise traz um efeito não significativo, mas em algum teste de médias é detectada diferença entre tratamentos, tal situação pode confundir os pesquisadores; por conseguinte, deve-se atentar apenas aos resultados da ANOVA. Caso não haja diferenças significativas, nenhum outro teste deve ser aplicado.

Com os adventos tecnológicos, diversos programas estatísticos foram (e são) constantemente criados e atualizados. Os *softwares* utilizados nos artigos foram SISVAR (36); R (18); SAS (14); Genes (10); Sigma-Plot (7); SPSS (4); Statistica, NTSYS, PoloPlus, Assistat, Selegen e Surfer 9.0 (2); Canoco, SAEG 9.1, WinStat, AgroStat, Minitab, Statgraphics Centurión e Origin 5.0 (1). Vale ressaltar que 17 pesquisas não citaram o programa empregado.

O SISVAR (FERREIRA, 2011) é empregado amplamente nas pesquisas agronômicas brasileiras por ser um *software* de fácil utilização, com ambiente agradável, executando a ANOVA tanto para delineamentos simples quanto para estruturas mais complexas, como as parcelas sub-subdivididas. Entretanto, o programa se limita a não testar as pressuposições do ANOVA (existe o teste de Shapiro-Wilk; porém, ele é aplicado aos dados, e não aos resíduos). Isso faz com que o analista as execute em outro aplicativo, o que contribui para o aumento no índice de trabalhos que não testaram essa pressuposição – 31 dos 36 artigos que utilizaram o SISVAR não as testaram de fato.

Por seu turno, o R é um *software* que expandiu o uso na comunidade científica por ser um programa livre e de código-fonte aberto. Assim, qualquer pesquisador consegue criar ou modificar modelos e distribuí-los gratuitamente sob a forma de pacotes.

Esse *software* é comumente empregado nas investigações internacionais e começa a ganhar destaque no contexto brasileiro. A linguagem de programação limita sobremaneira o programa R, o que ocasiona dificuldades e frustações para pesquisadores de áreas que não sejam relacionadas às Ciências da Computação, como as Ciências Agrárias. A dificuldade da interpretação da linguagem de códigos pode ser facilmente rompida, com ganhos excepcionais ao pesquisador, pois modelos mais parcimoniosos e complexos são encontrados apenas no R.

Esta pesquisa ratifica a elevada quantidade de programas computacionais para analisar os dados. Muitos deles apresentam interface amigável, sendo necessário carregar as informações e escolher o método estatístico a ser utilizado. Enquanto tais *softwares* poupam tempo e esforço, eles apresentam uso inadequado da estatística, pois facilmente o analista pode carregar seus dados no programa sem saber como a análise funciona ou porque determinado teste deve (ou não) ser apropriado (GARDENIER; RESNIK, 2002).

Outro ponto a ser questionado é a busca do p-valor menor que 0,05. Conforme DeMets (1999), 0,05 é um número arbitrariamente escolhido para o p-valor; logo, não há uma razão estatística ou filosófica sobre o fato de um p-valor de 0,051 se fundamentalmente diferente do p-valor 0,05. Entretanto, sob a pressão de publicar, pesquisadores podem decidir por manipular dados para a obtenção de um resultado "significativo", em que devem entender que p-valores são meramente convencionados e não invioláveis. Revistas de grande impacto já discutem sobre a arbitrariedade do p-valor 0,05, como *Nature* e *The American Statistician* (WASSERSTEIN et al., 2019). Vale dizer que há estímulos para "aposentar" o estatisticamente significativo e utilizar os intervalos de confiança como de compatibilidade (AMRHEIN et al., 2019).

Resultados dessa busca comprovam que a aplicação da estatística nas Ciências Agrárias é superficial, escassa e inerte ao tempo. Uma das metodologias mais utilizadas, a ANOVA, foi desenvolvida em 1925; assim, novos modelos, estatísticas e testes foram criados, e os adventos computacionais permitem que tais índices sejam analisados de forma rápida, concisa e correta. Então, por que a estatística ainda é negligenciada pelos pesquisadores das Ciências Agrárias?

A primeira justificava pode se fundamentar na seguinte situação: em qualquer ciência, há mistura de teorias ou a necessidade de outra base teórica; logo, o pesquisador tende a apresentar aversão, medo e preguiça de expandir conhecimentos. Outra justificativa reside na escassez de disciplinas em centros universitários que de fato apliquem a experimentação agrícola e a façam por meio de programas computacionais.

A revolução tecnológica exige que disciplinas como essa possuam interface completamente ligada a computadores, para que o acadêmico se familiarize com os métodos estatísticos e os *softwares*.

Evidentemente, muitos modelos possuem grande modelagem estatística que dificulta a interpretação prática na área de conhecimento do pesquisador. Isso reflete na necessidade de publicações que apliquem modelos estatísticos avançados em conjunto com exemplos práticos nas Agrárias, com a interface em algum programa estatístico. Assim, cada vez mais, os estudiosos terão estatísticas mais refinadas e aprenderão o correto uso para otimizar as pesquisas científicas.

Tanto a negligência perante a estatística quanto a manipulação dos resultados leva a erros de publicação. Textos com resultados incorretos favorecem o uso inadequado, por parte dos pesquisadores, como base em outras investigações, o que eleva o índice de trabalhos equivocados.

Esta pesquisa também questiona sobre a falta de artigos que incentivem e demonstrem a aplicação de técnicas estatísticas mais avançadas para a análise de dados agronômicos. Apesar de haver métodos estatísticos cada vez mais incisivos e confiáveis, a complexidade da modelagem gera repúdio e insegurança para que estudiosos fora do âmbito estatístico apliquem tais metodologias nas pesquisas.

Ademais, as revistas científicas precisam ser mais conscientes no que tange à publicação de resultados que sejam relevantes em um campo específico da pesquisa com escassez de publicações, mesmo com resultados não significativos na análise. Revisores devem dedicar atenção especial para a estatística de cada trabalho, a fim de reduzir o número de artigos com erros estatísticos.

4) CONSIDERAÇÕES FINAIS

Diante da negligência atinente à estatística na área das Ciências Agrárias, o objetivo desta tese é disseminar técnicas para a análise de dados agronômicos que ainda vem sendo utilizadas de modo incipiente por pesquisadores. O próximo capítulo citará exemplos práticos de modelos que se ajustam adequadamente aos dados que serão explicados de forma didática. Para auxiliar na escolha dos modelos que seguem distribuições diferentes da normal, a Tabela 1 sugere um método que não inclui outros mais complexos, como parcelas subdivididas que irão exigir componentes aleatórios

(MLGMs). Entretanto, o princípio selecionar o modelo mais apropriado segue a mesma metodologia dos MLGs.

A natureza dos dados agronômicos não segue perfeitamente uma distribuição normal, e a tentativa de ajustá-los falha com frequência. Informações que representam o número de sucessos de um determinado evento conforme as tentativas (por exemplo, a quantidade de sementes germinadas em um rolo com *n* sementes, o número de insetos mortos depois da aplicação de um produto em uma parcela com *n* insetos) seguem uma distribuição binomial; escalas de nota (para coloração de frutos ou severidade de doenças) se referem a uma distribuição multinomial; dados de contagem (número de insetos em uma planta, contagem de plantas infestantes por parcela) possuem distribuição de Poisson ou binomial negativa; e proporções (de uma doença afetando uma planta ou de folhas com fitotoxidez a determinado produto) se ajustam à distribuição beta.

Ao analisar a evolução da estatística desde a ANOVA até os MLGMs, verifica-se que Fisher e Mackenzie publicaram o primeiro uso da ANOVA para dados experimentais em 1923. Após a criação, Fisher estabeleceu modelos para estatística na pesquisa agrária que rapidamente passaram a ser utilizados na investigação experimental de várias áreas (1925, 1935), difundindo-se ainda mais a ANOVA na comunidade científica. Yates (1940) recuperou a informação interbloco – grande precursora para a metodologia mista – e estendeu o trabalho de Fisher para experimentos mais complexos, como as parcelas subdivididas. Já Bartlett sugeriu as primeiras transformações para quando os dados não atingissem a normalidade (1947). Nesse entremeio, Eisenhart (1947), Henderson (1953, 1963) e Harville (1976, 1977) iniciaram publicações para a modelagem mista, ao passo que Searle (1971) e Graybill (1976) integraram as matrizes algébricas com a teoria dos modelos lineares.

Com o desenvolvimento da interface computacional, *softwares* modernos foram criados, a exemplo do SAS. O comando no SAS "PROC GLM", com um pacote para aplicação de modelos normais lineares, foi introduzido em 1976 e, rapidamente, se tornou uma das ferramentas mais importantes para a análise de dados (apesar do nome, esse comando não aplica os MLGs). Contudo, limitações na função "PROC GLM" e em *softwares* similares eram visualizadas, principalmente para modelos inadequados, como as parcelas subdivididas e os dados não normais.

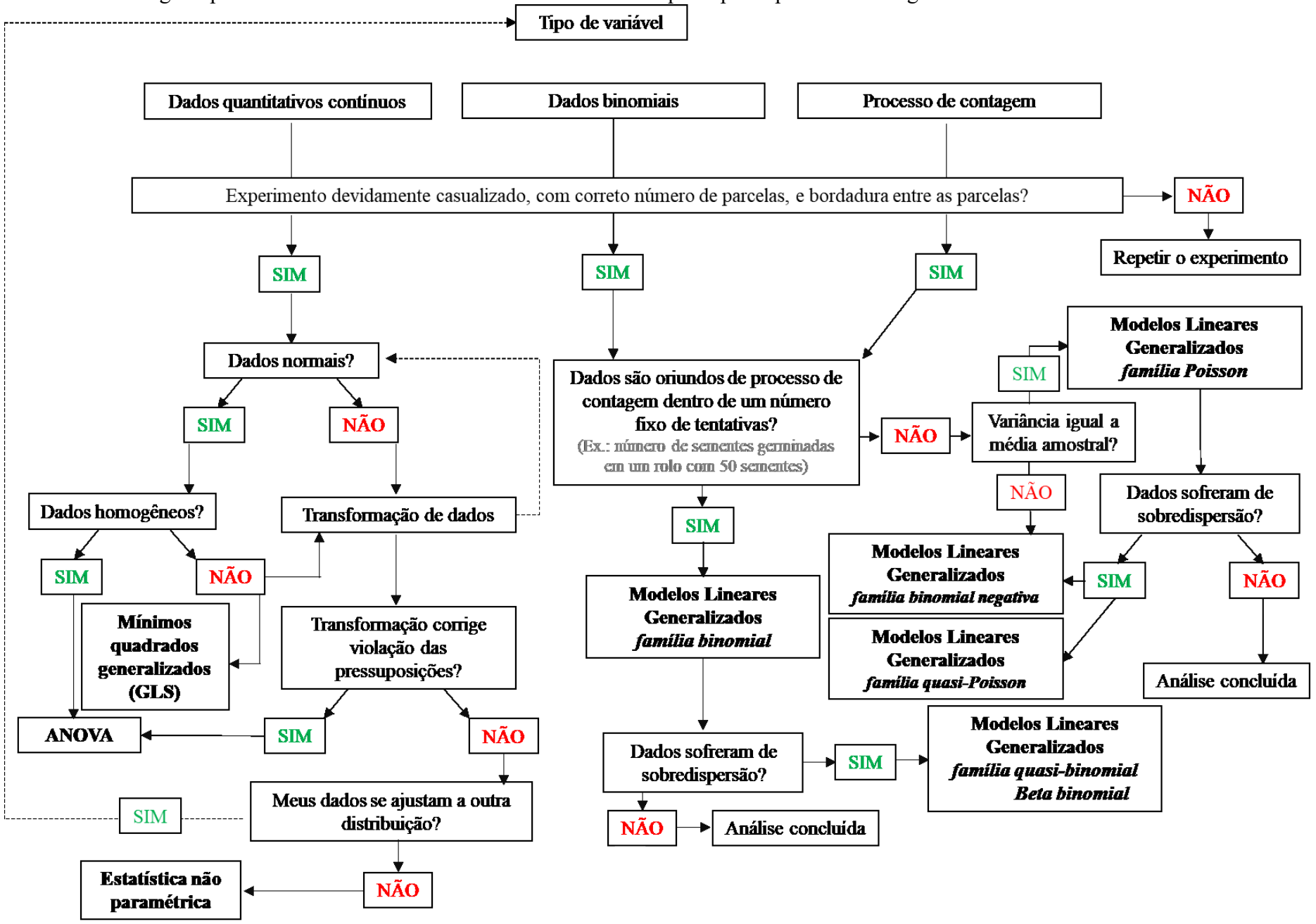
Nelder e Wedderburn (1972) introduziram os MLGs, o que constituiu um avanço para o tratamento de dados não normais. Enquanto a transformação alterava os dados para atender a pressuposições, Nelder e Wedderburn estenderam a base linear da ANOVA para

acomodar distribuições mais plausíveis aos dados. Em 1982, o Departamento de Agricultura dos Estados Unidos (USDA), por intermédio da *Supported University Statisticians of Southern Experiment Stations*, grupo responsável por desenvolver o SAS, iniciou um projeto para expandir a função "PROC GLM". O projeto divulgado em 1989, aliado à publicação de Laird e Ware (1982), se atentou à metodologia dos modelos mistos para várias áreas de pesquisa, incluindo as Ciências Agrárias.

Anteriormente a 1982, modelos mistos eram restritos a publicações específicas. Já no início dos anos 1990, a popularidade dessas técnicas imensa, tanto que, em 1992, o programa SAS introduziu as funções "PROC MIXED", que programava os modelos mistos para dados normais; e "PROC GENMOD", que aplicava os MLGs para dados não normais. Breslow e Clayton (1993) e Wolfinger e O'Connell (1993) integram modelos mistos na teoria dos MLGs, e, à mesma época, os adventos computacionais passavam por um elevado desenvolvimento.

Durante os anos 2000 surgiram *softwares* práticos para os MLGMs, em que a função "PROC GLIMMIX" foi criada em 2005 para o SAS. Outros pacotes para o R (`glmpql`, `gee`, `lme4` etc.) também apareceram, momento em que os programas computacionais conseguiram implantar todos os modelos estatísticos, desde os simples até os mais complexos, para dados normais e não normais.

TABELA 1. Chave geral para a escolha de técnicas estatísticas voltadas aos principais tipos de dados agronômicos.



5) REFERÊNCIAS

ABDI, H.; WILLIAMS, L. J. Tukey's Honestly Significant Difference (HSD) Test. **Encyclopedia of Research Design**, Thousand Oaks, v. 1, p. 1566–1571, 2010.

AHRENS, W. H.; COX, D. J.; BUDHWAR, G. Use of the arcsine and square root transformations for subjectively determined percentage data. **Weed Science**, [*s. l.*], v. 38, p. 452–458, 1990. DOI: https://doi.org/10.1017/S0043174500056824.

ANDERSON, T. W.; DARLING, D. A. A Test of Goodness of Fit. **Journal of the American Statistical Association**, United Kingdom, v. 49, n. 268, p.765–769, 1954. DOI: https://doi.org/10.1080/01621459.1954.10501232.

ARMHEIN, V.; GREENLAND, S.; McSHANE, B. Retire statistical significance. **Nature**, [*s. l.*], v. 567, p. 305–307, 2019. DOI: https://doi.org/ 10.1038/d41586-019-00857-9.

BAILAR, J. Science, statistics, deception. **Annals of Internal Medicine**, [*s. l.*], v. 104, p. 259– 260, 1986. DOI: https://doi.org/10.7326/0003-4819-104-2-259.

BARTLETT, M. S. The use of transformations. **Biometrics**, [*s. l.*], v. 3, p. 39–52, 1947. DOI: https://doi.org/10.2307/3001536.

BOX, G. E. P. Non-normality and tests on variances. **Biometrika**, North Carolina, v. 40, n. 3/4, p. 318–335, 1953. DOI: https://doi.org/10.2307/2333350.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, [*s. l.*], v. 88, p. 9–25, 1993. DOI: https://doi.org/10.1080/01621459.1993.10594284.

CALINSKI, T.; CORSTEN, L. C. A. Clustering Means in ANOVA by Simultaneous Testing. **Biometrics**, [*s. l.*], v. 41, n. 1, p. 39–48, 1985. DOI: https://doi.org/10.2307/2530641.

CARMER, S. G.; LIN, W. T. Type I error rates for divisive clustering methods for grouping means in the analysis of variance. **Communications in Statistics - Simulation and Computation**, [*s. l.*], v. 12, p. 451–466, 1983. DOI: https://doi.org/10.1080/03610918308812331.

CHEW, V. **Comparisons among treatment means in an analysis of variance**. Washington: USDA, 1977. 64 p.

COCHRAN, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. **Biometrics**, [*s. l.*], v. 3, p. 22–38, 1947. DOI: https://doi.org/10.2307/3001535.

CONOVER W. J.; JOHNSON M. E.; JOHNSON M. M. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding

data. **Technometrics**, [*s. l.*], v. 23, p. 351–361, 1981. DOI: https://doi.org/10.1080/00401706.1981.10487680.

COX, D.R.; SPJOTVOLL, E. On partitioning means into groups. **Scandinavian Journal of Statistics**, [*s. l.*], v. 9, p.147–152, 1982.

DeMETS, D. L. Statistics and ethics in medical research. **Science and Engineering Ethics**, [*s. l.*], v. 5, n. 1, p.97–117, 1999. DOI: https://doi.org/10.1007/s11948-999-0059-9.

DUFOUR, J. M.; FARHAT, A.; GARDIOL, L.; KHALAF, L. Simulation-based Finite Sample Normality Tests in Linear Regressions. **Econometrics Journal**, [*s. l.*], v. 1, p. 154–173, 1998. DOI: https://doi.org/10.1111/1368-423X.11009.

DUNNETT, C. W. A multiple comparison procedure for comparing several treatments with a control. **Journal of the American Statistical Association**, [*s. l.*], v. 50, p. 1096–1121, 1955. DOI: https://doi.org/10.1080/01621459.1955.10501294.

DURBIN, J.; WATSON, G. S. Testing for Serial Correlation in Least Squares Regression. **Biometrika**, [*s. l.*], v. 37, p. 409–428. 1950. DOI: https://doi.org/10.2307/2332391.

EISENHART, C. The assumptions underlying analysis of variance. **Biometrics**, [*s. l.*], v. 3, p. 1–21, 1947. DOI: https://doi.org/10.2307/3001534.

FERREIRA, D. F. Sisvar: a computer statistical analysis system. **Ciência e Agrotecnologia**, Lavras, v. 35, n. 6, p.1039–1042, 2011. DOI: https://doi.org/10.1590/S1413-70542011000600001.

FISHER, R. A.; MACKENZIE, W. A. Studies in crop variation: II. The manurial response of different potato varieties. **Journal of Agricultural Science**, [*s. l.*], v. 13, p. 311–320, 1923. DOI: https://doi.org/10.1017/S0021859600003592.

FISHER, R. A. **Statistical methods for research workers**. Edinburgh: Oliver and Boyd, 1925. 319 p.

FISHER, R. A. **The design of experiments**. Edinburgh: Oliver and Boyd, 1935. 236 p.

GARDENIER, J. S.; RESNIK, D. B. The misuse of statistics: Concepts, Tools and a Research Agenda. **Accountability in Research**, [*s. l.*], v. 9, p. 65–74, 2002. DOI: https://doi.org/10.1080/08989620212968.

GHASEMI, A.; ZAHEDIASL, S. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. **International Journal of Endocrinology and Metabolism**, [*s. l.*], v. 10, n. 2, p.486–489, 2012. DOI: https://doi.org/10.5812/ijem.3505.

GLASS, G. V.; PECKHAM P. D.; SANDERS, J. R. Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. **Review of Educational Research**, [*s. l.*], v. 42, p. 239–288, 1972. DOI: https://doi.org/10.3102/00346543042003237.

GRAYBILL, F. A. **Theory and application of the linear model**. North Scituate: Duxbury Press, 1976. 704 p.

HARVILLE, D. A. Extensions of the Gauss–Markov theorem to include the estimation of random effects. **Annals of Statistics**, [*s. l.*], v. 4, p. 384–395, 1976. DOI: https://doi.org/10.1214/aos/1176343414.

HARVILLE, D. A. Maximum likelihood approaches to variance component estimation and to related problems. **Journal of the American Statistical Association**, [*s. l.*], v. 72, p. 320–338, 1977. DOI: https://doi.org/10.1080/01621459.1977.10480998.

HENDERSON, C. R. Estimation of variance and covariance components. **Biometrics**, [*s. l.*], v. 9, p. 226–252, 1953. DOI: https://doi.org/10.2307/3001853.

HENDERSON, C. R. Selection index and expected genetic advance. *In*: HANSON, W. D.; ROBINSON, H. F. (ed.), **Statistical genetics and plant breeding**. Washington: National Research Council, 1963. 982. p. 141–163.

JELIHOVSCHI, E. G.; FARIA, J. C.; ALLMAN, L. B. ScottKnott: A package for performing the Scott-Knott Clustering algorithm in R. **Tendências em Matemática Aplicada e Computacional**, São Carlos, v. 15, n. 1, p. 3–17, 2014. DOI: https://doi.org/10.5540/tema.2014.015.01.0003.

JAEGER, T. F. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. **Journal of Memory and Language**, [*s. l.*], v. 59, p. 434–446, 2008. DOI: https://doi.org/10.1016/j.jml.2007.11.007.

JOLLIFFE, I. T. Cluster analysis as multiple comparison method. **Applied Statistics**, [*s. l.*], v. 1, p. 159–168, 1975.

JONES, D. Use, Misuse, and Role of Multiple-Comparison Procedures in Ecological and Agricultural Entomology. **Environmental Entomology**, [*s. l.*], v. 13, p. 635–649, 1984. DOI: https://doi.org/10.1093/ee/13.3.635.

KAO, L. S.; GREEN, C. E. Analysis of Variance: Is there a difference in means and what does it mean? **Journal of Surgical Research**, [*s. l.*], v. 144, p. 158–170, 2008. DOI: https://doi.org/10.1016/j.jss.2007.02.053.

KESKIN, S. Comparison of several Univariate Normality Tests regarding type I error rate and power of the test in simulation based on small samples. **Journal of Applied Science Research**, [*s. l.*], v. 2, n. 5, p. 296–300, 2006.

KOLMOGOROV, A. N. Sulla determinazione empirica di una legge di distribuzione. **Instituto Italiano degli Attuari**, [*s. l.*], v. 4, p. 83–91, 1933.

KOWALSKI, S. M.; POTCNER, K. J. How to recognize a split-plot experiment. **Quality Progress**, [*s. l.*], v. 36, n. 11, p. 60–66, 2003.

KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in one-criterion variance analysis. **Journal of the American Statistical Association**, [*s. l.*], v. 47, p. 583–621, 1952. DOI: https://doi.org/10.1080/01621459.1952.10483441.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. **Biometrics**, [*s. l.*], v. 38, p. 963–973, 1982. DOI: https://doi.org/10.2307/2529876.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. **Generalized Linear Models with random effects**. Nova Iorque: Chapman and Hall, 2006. 380 p. DOI: https://doi.org/10.1201/9781420011340.

LILLIEFORS, H. W. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. **Journal of American Statistical Association**, [*s. l.*], v. 62, n. 318, p. 399–402, 1967. DOI: https://doi.org/10.2307/2283970.

LIM, T. J.; LOH, W. Y. A comparison of tests of equality of variances. **Computacional Statistics and Data Analyses**, [*s. l.*], v. 22, n. 3, p.28–301, 1996. DOI: https://doi.org/10.1016/0167-9473(95)00054-2.

LITTLE, T. M.; HILLS, F. J. **Agricultural experimentations:** design and analysis. Nova Iorque: Wiley, 1978. 350 p.

McGUINNESS, K. A. Of rowing boats, ocean liners and tests of the ANOVA homogeneity of variance assumption. **Austral Ecology**, [*s. l.*], v. 26, n. 6, p.681–688, 2002. DOI: https://doi.org/10.1046/j.1442-9993.2002.01233.x.

MCKIGHT, P. E.; NAJAB, J. Kruskal-Wallis Test. **The Corsini Encyclopedia of Psychology**, [*s. l.*], p. 1-1, 2010. DOI: https://doi.org/10.1002/9780470479216.corpsy0491.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, [*s. l.*], v. 135, p. 370–384, 1972. DOI: https://doi.org/10.2307/2344614.

NELSON, L. A.; RAWLINGS, J. O. Ten common misuses of statistics in agronomic research and reporting. **Journal of Agronomic Education**, [*s. l.*], v. 12, p. 100–105, 1983.

NETER, J.; WASSERMAN, W.; KUTNER, M. H. **Applied Linear Statistical Models**. 2. ed. Homewood: Richard D. Irwin, 1985. 1408 p.

MENDES, M.; PALA, A. Type I error rate and power of three normality tests. **Pakistan Journal of Information and Technology**, [*s. l.*], v. 2, n. 2, p. 135–139, 2003. DOI: https://doi.org/10.3923/itj.2003.135.139.

OEHLERT, G. W. **Design and analysis of experiments**: Response surface design. Nova Iorque: W.H. Freeman and Co, 2000. 653 p.

PIMENTEL-GOMES, F. **Curso de estatística experimental**. 15. ed. Piracicaba: FEALQ, 2009. 451 p.

RAZALI, N. M.; WAH, Y. B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. **Journal of Statistical Modeling and Analytics**, Malaysia, v. 2, n. 1, p. 21–33, 2011.

RUXTON, G. D.; BEAUCHAMP, G. Time for some a priori thinking about post hoc testing. **Behavioral Ecology**, [*s. l.*], v. 19, n. 3, p. 690-693, 2008. DOI: https://doi.org/10.1093/beheco/arn020.

SCOTT, A. J.; KNOTT, M. A cluster analysis method for grouping means in the Analysis of Variance. **Biometrics**, [*s. l.*], v. 50, p. 507–512, 1974. DOI: https://doi.org/10.2307/2529204.

SEARLE, S. R. **Linear models**. Nova Iorque: John Wiley & Sons, 1971. 531 p.

SHAPIRO, S. S.; WILK, M. B. An Analysis of variance test for normality. **Biometrika**, [*s. l.*], v. 52, p. 591–611, 1965. DOI: https://doi.org/10.2307/2333709.

SHI, P. J.; SANDHU, H. S.; XIAO, H. J. Logistic regression is a better method of analysis than Linear regression of arcsine square root transformed proportional diapause data of *Pieris melete* (Lepidoptera: Pieridae). **Florida Entomologist**, [*s. l.*], v. 96, n. 3, p. 1183–1185, 2013. DOI: https://doi.org/10.1653/024.096.0361.

SOKAL, R. R.; ROHLF, F. J. **Biometry**: the principles and practice of statistics in biological research. 3. ed. Nova Iorque: W. H. Freeman, 1995. 937 p.

STEINSKOG, D. J.; TJOSTHEIM, D. B.; KVAMSTO, N. G. A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality. **American Meteorological Society**, [*s. l.*], v. 135, p. 1151–1157, 2007. DOI: https://doi.org/10.1175/MWR3326.1.

VORAPONGSATHORN, T.; TAEJAROENKUL, S.; VIWATWONGKASEM, C. A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions. **Songklanakarin Journal of Science Technology**, [*s. l.*], v. 26, n. 4, p. 537–547, 2004.

TUKEY, J. W. The problem of multiple comparisons. *In*: BRAUN, H. I. (ed.). **The collected works of John W. Tukey**: Vol. VIII Multiple comparisons: 1948-1983. Nova Iorque: Chapman and Hall, 1953. p. 469–475.

UNDERWOOD, A. J. **Experiments in ecology, their logical design and interpretation using Analysis of Variance.** Cambridge: Cambridge University Press, 1997. 524 p. DOI: https://doi.org/10.1017/CBO9780511806407.

WARTON, D.; HUI, F. The arcsine is asnine: the analysis of proportions in ecology. **Ecology**, [*s. l.*], v. 92, p. 3–10, 2011. DOI: https://doi.org/10.1890/10-0340.1.

WASSERSTEIN, R. L.; SCHIRM, A. L.; LAZAR, N. A. Moving to a World Beyond "p<0.05", **The American Statistician**, [*s. l.*], v. 73, p. 1–19, 2019. DOI: https://doi.org/10.1080/00031305.2019.1583913.

WILLAVIZE, S. A.; CARMER, S. G.; WALKER, W. M. Evaluation of cluster analysis for comparing treatment means. **Agronomy Journal**, [*s. l.*], v. 72, p. 317–320, 1980. DOI: https://doi.org/10.2134/agronj1980.00021962007200020016x.

WINER, B. J.; BROWN, D. R.; MICHELS, K. M. **Statistical Principles in Experimental Design** 3. ed. Nova Iorque: McGraw-Hill, 1991. 928 p.

WILSON, K.; HARDY, I. C. W. Statistical analysis of sex ratios: an introduction. *In*: HARDY, I. C. W. (ed.). **Sex ratios**: Concepts and Research Methods. Reino Unido: Cambridge University Press, 2002. p. 49–92.

WOLFINGER, R. D.; O'CONNELL, M. Generalized linear mixed models: A pseudo-likelihood approach. **Journal of Statistical Computation and Simulation**, [*s. l.*], v. 48, p. 233–243, 1993. DOI: https://doi.org/10.1080/00949659308811554.

YATES, F. The recovery of inter-block information in balanced incomplete block designs. **Annals of Eugenics**, [*s. l.*], v. 10, p. 317–325, 1940. DOI: https://doi.org/10.1111/j.1469-1809.1940.tb02257.x.

ZAR, J. H. **Biostatistical analysis**, 4. ed. Nova Jersey: Prentice Hall, 1998. 662 p.

CAPÍTULO II

AFASTANDO-SE DA NORMALIDADE E ANOVA: MODELOS LINEARES GENERALIZADOS PARA DADOS BINOMIAIS E DE CONTAGEM NO AMBIENTE COMPUTACIONAL R

Fábio Janoni Carvalho

Resumo: A estatística se tornou indispensável na pesquisa científica. Nesse contexto, a expressiva utilização da Análise de Variância (ANOVA) nas Ciências Agrárias despertou um alerta nos últimos anos pelos estatísticos porque, muitas vezes, os dados tratados não atendem a pressuposições. Dados com distribuições não normais são comuns em várias áreas da pesquisa agrária, como a porcentagem de sementes germinadas (binomial), a contagem de plantas infestantes por parcela (Poisson ou negativa binomial), o tempo para florescimento (exponencial ou gamma), a escala de uma doença (multinomial) e a proporção afetada de uma folha (beta). Apesar de haver modelos que comportem melhor a natureza e a variabilidade dos dados, como é o caso das distribuições binomial e de Poisson, eles ainda são pouco empregados, em virtude da dificuldade inerente à estatística e à falta de material que auxilie o pesquisador em uma linguagem compreensível à área. O objetivo deste capítulo foi mostrar, de forma didática, as principais técnicas para análise de dados binomiais e de contagem, assim como o tratamento de dados com excesso de zeros no ambiente computacional R para pesquisadores das Ciências Agrárias. O capítulo destaca a teoria e a análise no programa R a partir de três exemplos agronômicos: germinação de quebra-machado (dado binomial); controle biológico do pulgão-do-algodoeiro (dado de contagem); e controle de plantas infestantes (dado binomial). Em cada exemplo foram ajustados os modelos para corrigir a sobredispersão, com apresentação da análise de *deviance*, dos testes de comparação *post hoc* e das técnicas para verificar a qualidade de ajuste dos modelos. Ainda no controle de plantas infestantes, a análise de regressão logística foi ilustrada. Em se tratando da correção do excesso de zeros no banco de dados, ressaltam-se os modelos de zeros inflacionados para os dados do pulgão-do-algodoeiro. Os Modelos Lineares Generalizados (MLGs) oferecem qualidade de ajuste para dados binomiais, de contagem e com excesso de zeros. Esse tipo de dado é comumente encontrado nas Ciências Agrárias, e o presente capítulo une a teoria dos MLGs e a aplicação destes no ambiente computacional R, de forma didática e prática, aproximando o pesquisador a técnicas mais corretas para análise de dados, em razão da falta de materiais com esse objetivo.

Palavras-chave: binomial; binomial negativa; Modelos Zeros Inflacionados; Poisson; regressão logística; sobredispersão.

AWAY FROM DATA NORMALITY AND ANOVA: GENERALIZED LINEAR MODELS FOR BINOMIAL AND COUNTING DATA IN THE SOFTWARE R

Fábio Janoni Carvalho

Abstract: Statistics became indispensable in the scientific research. In this context, the expressive use of Analysis of Variance (ANOVA) in Agrarian Sciences has teased an alert in recent years by statisticians because the treated data do not often meet the ANOVA assumptions. Non-normal data is common in several areas of agrarian research, such as the percentage of germinated seeds (binomial), count of weeds per plot (Poisson or negative binomial), time to flowering (exponential or gamma), disease scale (multinomial) and leaf ratio affected by a disease (beta). Despite the existence of models that handle better the data characteristics and variability, for example the binomial and Poisson distributions, the use of these models is still incipient, given the difficulty inherent in statistics and the lack of material that guides the researcher in an understandable language to the area. The aim of this chapter was to provide, in a didactic form, the main techniques for binomial and counting data analysis, as well as the treatment of data with zero inflation in the software R for researchers of Agrarian Sciences. The chapter highlights the theory and the analysis in the software R with three agronomic examples: germination of *Peltogyne confertiflora* (binomial data); biological control of cotton aphid (counting data); and control of weed plants (binomial data). In each example, models to correct the overdispersion were adjusted, presenting the deviance analysis, *post hoc* comparison tests and techniques to verify the model adjustment. Furthermore, in the control of weed plants, logistic regression analysis was illustrated. In the case of correction of the zero inflation in the database, the zero inflated models for the data of the cotton aphid are emphasized. Generalized Linear Models (GLMs) provide adjustment quality for binomial, counting and zero inflated data. This type of data is commonly found in the Agrarian Sciences, and the present chapter approximates the GLMs theory and their application in the computational environment R in a didactic and practical way, bringing the researcher closer to the correct data analysis techniques, due the lack of materials with this scope.

Keywords: binomial; negative binomial; Zero Inflated Models; Poisson; logistic regression; overdispersion.

1) INTRODUÇÃO

A estatística se tornou indispensável em investigações científicas. A expressiva utilização da Análise de Variância (ANOVA) nas Ciências Agrárias despertou um alerta nos últimos anos porque, muitas vezes, os dados não atendem aos requisitos da ferramenta, principalmente em relação à normalidade de resíduos e à homocedasticidade. Diante disso, pesquisadores partem para a transformação dos dados ou até mesmo para estatísticas não paramétricas, mesmo com as severas críticas dos estatísticos em relação a esses procedimentos (ver Capítulo I). De fato, essa não é a melhor abordagem, dada a existência de modelos que conseguem tratar, com maior eficiência, as variâncias heterogêneas e os resíduos (dados) não normais (STROUP, 2012).

Antes de 1990, o Teorema do Limite Central garantia que, independentemente da distribuição dos dados, a distribuição da média amostral poderia ser assumida como aproximadamente normal. Uma vasta literatura considerando a robustez da ANOVA a desvios de normalidade se acumulou nessa década e, caso houvesse falhas nos pressupostos, transformações seriam recomendadas para estabilizar variâncias ou "normalizar" resíduos (MILLER, 1997).

A transformação angular consegue ser mais imprecisa que a ANOVA sem transformação, em virtude da perda do poder do teste na comparação dos tratamentos. Para dados naturalmente distribuídos segundo a binomial ou Poisson, que constantemente são "forçados" a seguirem distribuição normal, a ANOVA deveria ser considerada inaceitável para publicações científicas (STROUP, 2012).

De 1990 a 2000, avanços nas teorias estatísticas, aliados aos recursos computacionais, permitiram o uso de modelos mais refinados e completos, principalmente com os Modelos Lineares Generalizados (MLGs). Os MLGs estendem a teoria dos modelos lineares para acomodar dados que podem ser não normais, com variâncias heterogêneas e correlacionadas (LEE et al., 2006). Nesse contexto, a análise de dados agronômicos ainda permanece nos modelos anteriores a 1990, o que a torna antiquada e obsoleta.

Dados com distribuições não normais são comuns em várias áreas da pesquisa agrária. Exemplos incluem a porcentagem de sementes germinadas (binomial), a contagem de plantas infestantes por parcela (Poisson ou negativa binomial), o tempo para florescimento (exponencial ou gamma), a escala de uma doença (multinomial) e a

proporção afetada de uma folha (beta) (STROUP, 2015). A distribuição binomial está bastante presente em experimentos nos quais há determinada quantidade de tentativas (fixa ou não), em que se contabiliza o número de sucessos para o evento de interesse. A distribuição de Poisson envolve dados de contagem, ao considerar números inteiros que podem assumir qualquer valor, diferentemente da distribuição binomial, em que o máximo valor obtido se refere ao número de tentativas. Apesar disso, a distribuição normal não é apropriada para dados de contagem (por se tratar de números inteiros e não negativos), tampouco nem para dados binomiais (por se referir a uma série de eventos de sucesso ou fracasso – distribuição de Bernoulli). Isso desperta a necessidade de ajustar novos modelos para os dados (HOEF; BOVENG, 2007), em que as distribuições diferem da normal, pois são discretas, ao invés de contínuas.

Ademais, os dados ecológicos também contam com uma expressiva quantidade de zeros que prejudicam a análise dos dados. Modelos de zeros inflacionados se tornaram populares para a correção desses dados, sendo utilizados nas áreas de Entomologia (SAMPAIO et al., 2017), Fitopatologia (HAAS et al., 2011; GONZATTO JÚNIOR et al., 2017), Fitotecnia (CALAMA et al., 2011), Proteção de Plantas (YEŞILOVA et al., 2010), e Nematologia (DENDWOOD et al., 2008).

Mesmo com a existência de modelos que comportem melhor a natureza e a variabilidade dos dados, como as distribuições binomial e de Poisson, a utilização ainda é incipiente (ver Capítulo I), dada a dificuldade inerente à estatística e à falta de material que auxilie o pesquisador em uma linguagem compreensível à área. Diante disso, o objetivo deste capítulo é proporcionar aos pesquisadores das Ciências Agrárias uma abordagem didático-científica das principais técnicas para análise de dados binomiais e de contagem, assim como o excesso de zeros, em se tratando do ambiente computacional R.

2) METODOLOGIA E DISCUSSÃO

2.1 Ambiente R

Programas computacionais surgiram nas últimas décadas principalmente para a execução de técnicas estatísticas mais refinadas e com maior extensão de fórmulas. Dentre eles, o R se destaca não apenas por ser gratuito, mas pela adaptabilidade ao perfil do pesquisador. Assim, qualquer indivíduo pode fornecer pacotes adicionais ao *software*, acrescentando novas metodologias.

O R, por meio da linguagem de programação, permite processar dados, fazer cálculos, realizar análises estatísticas e construir gráficos. Possui um sistema planejado e coerente, sem se tornar inflexível e específico, estando embutido em outros programas estatísticos como Genes, MINITAB e SAS. O *software* pode ser baixado pelo *website* <http://www.r-project.org/>, e a região de CRAN utilizada pelo capítulo foi a da Universidade de São Paulo (USP), São Paulo (<http://www.vps.fmvz.usp.br/CRAN/>), na versão 3.5.0.

Alguns pesquisadores questionam sobre o motivo de as análises em programas estatísticos convencionais serem substituídas pelo R. A princípio, não deveriam, pois, se o estudioso se ambienta a um limite específico de testes estatísticos, não há razão para mudanças; logo, o motivo para se transformar é tomar conhecimento e vantagens dos novos (e antigos) métodos estatísticos, permitindo não só uma expansão na quantidade de análises possíveis, como também na qualidade proporcionada pelas novas análises aos dados (CRAWLEY, 2007). Outros pretextos são a qualidade e a confiabilidade do programa, além do grande grupo de suporte *online* disponível. Por fim, a tecnologia é gratuita, em que novos pacotes com modelos específicos para cada área de pesquisa são lançados constantemente.

Por ser um programa de linguagem, as atualizações dos pacotes que integram o R podem ocasionar mudanças de fórmulas; por esse motivo, apresentaram-se versões dos pacotes utilizadas neste capítulo (Tabela 1). Antes de iniciar qualquer análise estatística de interesse, os pacotes do R devem ser instalados com êxito pelo caminho: "*Pacotes* : *Instalar pacote(s)* : *CRAN mirror* : *Packages* (escolher o pacote de interesse) : Ok". A instalação pode também ser realizada pelo comando `install.packages("Nome do`

pacote"). Para a aplicação de alguma função do pacote e este seja carregado pelo *software*, é necessária a execução antecipada do comando `library(Nome do pacote)`.

TABELA 1. Lista de pacotes no programa R para a aplicação dos modelos apresentados.

Pacote	*Versão*	*Autores*	*Utilização*
`stats`	3.5.0	Pacote básico do programa	MLG Coeficientes do modelo Teste de hipóteses
`car`	3.0-0	Fox e Weisberg (2011)	Teste de hipóteses
`AER`	1.2-5	Kleiber e Zeiles (2008)	Teste de sobredispersão no modelo Poisson
`MASS`	7.3-5	Venables e Ripley (2002)	MLG distribuição binomial negativa
`pscl`	1.5.2	Jackman (2017)	Modelagem Hurdle (ZTP e ZTBN) Modelagem de Zeros Inflacionados (ZIP e ZIBN)
`emmeans`	1.2.2	Lenth (2018)	Comparação de médias
`multcompView`	0.1-7	Graves et al. (2015)	Exibe as letras do teste de médias

2.2 Distanciando-se da normalidade: Modelos Lineares Generalizados (MLG)

Modelos de regressão para dados que não seguem a distribuição normal podem ser representados pelos MLGs (NELDER; WEDDERBURN, 1972), os quais generalizam modelos lineares clássicos para variáveis contínuas, de forma que toda a estrutura para estimação e predição pode ser estendida para modelos com outras distribuições aplicáveis pertencentes à família exponencial, tais como binomial, Poisson, beta e gamma (DOBSON; BARNETT, 2008). A descrição detalhada da estrutura dos MLGs pode ser obtida em vasta literatura, com destaque para McCullagh e Nelder (1989), Dobson e Barnett (2008) e Lee et al. (2006).

Os MLGs descrevem a dependência de uma variável escalar $y_i (i = 1, \dots, n)$ em um vetor de regressores x_i. A distribuição condicional de $y_i|x_i$ pertence à família exponencial com probabilidade da função densidade definida como:

$$f(y; \lambda, \phi) = \exp\left(\frac{y \cdot \lambda - b(\lambda)}{\phi} + c(y, \phi)\right) \quad (1)$$

em que: λ: parâmetro canônico; ϕ: parâmetro de dispersão.

As funções $b(\cdot)$ e $c(\cdot)$ são conhecidas e determinadas de acordo com a família de distribuição. A dependência da média condicional μ_i aos regressores x_i é especificada por:

$$\begin{aligned} g(\mu_i) &= x_i^T \beta \\ g(\mu) &= \eta = X\beta \end{aligned} \tag{2}$$

em que: $g(\cdot)$: função de ligação (η); $X = (x_1, x_2, \ldots, x_n)^T$: matriz do modelo composta pelos componentes explanatórios; x_i^T: i-ésima linha da matriz experimental X; β: vetor dos coeficientes de regressão estimados.

A função de ligação estabelece a relação entre a média e o modelo proposto em determinada escala. Neste capítulo, optou-se por utilizar apenas as funções de ligações canônicas específicas de cada distribuição, pois elas naturalmente tendem a simplificar o modelo (MYERS et al., 2002) e buscar ao máximo modelos completos que, ao mesmo tempo, sejam simples. Entretanto, as funções de ligação podem ser facilmente alteradas e os modelos, criados com as diferentes funções de ligação, se comparados quanto ao ajuste.

Nesse contexto, o R implantou a função `glm()` no pacote `stats`, que é bastante flexível e absorve a estrutura dos MLGs (CHAMBERS; HASTIE, 1992). A fórmula foi originalmente implementada por Simon Davies, da Universidade de Auckland, e extensamente reescrita pelos membros da comunidade R. A fórmula é definida por:

```
glm (formula, family = (link=), data =, ..., offset, ...)
```

em que: `formula` é um objeto que anuncia a estrutura do modelo; `family` é a descrição da distribuição dos dados; `link` é a função de ligação a ser utilizada (se não especificado, o sistema usa a função canônica); `data` é o banco de dados; `offset` é um vetor numérico utilizado para especificar um componente conhecido para ser incluído no preditor linear durante a modelagem.

O delineamento do modelo é especificado no parâmetro `formula`, em que o operador "~" expressa a relação entre variável e os fatores, e os operadores "+" e "*" ou ":" indicam a relação entre os fatores, em que "+" é a adição de um fator e "*" ou ":" a interação destes fatores. Quando uma interação é expressa com "*", o programa entende que os fatores isolados também estão inclusos. A Tabela 2 apresenta as combinações mais

comuns para experimentos agrícolas. Vale ressaltar que a adição de mais fatores segue a mesma lógica apresentada.

TABELA 2. Estrutura básica para os fatores em um Delineamento Inteiramente Casualizado (DIC) e um Delineamento de Blocos Casualizados (DBC) no argumento `formula`.

Delineamento		Estrutura do argumento `formula`
DIC	1 fator	`variável ~ fator1`

	2 fatores	`variável ~ fator1*fator2`
		`variável ~ fator1 + fator2 + fator1:fator2`
		`variável ~ fator1 + fator2 + fator1*fator2`

DBC	1 fator	`variável ~ fator1 + bloco`

	2 fatores	`variável ~ fator1*fator2 + bloco`
		`variável ~ fator1 + fator2 + fator1:fator2 + bloco`
		`variável ~ fator1 + fator2 + fator1*fator2 + bloco`

A função `glm` permite utilizar as seguintes famílias de distribuição: normal (`family = gaussian`), binomial (`family = binomial`), Poisson (`family = poisson`), Gama (`family = Gama`) e normal inversa (`family = inverse.gaussian`). Os modelos quasi também podem ser atribuídos para a distribuição normal (`family = quasi`), binomial (`family = quasibinomial`) e de Poisson (`family = quasipoisson`).

Como parte fundamental dos MLGs, a família exponencial dos dados decide as fórmulas estatísticas do modelo e a função de ligação. Para dados de proporção é utilizada a distribuição binomial, que expande a distribuição de Bernoulli. Quando os dados são provenientes de contagem, emprega-se a distribuição de Poisson (CHAMBERS; HASTIE, 1992). Vale ressaltar que várias alterações foram realizadas nas distribuições para melhor adequação ao banco de dados, o que gera novas distribuições como a binomial negativa e a beta binomial, que não são encontradas na função `glm`.

Com o escopo de testar a significância dos efeitos do modelo, a análise de *deviance* (ANODEV) é utilizada nos MLGs. O *deviance* é a estatística do *log* da taxa de máxima de verossimilhança (LEE et al., 2006), descrito como:

$$S_p = 2\left(\hat{\lambda}_n - \hat{\lambda}_p\right) = \frac{D_p}{\phi} = \frac{2}{\phi}\sum_{i=1}^{n}\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) + b\left(\hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right)\right] \quad (3)$$

em que: $\hat{\lambda}_n$: valor máximo do logaritmo da função de verossimilhança no modelo saturado; $\hat{\lambda}_p$: valor máximo do logaritmo da função de verossimilhança no modelo corrente; S_p: desvio escalonado; D_p: desvio; $\hat{\phi}$: parâmetro de dispersão estimado; $\tilde{\theta} = q(y_i)$: estimativa de máxima verossimilhança do parâmetro canônico sob o modelo

saturado; $\hat{\theta} = q(\hat{\mu}_i)$: estimativa de máxima verossimilhança do parâmetro canônico sob o modelo corrente.

Para que os resultados sejam informados, basta solicitar a função `summary()` e especificar o modelo criado – a função `coef()` do pacote `stats` estabelece os coeficientes para os parâmetros do modelo. Para o teste de hipóteses, deve-se recorrer à função `anova()`, também do pacote `stats`, na forma:

```
anova(object, test="F")
```

onde: `object` é o modelo criado; `test` é o teste de hipótese a ser utilizado (opções: "`Chisq`", "`LRT`", "`Rao`", "`F`" ou "`Cp`").

2.3 Distribuição binomial

Criada por James Bernoulli, a distribuição binomial surge de um conjunto de *n* tentativas idênticas e independentes, com a probabilidade de sucesso dada por $P(Y_i = 1) = \pi$ e probabilidade de fracasso dada por $P(Y_i = 0) = 1 - \pi$, considerando o sucesso como o acontecimento do evento de interesse, e o fracasso, o não acontecimento ou o acontecimento inverso do esperado. A distribuição é definida por:

$$f(y;\pi) = \binom{N}{y} \times \pi^y \times (1-\pi)^{N-y} \quad (4)$$

em que: *Y*: número de sucessos (por exemplo, a semente germinou, um produto matou o inseto, a planta ficou doente); *N*: número de tentativas (como número total de sementes, insetos ou plantas na parcela); π: probabilidade do evento de sucesso..

A esperança e a variância da distribuição são representadas por $E(Y) = N \times \pi$ e $var(Y) = N \times \pi \times (1-\pi)$. Para a distribuição binomial no R, a função `glm` é especificada em `family=binomial` – a função de ligação canônica é automaticamente ajustada para `logit`. Para alterar a função de ligação, é necessário acrescentar ainda `link=linkfunction`, substituindo `linkfunction` por `probit`, `cauchit`, `log` ou `cloglog`. Com a função de ligação logit, a média será modelada em função dos preditores lineares na forma:

$$\mu_i = \frac{1}{1+e^{-(\alpha+\beta_1 x_1 \cdots \beta_n x_n)}} \quad (5)$$

em que: α é o intercepto do modelo; $\beta_1, \ldots, \beta_n$ são os parâmetros estimados para cada fator anunciado ao modelo, incluindo as interações; $x_1, \ldots, x_n$ são os fatores fixos do modelo, incluindo as interações.

2.3.1 A distribuição binomial na germinação de *Peltogyne confertiflora*

A espécie florestal *Peltogyne confertiflora* (Mart. ex Hayne Benth.), também conhecida como quebra-machado, roxinho, jatobá-roxo ou guarabu-roxo, pertence à família Fabaceae e é encontrada nos domínios fitogeográficos da Caatinga, do Cerrado e da Mata Atlântica (SILVA, 1976). Com o intuito de testar um método pré-germinativo para a espécie, testaram-se quatro lotes com potenciais germinativos diferentes. O experimento foi realizado em delineamento inteiramente casualizado contendo quatro lotes (L_1, L_2, L_3 e L_4) e 16 repetições por lote. Cada repetição consistiu em 25 sementes por rolo (n=25). Os dados (Anexo A) constam de quatro colunas, sendo a coluna "`lote`" referente aos quatro lotes; "`rep`", às 16 repetições; "`germ`", ao número de sementes germinadas; e "`ngerm`", ao número de sementes não germinadas.

Convém salientar que a germinação das sementes seguiu uma distribuição binomial, em que o número de sementes que germinaram correspondeu ao evento de sucesso diante de um número de tentativas igual a 25 sementes. Para que os dados sejam lidos no R, os seguintes comandos são executados:

```
1     dados<-read.table("exemplo1.txt", header=T)
2     attach(dados)
3     resp<-cbind(germ, ngerm)
4     str(dados)
```

As linhas 1 e 2 importam os dados do experimento, ao passo que a linha 3 combina as colunas que contêm as sementes germinadas com as não germinadas, criando uma variável chamada de "`resp`". O comando é realizado, pois, na modelagem binomial, é necessário o número de acertos e erros do experimento – a função `str()` mostra os níveis de cada variável para certificar que a leitura foi correta. O MLG com distribuição binomial e função de ligação logit é executado com o comando:

```
5       bin<-glm(resp ~ lote, family = binomial, data = dados)
6       summary(bin)
7       anova(bin, test="Chisq")
```

Já a linha 5 executa o MLG e o denomina como "`bin`". A linha 6 realiza a análise de *deviance* (ANODEV) para testar a diferença entre os fatores pelo teste de qui-quadrado. Para os modelos binomial e Poisson, o teste de qui-quadrado é feito para as inferências sob os fatores, ao invés do teste de *F*, empregado para dados normais, pois o parâmetro de dispersão é estimado a partir dos dados. A função exibe o seguinte resultado:

```
Call:
glm(formula = resp ~ lote, family = binomial, data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4001  -0.7490  -0.1253   0.6319   4.8108

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7002     0.1030  -6.797 1.07e-11 ***
loteL2        2.5720     0.1835  14.013  < 2e-16 ***
loteL3        1.2217     0.1460   8.369  < 2e-16 ***
loteL4        0.8606     0.1438   5.985 2.17e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354.131  on 63  degrees of freedom
Residual deviance:  95.275  on 60  degrees of freedom
AIC: 319.23

Number of Fisher Scoring iterations: 4
```

Da saída, os coeficientes do modelo são apresentados. Caso os lotes fossem descritos apenas por números (1 a 4), o R os entenderia como um fator quantitativo e, por conseguinte, apenas um coeficiente seria demonstrado, por isso a nomeação de `L1` a `L4`. O primeiro tratamento do fator, na ordem alfabética, é sempre fixado para que os demais coeficientes sejam mensurados – os coeficientes são muito importantes para estimar equações de variáveis quantitativas. Sendo assim, uma regressão logística foi executada para a modelagem dos dados para encontrar a média do lote. Para *P. confertiflora*, a estimação do percentual médio de cada lote se estabeleceu na forma logística (Tabela 3).

Equações possibilitam somente encontrar o valor da média da probabilidade de sucesso de germinação de cada lote. A informação entre parênteses na saída relata que o parâmetro de dispersão foi igual a um. Para os modelos binomial e Poisson, o parâmetro de dispersão é sempre um, o que pode configurar uma sub ou sobredispersão dos dados.

TABELA 3. Relação entre a média da probabilidade de germinação para cada lote de *Peltogyne confertiflora*, de acordo com os parâmetros estimados de um MLG com distribuição binomial e função de ligação logit.

Lote	Estimativa da média
L_1	$\mu_{L1} = \frac{1}{1+e^{-(-0.7002)}} = 0.3318$
L_2	$\mu_{L2} = \frac{1}{1+e^{-(-0.7002+2.572)}} = 0.8667$
L_3	$\mu_{L3} = \frac{1}{1+e^{-(-0.7002+1.2217)}} = 0.6275$
L_4	$\mu_{L4} = \frac{1}{1+e^{-(-0.7002+0.8606)}} = 0.5400$

Para cada estimativa de um parâmetro ao modelo, há uma estatística com "`z-value`", que se refere ao teste de Wald com distribuição z bicaudal para testar a hipótese de estimativa nula. A hipótese nula (H_0) é de que a estimativa possui distribuição normal com média zero e desvio padrão igual a um – a coluna "`Pr(>|z|)`" mostra a significância do teste.

A função `summary` ainda apresenta *deviances* do modelo nulo e residual, mas o teste com o *deviance* dos tratamentos (ANODEV) é visualizado apenas com a função `anova`. Por fim, o valor de AIC é apresentado (319.23), e ao executar a linha 7, o programa gera o seguinte resultado:

```
Analysis of Deviance Table

Model: binomial, link: logit
Response: resp
Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    63     354.13
lote  3   258.86         60      95.28 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

É possível concluir que o fator lote foi expressivo a 0.001 de significância, mostrando que o método pré-germinativo foi eficiente em distinguir os lotes. Após o teste de significância dos efeitos, pretende-se verificar o lote com o maior número de sementes germinadas, por meio de um teste de médias. Para isso, os pacotes `emmeans` e `multcompView` são utilizados:

```
8     library(emmeans)
9     library(multcompView)
10    media <- emmeans(bin, ~ lote)
```

```
11    medfin<-regrid(media)
12    cld(medfin, alpha=0.05, Letters=letters, adjust="tukey")
```

Linhas 8 e 9 carregam os pacotes; a 10 calcula a média e o erro padrão dos tratamentos na escala logit; a 11 converte os valores logit para a escala original; e a 12 realiza comparações múltiplas com ajuste de Tukey para chegar os valores das médias. O script apresenta a saída:

```
> cld(medfin, alpha=0.05, Letters=letters, adjust="tukey")
 lote      prob         SE  df asymp.LCL asymp.UCL .group
 L1   0.3317647 0.02283943 Inf 0.2748736 0.3886558  a
 L4   0.5400000 0.02491987 Inf 0.4779267 0.6020733   b
 L3   0.6275000 0.02417353 Inf 0.5672858 0.6877142   b
 L2   0.8666667 0.01755410 Inf 0.8229409 0.9103925    c

Confidence level used: 0.95
Conf-level adjustment: sidak method for 4 estimates
P value adjustment: tukey method for comparing a family of 4 estimates
significance level used: alpha = 0.05
```

As médias das probabilidades são apresentadas para cada lote (mesmos valores obtidos da Tabela 3). Ao invés de percentuais de germinação, o sistema gera valores da probabilidade para o evento de sucesso (π_i). Caso haja a necessidade de demonstração em percentuais, basta multiplicar os resultados por 100 – o teste de médias com as respectivas letras pode ser visualizado na Tabela 4:

TABELA 4. Porcentagem de germinação dos diferentes lotes de *Peltogyne confertiflora* ajustados a um MLG com distribuição binomial e função de ligação logit.

Lote	Germinação (%)
L_1	33.18±5.7 c
L_2	86.67±4.4 a
L_3	62.75±6.0 b
L_4	54.00±6.2 b

Médias, acompanhadas pelos intervalos de confiança ajustados pelo método de Šidák e seguidas por letras distintas na coluna, diferem-se entre si pelo teste de Tukey a 0.05 de significância.

O programa utiliza a correção de Šidák (ŠIDÁK, 1967) para ajustar os intervalos de confiança. Tal método é utilizado para contra-atacar a problemática de múltiplas comparações, controlando a taxa de erro da família dos testes (do inglês *familywise error rate – FWER*), referente à probabilidade de rejeição incorreta de ao menos uma das hipóteses nulas que compõem a família. Os novos intervalos de confiança são calculados por:

$$\alpha_S = 1(1-\alpha)^{1/m} \quad (6)$$

em que: α_S: probabilidade ajustada de Šidák; m: número de comparações realizada entre as médias.

Para *P. confertiflora*, α_S é calculado para ajustar os intervalos de confiança. O valor foi obtido da seguinte forma:

$$\alpha_S = 1(1 - 0.05)^{1/4}$$
$$\alpha_S = 1(0.95)^{0.25}$$
$$\alpha_S = 0.012741$$

Nesses termos, o valor de α_S é dividido por 2 (distribuição bicaudal), obtendo-se 0.006371. Com a distribuição normal padrão, coleta-se o valor de z (2.490915). O programa calcula automaticamente esse valor e informa os limites inferior e superior do intervalo de confiança para a média estimada (`asymp.LCL` e `asymp.UCL`). Por exemplo, para o lote L_1 o erro é estimado em ± 0.02283943 × 2.490915.

2.4 Distribuição de Poisson

A distribuição de Poisson é utilizada para variáveis aleatórias discretas e descreve a probabilidade de uma série de eventos ocorrerem em certo período ou espaço, caso cada evento seja independente do anterior. Descrita por Siméon-Denis Poisson em 1837, foi a primeira distribuição que descreve os processos de contagem. A função de distribuição é definida por:

$$\begin{cases} f(y;\mu) = \frac{\mu^y \times e^{-\mu}}{y!}, \text{para } y \geq 0 \text{ e inteiro} \\ 0, \quad \text{para } y < 0 \end{cases} \tag{7}$$

Para valores abaixo das médias, a curva de densidade é distorcida, mas, à medida que a média se torna grande, a distribuição fica simétrica (Figura 1) – apesar de y ser inteiro, a média μ pode assumir valores decimais. Uma particularidade importante dessa distribuição é que a esperança e a variância são iguais à média amostral ($E(Y) = \mu$ e $var(Y) = \mu$. Nota-se vantagem da distribuição Poisson para dados de contagem devido à imposição de probabilidades iguais a zero para dados negativos.

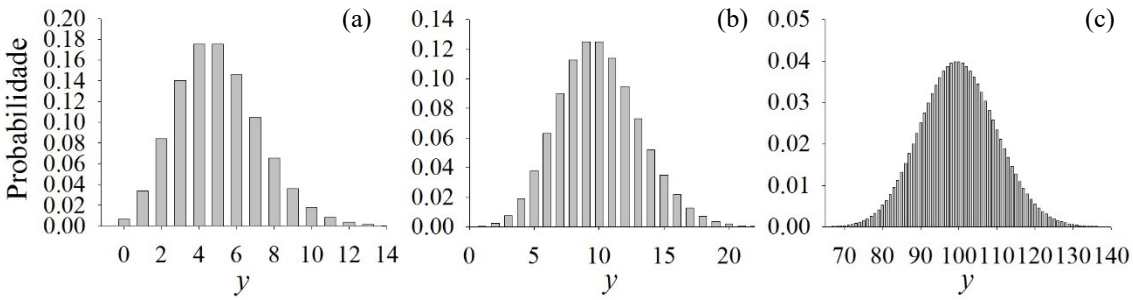


FIGURA 1. Probabilidade de Poisson para μ=5 (a), μ=10 (b) e μ=100 (c).

No ambiente R, a distribuição de Poisson é especificada na função `glm` como `family=poisson`, e a função de ligação canônica é automaticamente ajustada para `log`. Para alterar a função de ligação, deve-se acrescentar ainda `link=linkfunction`, substituindo `linkfunction` por `identity` ou `sqrt`. A relação entre a média e os preditores para função log se dará deste modo:

$$\mu_i = e^{(\alpha+\beta_1 x_1 \cdots \beta_n x_n)} \tag{8}$$

A distribuição foi modificada por diversos autores com acréscimos de parâmetros que exploram a diversidade dos dados, corrigindo problemas de truncamento esquerdo (COHEN, 1960), direito e duplo (COHEN, 1961) e outras generalizações (CONSUL; JAIN, 1973; CONSUL, 1989; CHANDRA et al., 2013). Apesar de algumas dessas generalizações tentarem suavizar desvios da variância em relação à média, a distribuição de Poisson se limita a modelar apenas dados que possuem média e variância iguais ou bem próximas. Na realidade, a distribuição de Poisson dificilmente servirá para dados de contagem ecológicos pela violação desse pressuposto. Para a maioria dos dados ecológicos, a variância é extremamente superior à média, fenômeno que causa a sobredispersão dos dados (ZUUR et al., 2009).

2.4.1 A distribuição de Poisson no controle biológico do pulgão-do-algodoeiro

O pulgão-do-algodoeiro (*Aphis gossypii* Glover) é uma praga cosmopolita amplamente distribuída em regiões tropicais que causa danos ao algodoeiro não só diretos, como também pela transmissão de viroses (KERSTING et al., 1999). Três formulações de produtos biológicos (`B`, `C`, `D`) e uma química (`A`) foram testadas para controle desse pulgão. Além das formulações foi utilizada uma testemunha (`cont`) sem nenhuma aplicação (Anexo B). Como a variação na flutuação populacional do inseto é

alta, dez blocos foram utilizados, totalizando 50 parcelas experimentais. Aos 30 dias, ninfas dos pulgões criadas em laboratório foram distribuídas na área em todas as parcelas; já aos 45 dias, os produtos foram aplicados e, aos 50 dias, contabilizou-se a quantidade de número de pulgões por meio da média da contagem dos pulgões em cinco plantas aleatórias da parcela – o número de pulgões foi inicialmente ajustado à distribuição de Poisson. A abertura dos dados (linhas 1 a 3) e o MLG com distribuição Poisson e função de ligação log (linhas 4 a 6) são executados com os comandos:

```
1     dados<-read.table("exemplo2.txt", header=T)
2     attach(dados)
3     str(dados)
4     poi<-glm(cont ~ trat + bloco, family = poisson, data = dados)
5     summary(poi)
6     anova(poi, test="Chi")
```

A saída do R identifica que todos os coeficientes do modelo foram significativos, assim como a significância para o efeito dos tratamentos (*p-valor*<0.001).

```
Call:
glm(formula = cont ~ trat + bloco, family = poisson, data = dados)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-16.0236   -3.9092   -0.6915    1.6121   12.9630

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.42269    0.06213  55.089  < 2e-16 ***
tratB       -2.24568    0.15339 -14.640  < 2e-16 ***
tratC       -2.28916    0.15644 -14.633  < 2e-16 ***
tratcont     2.79003    0.04889  57.063  < 2e-16 ***
tratD        0.33712    0.06213   5.426 5.76e-08 ***
blocoII     -0.33544    0.06436  -5.212 1.87e-07 ***
blocoIII    -2.37232    0.14229 -16.673  < 2e-16 ***
blocoIV      0.52829    0.05240  10.082  < 2e-16 ***
blocoIX     -3.47093    0.23933 -14.502  < 2e-16 ***
blocoV       1.43228    0.04625  30.965  < 2e-16 ***
blocoVI     -0.63772    0.07068  -9.023  < 2e-16 ***
blocoVII     0.93511    0.04904  19.068  < 2e-16 ***
blocoVIII    0.74366    0.05048  14.732  < 2e-16 ***
blocoX       0.45871    0.05309   8.640  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 25838.4  on 49  degrees of freedom
Residual deviance:  1663.7  on 36  degrees of freedom
AIC: 1895.9

Number of Fisher Scoring iterations: 7

Analysis of Deviance Table
```

```
Model: poisson, link: log
Response: cont
Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     49    25838.4
trat   4  18044.9        45     7793.5 < 2.2e-16 ***
bloco  9   6129.8        36     1663.7 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Na distribuição de Poisson, a relação entre a média e os parâmetros se dá na escala log, podendo ser obtida pelos coeficientes gerados (Tabela 5) – cabe ressaltar que o intercepto fixa o primeiro tratamento por ordem alfabética. Como o experimento é feito em blocos casualizados, o intercepto representa o efeito conjunto tanto do tratamento `A` como do bloco `I`. Para o cálculo das médias, é necessário realizar primeiramente o comando `media <- emmeans(poi, ~trat)`, para calcular o coeficiente isolado do tratamento `A`, o que resulta em 3.1508547. Mais uma vez, a apresentação do cálculo das médias tem caráter apenas ilustrativo, não sendo necessário realizá-lo, uma vez que, pelo comando `regrid`, os mesmos valores são obtidos.

TABELA 5. Relação entre a média da contagem de pulgões-do-algodoeiro (*Aphis gossypii*) por planta com a aplicação de diferentes produtos, de acordo com os parâmetros estimados de um MLG com distribuição de Poisson e função de ligação log.

Tratamento	Relação log para média
A	$\mu = e^{3.1508547} = 23.36$
B	$\mu = e^{3.1508547-2.24568} = 2.47$
C	$\mu = e^{3.1508547-2.28916} = 2.37$
D	$\mu = e^{3.1508547+0.33712} = 32.72$
Controle	$\mu = e^{3.1508547+2.79003} = 380.27$

Para a comparação das médias e as respectivas estimativas, as linhas 7 e 8 carregam os pacotes, enquanto 9 a 11 realizam os testes.

```
7     library (emmeans)
8     library(multcompView)
9     media <- emmeans(poi, ~trat)
10    medfin <- regrid(media)
11    cld(medfin,alpha=0.05, Letters=letters, adjust="tukey")
```

A função `summary` mostra as comparações entre cada par de médias, e `cld` calcula as médias com os respectivos erros padrão, aplicando o teste de médias.

```
> summary(pairs(media), type = "response")
 contrast        ratio           SE  df z.ratio p.value
 A / B     9.446808511 1.4490568667 Inf  14.640  <.0001
 A / C     9.866666667 1.5435728423 Inf  14.633  <.0001
 A / cont  0.061419283 0.0030030097 Inf -57.063  <.0001
 A / D     0.713826367 0.0443490678 Inf  -5.426  <.0001
 B / C     1.044444444 0.2178331928 Inf   0.208  0.9996
 B / cont  0.006501591 0.0009514324 Inf -34.411  <.0001
 B / D     0.075562701 0.0114307939 Inf -17.073  <.0001
 C / cont  0.006224927 0.0009308411 Inf -33.967  <.0001
 C / D     0.072347267 0.0111682103 Inf -17.013  <.0001
 cont / D 11.622186495 0.4856417429 Inf  58.702  <.0001

Results are averaged over the levels of: bloco
P value adjustment: tukey method for comparing a family of 5 estimates
Tests are performed on the log scale
> medfin<-regrid(media)
> cld(medfin,alpha=0.05, Letters=letters, adjust="tukey")
 trat       rate         SE  df  asymp.LCL  asymp.UCL .group
 C      2.367164  0.3586705 Inf   1.445824   3.288503  a
 B      2.472371  0.3668153 Inf   1.530110   3.414633  a
 A     23.356017  1.2767143 Inf  20.076441  26.635594   b
 D     32.719466  1.5839495 Inf  28.650675  36.788257    c
 cont 380.271732 11.2430820 Inf 351.390917 409.152547     d

Results are averaged over the levels of: bloco
Confidence level used: 0.95
Conf-level adjustment: sidak method for 5 estimates
P value adjustment: tukey method for comparing a family of 5 estimates
significance level used: alpha = 0.05
```

Com esse modelo, concluiu-se que os produtos B e C foram eficientes no controle do pulgão-do-algodoeiro. A aplicação química (produto A) teve desempenho intermediário, e o produto D não apresentou desempenho tão eficiente em relação aos demais (Tabela 6). Entretanto, esse modelo feriu o pressuposto de que a variância dos dados é igual à média, o que pode acarretar na sobredispersão.

TABELA 6. Número médio de pulgões-do-algodoeiro (*Aphis gossypii*) por planta com a aplicação de diferentes produtos ajustados a um MLG com distribuição de Poisson e função de ligação log.

Produto	Pulgões por planta
A	23.36±3.28 b
B	2.47±0.94 a
C	2.37±0.92 a
D	32.72±4.07 c
Controle	380.27±28.88 d

Médias, acompanhadas pelos intervalos de confiança ajustados pelo método de Šidák e seguidas por letras distintas na coluna, diferem-se entre si pelo teste de Tukey a 0.05 de significância.

2.5 Sub e sobredispersão nos modelos binomial e Poisson

Sobredispersão indica que a variância dos dados é maior que a média, ao passo que, na subdispersão, a variância dos dados é menor que a média. Como a variância para

as distribuições binomial e Poisson é modelada exclusivamente pela média, quando a variância for superior à média, ocorre sobredispersão, e o modelo se torna impreciso. Nesses casos, a sobredispersão pode ser aparente, em virtude da falta de covariáveis ou interações não estudadas, *outliers*, efeitos não lineares considerados lineares no componente sistemático e inadequação da função de ligação; e real, quando não é possível identificar nenhuma causa anterior. Essa última ocorre devido ao fato de a variação dos dados ser, por natureza, maior que a média, e quando há observações agrupadas ou correlação entre elas (HILBE, 2011).

A sobredispersão de um modelo pode ser baseada na aproximação χ^2 do *deviance* residual (D_{res}). Se existir sobredispersão, $\frac{D_{res}}{\phi}$ possui distribuição qui-quadrado com $n - p$ graus de liberdade, o que estima o parâmetro de dispersão:

$$\hat{\phi} = \frac{D_{res}}{gl_{res}};$$
$$\begin{cases} \text{se } \hat{\phi} > 1, \text{dados sofrem de sobredispersão} \\ \text{se } \hat{\phi} < 1, \text{dados sofrem de subdispersão} \end{cases} \quad (9)$$

De fato, a verificação da sobredispersão para os modelos binomial e de Poisson é obrigatória após a modelagem, pois, caso a sub ou sobredispersão seja constatada, métodos estatísticos devem ser empregados para ser corrigida. Dificilmente, a estimação do parâmetro de dispersão será igual a um; logo, cabe ao pesquisador inferir se o desvio do valor estimado é elevado de maneira suficiente para produzir uma sobredispersão aos dados.

Para dados com distribuição de Poisson, a função `dispersiontest` do pacote `AER` testa a hipótese nula de equidispersão no modelo Poisson contra a hipótese alternativa de sub ou sobredispersão (CAMERON; TRIVEDI, 1990) – tal função não pode ser utilizada para dados binomiais. Em um MLG binomial, testa-se a sobredispersão pelo teste de qui-quadrado ao verificar se o modelo quasi-binomial difere do modelo binomial.

2.5.1 Sobredispersão na germinação de *Peltogyne confertiflora*

Para a germinação de *P. confertiflora*, a estimação da sobredispersão foi maior que um ($\hat{\phi} = \frac{95.275}{60} = 1.588$); porém, não é uma sobredispersão expressiva. Mesmo assim, será ajustado o modelo de distribuição quasi-binomial para comparar o modelo binomial com o quasi-binomial por meio do teste qui-quadrado.

2.5.2 Sobredispersão no controle biológico do pulgão-do-algodoeiro

Para os dados da contagem de pulgões no modelo Poisson, estimou-se a sobredispersão como $\hat{\phi} = \frac{1663.7}{36} = 46.213$. A sobredispersão foi tão elevada que, dificilmente, a falha do modelo será corrigida de maneira integral. A hipótese de sobredispersão foi verificada no R:

```
1     library(AER)
2     dispersiontest(poi,trafo=1)
```

Da saída, a hipótese da sobredispersão dos dados foi aceita (*p-valor*<0.01), o que foi justificado pela alta variação dos valores que ocorre em razão do rápido potencial reprodutivo da espécie. Nesse sentido, a sobredispersão pode ser corrigida com a distribuição quasi-Poisson ou binomial negativa.

```
Overdispersion test

data:  poi
z = 3.5498, p-value = 0.0001928
alternative hypothesis: true alpha is greater than 0
sample estimates:
 alpha
42.52358
```

2.6 Quasi-verossimilhança

Uma das alternativas para modelagem de dados sobredispersos foi proposta por Wedderburn (1974), que definiu as funções de quasi-verossimilhança para exibir propriedades semelhantes ao *log* da máxima verossimilhança, mas sem possuir correspondência com nenhuma distribuição de probabilidade – os modelos são caracterizados apenas por média e variância. A formulação de um modelo quasi deixa os parâmetros em um estado natural e interpretável e permite diagnósticos padrões sem a perda de eficiência no ajuste de algoritmos (HOEF; BOVENG, 2007).

Com essa abordagem, as distribuições binomial ou de Poisson não são explicitamente especificadas. Apesar de a distribuição não ser especificada, a mesma estrutura do modelo para a função de ligação e preditor é mantida (ZUUR et al., 2009). A função de quasi-verossimilhança permite incluir um fator multiplicativo, isto é, o

parâmetro de sobredispersão (do inglês *overdispersion parameter* ou *scale parameter*), o qual é estimado a partir dos dados.

2.6.1 Distribuição quasi-binomial (DQB)

Consul desenvolveu a distribuição quasi-binomial em 1974 e detalhou as propriedades em 1990. A função de probabilidade do modelo quasi-binomial leva em consideração o parâmetro de dispersão estimado:

$$f(y;\pi) = \binom{n}{y} \times \pi\ (\pi + y\hat{\phi})^{y-1} \times \left(1 - \pi - y\hat{\phi}\right)^{n-y} \qquad (10)$$

Quando $\hat{\phi} = 0$, o modelo se iguala ao da distribuição binomial. A esperança e a variância da distribuição são representadas por $E(Y) = n \times \pi$ e $var(Y) = \hat{\phi} \times n \times \pi \times (1 - \pi)$ e, para a distribuição quasi-binomial no R, a função `glm` é especificada em `family=quasibinomial`.

2.6.1.1 Corrigindo a sobredispersão na germinação de *Peltogyne confertiflora*

Comandos executados anteriormente no R para o modelo binomial são os mesmos para o modelo quasi-binomial, alterando apenas o parâmetro `family` para `quasibinomial`. Da mesma forma, os coeficientes do modelo são estimados, e o teste de significância dos fatores é executado pelo comando `anova`.

```
1     dados<-read.table("exemplo1.txt", header=T)
2     attach(dados)
3     resp<-cbind(germ, ngerm)
4     str(dados)
5     quasibin<-glm(resp~lote, family = quasibinomial, data = dados)
6     summary(quasibin)
7     anova(quasibin, test="Chisq")
8     library (emmeans)
9     library(multcompView)
10    media <- emmeans(quasibin, ~ lote)
11    summary(pairs(media), type = "response")
12    medfin<-regrid(media)
13    cld(medfin, alpha=0.05, Letters=letters, adjust="tukey")
```

A saída diferiu do modelo binomial, pois houve alteração na significância de fatores, coeficientes e erros padrão. Além disso, o AIC para os modelos quasi não foi calculado, pois eles não possuem verossimilhança. A significância do fator `lote` foi mantida no modelo quasi-binomial:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4001  -0.7490  -0.1253   0.6319   4.8108

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7002     0.1294  -5.413 1.14e-06 ***
loteL2        2.5720     0.2305  11.160 2.85e-16 ***
loteL3        1.2217     0.1833   6.665 9.28e-09 ***
loteL4        0.8606     0.1806   4.766 1.23e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.576621)

    Null deviance: 354.131  on 63  degrees of freedom
Residual deviance:  95.275  on 60  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4


Analysis of Deviance Table
Model: quasibinomial, link: logit
Response: resp
Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    63     354.13
lote  3   258.86         60      95.28 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



lote      prob         SE  df asymp.LCL asymp.UCL .group
 L1   0.3317647 0.02867801 Inf 0.2603302 0.4031992  a
 L4   0.5400000 0.03129028 Inf 0.4620586 0.6179414   b
 L3   0.6275000 0.03035314 Inf 0.5518929 0.7031071   b
 L2   0.8666667 0.02204156 Inf 0.8117630 0.9215703    c

Confidence level used: 0.95
Conf-level adjustment: sidak method for 4 estimates
P value adjustment: tukey method for comparing a family of 4 estimates
significance level used: alpha = 0.05
```

Resultados obtidos pelo teste de médias foram mantidos (Tabela 7), e apenas o erro-padrão foi alterado, sofrendo um pequeno aumento.

TABELA 7. Porcentagem de germinação dos diferentes lotes de *Peltogyne confertiflora*, ajustados aos modelos binomial e quasi-binomial, com função de ligação logit.

	Binomial	**Quasi-binomial**
Lote	**Germinação (%)**	
L_1	33.18±5.7 c	33.18±7.1 c
L_2	86.67±4.4 a	86.67±5.5 a
L_3	62.75±6.0 b	62.75±7.6 b
L_4	54.00±6.2 b	54.00±7.8 b

Médias, acompanhadas pelos intervalos de confiança ajustados pelo método de Šidák e seguidas por letras distintas na coluna, diferem-se entre si pelo teste de Tukey a 0.05 de significância.

O modelo binomial previamente ajustado (Tópico 2.3.1) foi comparado ao modelo quasi-binomial. Para isso, executou-se o seguinte comando:

```
14    pchisq(summary(quasibin)$dispersion*bin$df.residual,
      bin$df.residual, lower = F)
```

Diante disso, o comando apresentou o *p-valor* de 0.0029, mostrando diferença entre os modelos, em que se opta pelo ajuste ao modelo quasi-binomial. A análise dos resíduos após o ajuste dos modelos é importante para inferir sobre a qualidade desse ajuste. Há três tipos principais de resíduos para a checagem do modelo nos MLGs: os ordinários; os de Pearson e os desviados. Entretanto, encontram-se na literatura diversos outros tipos de resíduos (MCCULLAGH; NELDER, 1989).

Neste capítulo será apresentado apenas o cálculo dos resíduos de Pearson, aplicável a todos os modelos, pois os resíduos desviados para os modelos de zeros inflacionados só podem ser obtidos pela técnica de *bootstrapping*, e os resíduos ordinários não consideram a dispersão dos dados – aqui, os resíduos são plotados em relação aos valores preditos do modelo. O resíduo de Person por observação ($\hat{\varepsilon}_i^P$) é obtido por:

$$\hat{\varepsilon}_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{var(Y_i)}} \tag{11.1}$$

$$\hat{\varepsilon}_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} var(Y_i)}} \tag{11.2}$$

Enquanto a Equação 11.1 é utilizada para os modelos com distribuição binomial, Poisson e binomial negativa, a Equação 11.2 é utilizada para os modelos quasi, em que a variância é multiplicada pelo parâmetro de dispersão estimado, e isso reflete na redução do valor dos resíduos. Para o modelo binomial, o gráfico foi criado com os seguintes comandos:

```
15    Resíduos<- resid(bin, type = "pearson")
16    Valores_preditos<-predict(bin, type = "response")
17    plot(x = Valores_preditos, y = Resíduos, main = "Modelo
binomial", xlim=c(0,1), ylim=c(-2,6))
```

A linha 15 calcula os resíduos de Pearson; a 16 estima os valores preditos; e a 17 plota em um gráfico os dois valores, considerando o intervalo de 0 a 1 para os valores preditos (`xlim=c(0,1)`), e de -2 a 6, aos resíduos de Pearson (`ylim=c(-2,6)`). Como a função `resid` não considera o parâmetro de dispersão para escalar os resíduos de Pearson, os seguintes comandos foram executados:

```
18    res<-resid(quasibin, type = "response")
19    Valores_preditos <- predict(quasibin, type = "response")
20    Resíduos <- res / sqrt(1.576621 * Valores_preditos)
21    plot(x = Valores_preditos, y = Resíduos, main = "Modelo quasi-
      binomial",xlim=c(0,1), ylim=c(-2,6))
```

Por sua vez, a linha 18 calcula os resíduos ordinários ($y_i - \hat{\mu}_i$), e a linha 20 calcula os resíduos de Pearson, inserindo o parâmetro de dispersão na fórmula (1.576621) – a plotagem foi realizada na mesma escala dos resíduos do modelo binomial (Figura 2). A análise gráfica mostrou como o parâmetro de dispersão auxilia na modelagem da variância dos dados e, como os resíduos de Pearson levaram em consideração a variância dos dados, o modelo quasi é vantajoso por adicionar um parâmetro para controlar a sobredispersão dos dados. Se a sobredispersão não fosse considerada, seria interessante remover *outliers* no banco de dados para posteriormente ajustar os modelos (o modelo binomial possui observações com resíduo muito elevado).

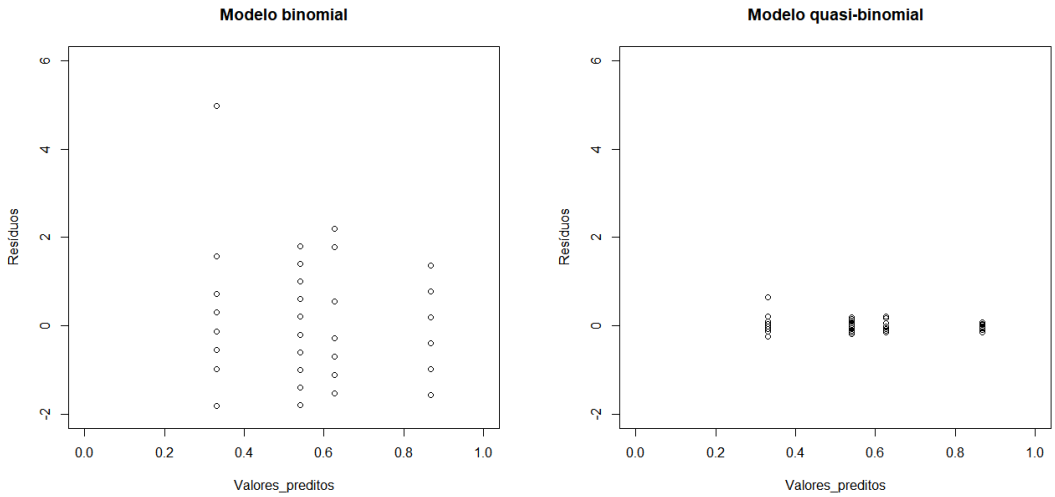


FIGURA 2. Resíduos de Pearson plotados em relação aos valores preditos da germinação de diferentes lotes de *Peltogyne confertiflora*, ajustados aos modelos binomial e quasi-binomial com função de ligação logit.

2.6.2 Distribuição quasi-Poisson (DQP)

O modelo quasi-Poisson insere um novo parâmetro ao modelo e, com isso, os erros padrão são multiplicados por $\sqrt{\hat{\phi}}$, tornando menor a significância dos parâmetros. Assume-se para o modelo quasi-Poisson que $E(Y) = \mu$ e $var(Y) = \hat{\phi}\mu$; se os parâmetros do modelo de Poisson forem altamente significativos, a introdução do parâmetro de dispersão para corrigir a sobredispersão não afetará o modelo. Porém, outras metodologias devem ser utilizadas, caso $\hat{\phi}$ seja maior que 15 (ZUUR et al., 2009). Para R, o modelo quasi-Poisson pode ser também estimado com a função `glm()`, ajustando-se para `family = quasipoisson`.

2.6.2.1 Corrigindo a sobredispersão no controle biológico do pulgão-do-algodoeiro

Com a intenção de ajustar a distribuição quasi-Poisson, o parâmetro `family` é alterado para `quasipoisson`. Os coeficientes do modelo foram estimados e se executou o teste de significância dos fatores pelo comando `anova`.

```
1     dados<-read.table("exemplo2.txt", header=T)
2     attach(dados)
3     str(dados)
4     quasipoi<-glm(cont ~ trat + bloco, family = quasipoisson, data =
dados)
```

```
5     summary(quasipoi)
6     anova(quasipoi, test="Chi")
7     library (emmeans)
8     library(multcompView)
7     media <- emmeans(quasipoi, ~trat)
8     summary(pairs(media), type = "response")
9     medfin<-regrid(media)
10    medfin
11    cld(medfin,alpha=0.05, Letters=letters, adjust="tukey")
```

O parâmetro de dispersão, igual a um na distribuição de Poisson, foi estimado em 61.41471. Apesar do aumento do parâmetro, a significância do fator em questão foi mantida. Entretanto, o elevado valor atribuído a $\hat{\phi}$ apontou para a necessidade de empregar outras técnicas estatísticas.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-16.0236   -3.9092   -0.6915    1.6121   12.9630

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4227     0.4869   7.030 3.00e-08 ***
tratB        -2.2457     1.2021  -1.868 0.069898 .
tratC        -2.2892     1.2260  -1.867 0.070038 .
tratcont      2.7900     0.3832   7.282 1.41e-08 ***
tratD         0.3371     0.4869   0.692 0.493135
blocoII      -0.3354     0.5044  -0.665 0.510272
blocoIII     -2.3723     1.1151  -2.128 0.040297 *
blocoIV       0.5283     0.4106   1.287 0.206461
blocoIX      -3.4709     1.8756  -1.851 0.072452 .
blocoV        1.4323     0.3625   3.951 0.000347 ***
blocoVI      -0.6377     0.5539  -1.151 0.257164
blocoVII      0.9351     0.3843   2.433 0.020065 *
blocoVIII     0.7437     0.3956   1.880 0.068237 .
blocoX        0.4587     0.4161   1.102 0.277568
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 61.41471)

    Null deviance: 25838.4  on 49  degrees of freedom
Residual deviance:  1663.7  on 36  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7

Analysis of Deviance Table
Model: quasipoisson, link: log
Response: cont
Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     49    25838.4
trat   4  18044.9         45     7793.5 < 2.2e-16 ***
bloco  9   6129.8         36     1663.7 < 2.2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


trat       rate        SE  df   asymp.LCL  asymp.UCL .group
 C      2.367164  2.810812 Inf  -4.8531476   9.587475  a
 B      2.472371  2.874641 Inf  -4.9119011   9.856644  a
 A     23.356017 10.005296 Inf  -2.3452176  49.057252  a
 D     32.719466 12.413022 Inf   0.8333509  64.605580  a
 cont 380.271732 88.109266 Inf 153.9398935 606.603570   b

Results are averaged over the levels of: bloco
Confidence level used: 0.95
Conf-level adjustment: sidak method for 5 estimates
P value adjustment: tukey method for comparing a family of 5 estimates
significance level used: alpha = 0.05
```

Pela alta sobredispersão, foi possível verificar o aumento nos erros padrão. No modelo quasi-Poisson, os tratamentos foram diferentes apenas do controle, ao contrário do que foi observado pelo modelo Poisson. Em razão do alto valor atribuído ao parâmetro de dispersão, os dados também serão ajustados à distribuição binomial negativa que, por sua vez, insere um parâmetro extra ao modelo.

TABELA 8. Número médio de pulgões-do-algodoeiro (*Aphis gossypii*) por planta com a aplicação de diferentes produtos ajustados aos modelos de Poisson e quasi-Poisson com função de ligação log.

	Poisson	**Quasi-Poisson**
Produto	**Pulgões por planta**	
A	23.36±3.28 b	23.36±25.70 a
B	2.47±0.94 a	2.47±7.38 a
C	2.37±0.92 a	2.37±7.22 a
D	32.72±4.07 c	32.72±31.89 a
Controle	380.27±28.88 d	380.27±226.33 b

Médias, acompanhadas pelos intervalos de confiança ajustados pelo método de Šidák e seguidas por letras distintas na coluna, diferem-se entre si pelo teste de Tukey a 0.05 de significância.

2.7 Distribuição binomial negativa (BN)

A distribuição binomial negativa é a combinação das distribuições de Poisson e gamma, assumindo que os valores de *y* possuem distribuição Poisson e que a média μ segue distribuição gamma. Descrita por Fisher em 1941, Bliss e Fisher (1953) apontaram a necessidade desta distribuição para dados ecológicos oriundos de contagem. A distribuição possui dois parâmetros: μ e *k*.

$$E(Y_i) = \mu_i \qquad \text{com } Y_i \sim BN(\mu_i, k) \qquad (12)$$
$$var(Y_i) = \mu_i + \frac{\mu_i^2}{k}$$

Nesse contexto, o segundo termo da variância (μ_i^2/k) determina o tamanho da sobredispersão – quanto menor o valor de *k*, maior a sobredispersão. Se o valor de *k* for bem maior que μ_i^2, o termo se aproxima de zero e a variância é igual a média, assim como na distribuição de Poisson (ZUUR et al., 2009). A regressão log-linear BN especifica que a probabilidade de distribuição da contagem é binomial negativa com média μ e parâmetro *k* com $\log(\mu) = X'\beta$. Assim, com a função de ligação canônica logarítmica, os valores ajustados serão sempre não negativos:

$$\log(\mu_i) = \eta(\beta_1 x_1, \cdots, \beta_n x_n) \text{ ou } \mu_i = e^{\eta(\beta_1 x_1, \cdots, \beta_n x_n)} \quad (13)$$

O MLG com distribuição binomial negativa pode ser ajustado pela função `glm.nb()` do pacote `MASS`, da mesma forma que a função `glm`, sem a necessidade de especificar a distribuição na função. Outras funções de ligação disponíveis para a distribuição são a identidade (`link=identity`) e a raiz quadrada (`link=sqrt`). O pacote `MASS` também fornece a opção de especificar a família como `negative.binomial()` e ajustá-la na função `glm()`, se o argumento *k* for especificado. Se *k* for desconhecido e deva ser estimado pelos dados, não há a possibilidade de usar a função `glm()`. Para esses casos, apenas a função `glm.nb()` pode ser empregada.

2.7.1 A distribuição binomial negativa no controle biológico do pulgão-do-algodoeiro

Na busca de um modelo que corrija a alta sobredispersão dos dados, a contagem do número de pulgões se ajustou ao modelo binomial negativo.

```
1     dados<-read.table("exemplo2.txt", header=T)
2     attach(dados)
3     str(dados)
4     library(MASS)
5     bn<-glm.nb(cont ~ trat + bloco, data = dados)
6     summary(bn)
7     anova(bn, test="Chi")
```

A linha 4 carrega o pacote `MASS`; 5 ajusta o modelo binomial negativo, chamando-o de `bn`; e 6 e 7 apresentam coeficientes e análise de *deviance*. O modelo atribuiu o valor de *k* como 0.608 e manteve as diferenças significativas para os tratamentos.

```
glm.nb(formula = cont ~ trat + bloco, data = dados, init.theta = 0.6081288121,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4846  -0.7883  -0.2277   0.1797   1.3231

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.0076     0.6950   5.767 8.09e-09 ***
tratB        -2.5851     0.6117  -4.226 2.37e-05 ***
tratC        -2.6681     0.6140  -4.346 1.39e-05 ***
tratcont      2.2315     0.5775   3.864 0.000112 ***
tratD         0.2863     0.5798   0.494 0.621509
blocoII      -0.3949     0.8404  -0.470 0.638453
blocoIII     -1.6672     0.8724  -1.911 0.055997 .
blocoIV       0.5032     0.8307   0.606 0.544672
blocoIX      -2.4896     0.9087  -2.740 0.006147 **
blocoV        0.5041     0.8307   0.607 0.543986
blocoVI       0.0487     0.8347   0.058 0.953478
blocoVII      0.3533     0.8319   0.425 0.671057
blocoVIII     0.4497     0.8311   0.541 0.588470
blocoX        0.3801     0.8316   0.457 0.647631
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.6081) family taken to be 1)

Null deviance: 169.083  on 49  degrees of freedom
Residual deviance: 56.957  on 36  degrees of freedom
AIC: 456.24

Number of Fisher Scoring iterations: 1

              Theta:  0.608
          Std. Err.:  0.136

2 x log-likelihood:  -426.238

Analysis of Deviance Table
Model: Negative Binomial(0.6081), link: log
Response: cont
Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     49    169.083
trat   4   96.373        45     72.710  < 2e-16 ***
bloco  9   15.754        36     56.957  0.07221 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In anova.negbin(bn, test = "Chi") :
tests made without re-estimating 'theta'
```

Diferenças significativas detectadas para o fator tratamento, as médias foram desdobradas e comparadas:

```
8     library (emmeans)
9     library(multcompView)
10    media <- emmeans(bn, ~trat)
11    medfin<-regrid(media)
12    cld(medfin,alpha=0.05, Letters=letters, adjust="tukey")
```

Dos comandos, apresenta-se a saída com o teste de médias.

```
trat   response         SE  df   asymp.LCL  asymp.UCL .group
 C      3.028931   1.384764 Inf  -0.5281986   6.586061  a
 B      3.290997   1.494269 Inf  -0.5474259   7.129419  a
 A     43.653365  17.926563 Inf  -2.3957293  89.702460  a
 D     58.121722  23.796090 Inf  -3.0047979 119.248241  a
 cont 406.570564 165.102540 Inf -17.5387592 830.679887  a

Results are averaged over the levels of: bloco
Confidence level used: 0.95
Conf-level adjustment: sidak method for 5 estimates
P value adjustment: tukey method for comparing a family of 5 estimates
significance level used: alpha = 0.05
```

É visível o aumento das médias dos tratamentos e do erro padrão (Tabela 9), e, apesar de diferenças visivelmente discrepantes, o teste não detectou diferenças entre os produtos e o controle. A distribuição binomial negativa é adequada aos dados, mas não corrigiu o problema de excesso de zeros no banco de dados. Assim, para que o modelo esteja adequado, os zeros precisam também ser modelados.

TABELA 9. Número médio de pulgões-do-algodoeiro (*Aphis gossypii*) por planta após a aplicação de produtos, ajustado aos modelos de Poisson, quasi-Poisson (QP) e binomial negativa (BN).

Produto	**Poisson**	**QP**	**BN**
	Pulgões por planta		
A	23.36±3.28 b	23.36±25.70 a	43.65±46.05 a
B	2.47±0.94 a	2.47±7.38 a	3.29±3.84 a
C	2.37±0.92 a	2.37±7.22 a	3.03±3.56 a
D	32.72±4.07 c	32.72±31.89 a	58.12±61.13 a
Controle	380.27±28.88 d	380.27±226.33 b	406.57±424.11 a

Todos os modelos foram ajustados com a função de ligação logarítmica. Médias, acompanhadas pelos intervalos de confiança ajustados pelo método de Šidák e seguidas por letras distintas na coluna, diferem-se entre si pelo teste de Tukey a 0.05 de significância.

2.8 Modelo fatorial com tratamento quantitativo – Aplicação da Regressão Logística

Com o objetivo de expandir a teoria dos MLGs para ensaios fatoriais com tratamentos quantitativos, será apresentado um exemplo com controle de plantas infestantes (Anexo C). O experimento descreve uma situação muito comum na área da Fitotecnia, na qual doses de determinado produto são utilizadas para o controle de pragas, patógenos ou plantas infestantes. Nesse caso, as doses foram ajustadas a uma regressão logística.

Nesse caso, caruru-roxo (*Amaranthus hybridus*), corda-de-viola (*Ipomoea grandifolia*) e amendoim-bravo (*Euphorbia heterophylla*) são plantas infestantes de grande impacto econômico em diversas culturas de interesse. Para controlá-las, um herbicida foi desenvolvido e testado em sete concentrações do ingrediente ativo: 0 mg L^{-1}, 5 mg L^{-1}, 10 mg L^{-1}, 15 mg L^{-1}, 20 mg L^{-1}, 25 mg L^{-1} e 30 mg L^{-1}. As parcelas experimentais foram dispostas em blocos casualizados, com quatro blocos, em um esquema fatorial 3 x 7, e cada planta infestante foi semeada com a condução de 40 plantas por parcela em uma área de 4m² (2 x 2m). Plantas infestantes foram semeadas em campo, e qualquer outra espécie de planta infestante que germinasse na parcela foi eliminada por arranquio. Aplicou-se o herbicida quando as plantas apresentavam dois pares de folhas, e, 12 dias após a semeadura, foi verificada a eficiência dos herbicidas, contando-se o número de plantas mortas. No arquivo, a coluna "`morta`" representa a quantidade de plantas que morreram com a aplicação do herbicida, e a coluna "`viva`" se refere às que sobreviveram. Inicialmente, o modelo binomial foi ajustado com os seguintes comandos:

```
1     dados<-read.table("exemplo3.txt", header=T)
2     str(dados)
3     dados <- transform(dados,d=Dose, D=factor(Dose))
4     str(dados)
5     attach(dados)
6     resp<-cbind(morta, viva)
```

Os comandos das linhas 3 e 4 são necessários para R entender que o fator dose é quantitativo, e não qualitativo. O comando `resp` da linha 6 combina os eventos de sucesso (quantidade de plantas mortas) e de fracasso (número de plantas vivas). Assim, o MLG com distribuição binomial e função de ligação logit pode ser executado desta maneira:

```
7     bin<-glm(resp~Especie*D+Bloco, family=binomial, data=dados)
8     summary(bin)
9     anova(bin, test="Chisq")
```

Com os resultados obtidos do modelo binomial foi verificado o efeito significativo da interação. Entretanto, estima-se primeiramente se há sobredispersão nos dados: $\hat{\phi} = \frac{25.796}{60} = 0.430$.

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2349.770  on 83  degrees of freedom
```

```
Residual deviance:   25.796  on 60  degrees of freedom
AIC: 272.08
```

Mesmo com menor frequência, os dados também podem apresentar subdispersão, fenômeno inverso ao da sobredispersão (quando $\hat{\phi}$<1). Da mesma forma, a subdispersão precisa ser modelada e corrigida no modelo, e, para isso, utilizou-se a distribuição quasi-binomial. Os dados foram ajustados a um MLG com distribuição quasi-binomial e função de ligação logit.

```
10    quasibin<-glm(resp~Especie*D+Bloco,family=quasibinomial,
      data=dados)
11    pchisq(summary(quasibin)$dispersion*bin$df.residual,
      bin$df.residual, lower = F)
```

O comando apresenta o teste de qui-quadrado com *p-valor* de 0.999, mostrando que os modelos não diferem, com opção pelo modelo binomial. Para estudar a interação, primeiramente se fixou o fator dose, e as plantas infestantes foram estudadas em cada dose. Os pacotes para comparação de médias foram carregados e se conferiu a identificação dos tratamentos pelo programa com as linhas 14 e 15.

```
12    library(emmeans)
13    library(multcompView)
14    referencia <- ref_grid(bin)
15    referencia
```

```
> referencia
'emmGrid' object with variables:
    Especie = Amend, Caruru, Corda
    D = 0, 5, 10, 15, 20, 25, 30
    Bloco = I, II, III, IV
Transformation: "logit"
```

Todos os níveis de espécies e doses foram identificados corretamente – o teste de médias irá fixar a dose e verificar a espécie que foi melhor controlada. Ao executar o teste de médias (e como a interação foi significativa), adiciona-se o fator fixado na função `emmeans`, colocando, após o fator estudado: `by = "D"`. Se não ocorrer essa especificação, a função irá desdobrar o fator `Especie` isolado, como se a interação não fosse significativa. Para cada dose, o programa apresenta a espécie que possuiu maior porcentagem de controle (Tabela 10).

```
16    media <- emmeans(bin, "Especie", by = "D")
17    medfin<-regrid(media)
18    cld(medfin, alpha=0.05, Letters=letters, adjust="tukey")
```

```
D = 0:
 Especie         prob           SE  df     asymp.LCL    asymp.UCL .group
 Caruru  6.782998e-11 3.938481e-07 Inf -9.403381e-07 9.404738e-07  a
 Corda   6.782999e-11 3.938481e-07 Inf -9.403382e-07 9.404738e-07  a
 Amend   6.782999e-11 3.938481e-07 Inf -9.403382e-07 9.404738e-07  a

D = 5:
 Especie         prob           SE  df     asymp.LCL    asymp.UCL .group
 Caruru  9.960980e-02 2.364918e-02 Inf  4.314176e-02 1.560779e-01  a
 Corda   3.120585e-01 3.666067e-02 Inf  2.245225e-01 3.995946e-01   b
 Amend   3.746770e-01 3.831002e-02 Inf  2.832027e-01 4.661513e-01   b

D = 10:
 Especie         prob           SE  df     asymp.LCL    asymp.UCL .group
 Caruru  1.058446e-01 2.429555e-02 Inf  4.783322e-02 1.638560e-01  a
 Amend   4.937284e-01 3.957920e-02 Inf  3.992237e-01 5.882332e-01   b
 Corda   5.062625e-01 3.957924e-02 Inf  4.117577e-01 6.007674e-01   b

D = 15:
 Especie         prob           SE  df     asymp.LCL    asymp.UCL .group
 Caruru  2.432345e-01 3.393125e-02 Inf  1.622156e-01 3.242535e-01  a
 Amend   6.315799e-01 3.817804e-02 Inf  5.404207e-01 7.227390e-01   b
 Corda   7.004559e-01 3.624139e-02 Inf  6.139210e-01 7.869909e-01   b

D = 20:
 Especie         prob           SE  df     asymp.LCL    asymp.UCL .group
 Caruru  4.310619e-01 3.920152e-02 Inf  3.374589e-01 5.246648e-01  a
 Amend   7.505100e-01 3.422489e-02 Inf  6.687899e-01 8.322301e-01   b
 Corda   9.937835e-01 6.197517e-03 Inf  9.789855e-01 1.008582e+00    c

D = 25:
 Especie         prob           SE  df     asymp.LCL    asymp.UCL .group
 Caruru  7.192305e-01 3.555020e-02 Inf  6.343459e-01 8.041150e-01  a
 Corda   9.813468e-01 1.067054e-02 Inf  9.558684e-01 1.006825e+00   b
 Amend   9.937835e-01 6.197517e-03 Inf  9.789855e-01 1.008582e+00   b

D = 30:
 Especie         prob           SE  df     asymp.LCL    asymp.UCL .group
 Caruru  9.937835e-01 6.197517e-03 Inf  9.789855e-01 1.008582e+00  a
 Corda   9.937835e-01 6.197517e-03 Inf  9.789855e-01 1.008582e+00  a
 Amend   1.000000e+00 3.938344e-07 Inf  9.999991e-01 1.000001e+00  a

Results are averaged over the levels of: Bloco
Confidence level used: 0.95
Conf-level adjustment: sidak method for 3 estimates
P value adjustment: tukey method for comparing a family of 3 estimates
significance level used: alpha = 0.05
```

Da porcentagem de controle das plantas infestantes, o caruru apresentou maior resistência ao produto, se comparado a outras espécies; porém, na dosagem de 30 mg L^{-1}, tal produto foi eficiente no controle das três plantas. Diante de cada espécie, ajustou-se uma regressão logística, e, caso o modelo quasi-binomial fosse escolhido, os intervalos de confiança para as médias seriam menores do que os demonstrados para o modelo binomial.

TABELA 10. Porcentagem média do controle de três plantas infestantes submetidas a doses de herbicida.

Espécie	*Dose (mg L^{-1})*			
	0	*5*	*10*	*15*
Caruru	0.0±0.0 a	10.0±5.7 b	10.6±5.8 b	24.3±8.1 b
Corda de viola	0.0±0.0 a	31.2±8.8 a	50.6±9.5 a	70.1±8.7 a
Amendoim	0.0±0.0 a	37.5±9.2 a	49.4± 9.5 a	63.2±9.1 a
	20	*25*	*30*	
Caruru	43.1±9.4 c	71.9±8.5 b	99.4±1.5 a	
Corda de viola	99.4±1.5 a	98.1±2.5 a	99.4±1.5 a	
Amendoim	75.1±8.2 b	99.4±1.5 a	100.0±0.0 a	

Médias, acompanhadas pelos intervalos de confiança ajustados pelo método de Šidák e seguidas por letras distintas na coluna diferem-se entre si pelo teste de Tukey a 0.05 de significância. Modelo ajustado: MLG com distribuição binomial e função de ligação logit.

Para o ajuste da regressão em cada espécie, deve-se separar os dados. O código abaixo separa os dados para amendoim-bravo (linhas 23 a 25), caruru-roxo (linhas 26 a 28) e corda de viola (linhas 29 a 31).

```
21    DOSE<-D[1:28]
22    dose<-Dose[1:28]

23    morta1<-morta[1:28]
24    viva1<-viva[1:28]
25    resp1<-cbind(morta1, viva1)

26    morta2<-morta[29:56]
27    viva2<-viva[29:56]
28    resp2<-cbind(morta2, viva2)

29    morta3<-morta[57:84]
30    viva3<-viva[57:84]
31    resp3<-cbind(morta3, viva3)
```

Aqui, especifica-se o intervalo relativo à forma como as repetições das dosagens aparecem nos dados (coluna `Dose`, valores de 1 a 28; 29 a 56; 57 a 84). Como há sete doses testadas em quatro blocos, cada espécie possui 28 parcelas no total, em que se dividem os dados de 1 a 28, 29 a 56 e 57 a 84 – tal classificação foi possível porque o fator qualitativo `Especie` foi fixado na ordenação dos fatores. Em seguida, ajustaram-se novos MLGs para cada espécie, com vistas a obter os coeficientes da regressão. A distribuição binomial e a função de ligação logit são mantidas.

```
32    bin_amendoim<-glm(resp1~dose,family=binomial, data=dados)
33    summary(bin_amendoim)

34    bin_caruru<-glm(resp2~dose,family=binomial, data=dados)
35    summary(bin_caruru)

36    bin_corda<-glm(resp3~dose,family=binomial, data=dados)
37    summary(bin_corda)
```

Diante da saída dos comandos, obtêm-se os valores dos coeficientes da equação logística para cada espécie, e as equações são ajustadas (Tabela 11).

```
> summary(bin_amendoim)
Call:
glm(formula = resp1 ~ dose, family = binomial, data = dados)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.849  -1.234   0.511   1.333   2.774

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.23709    0.15951  -14.03   <2e-16 ***
dose         0.20252    0.01149   17.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 671.100  on 27  degrees of freedom
Residual deviance:  91.914  on 26  degrees of freedom
AIC: 162.58
```

```
> summary(bin_caruru)
Call:
glm(formula = resp2 ~ dose, family = binomial, data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9321  -0.9995  -0.3100   1.1354   2.8974

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37674    0.25471  -17.18   <2e-16 ***
dose         0.22163    0.01267   17.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 677.026  on 27  degrees of freedom
Residual deviance:  57.963  on 26  degrees of freedom
AIC: 136.63
```

```
> summary(bin_corda)
Call:
glm(formula = resp3 ~ dose, family = binomial, data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-2.5534  -1.3979   0.1249   1.1248   2.2128

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.74170    0.19182  -14.29   <2e-16 ***
dose         0.27522    0.01605   17.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 805.114  on 27  degrees of freedom
Residual deviance:  65.748  on 26  degrees of freedom
AIC: 126.7
```

Convém salientar que a regressão logística estima os parâmetros via máxima verossimilhança (MLE), diferentemente da regressão linear, que o faz por meio de quadrados mínimos ordinários (OLS). Por esse motivo, o coeficiente de determinação não pode ser calculado (R^2).

TABELA 11. Regressões logísticas para o controle das plantas infestantes amendoim-bravo, caruru-roxo e corda de viola, sendo y a probabilidade de controle e x a dosagem do produto em mg L^{-1}.

Amendoim-bravo	$y = \frac{1}{1 + e^{-(-2.2371+0.2025x)}}$
Caruru-roxo	$y = \frac{1}{1 + e^{-(-4.3767+0.2216x)}}$
Corda de viola	$y = \frac{1}{1 + e^{-(-2.7417+0.2752x)}}$

Há na literatura outras abordagens para calcular o chamado pseudo-R^2, que consegue representar também a qualidade do ajuste da equação. Todavia, para os modelos quasi, em que não se estima a máxima verossimilhança, há dificuldades para mensurar um pseudo-R^2 adequado. Portanto, optou-se pela redução do *deviance* (D^2), também chamado de *deviance* explicado, que permite inferir sobre a qualidade do ajuste da equação pela fórmula:

$$D^2 = \frac{(D_{nulo} - D_{res})}{D_{nulo}} \quad (14)$$

em que D_{nulo} é o *deviance* do modelo nulo.

Para as equações ajustadas, os valores de D^2 foram calculados (Tabela 12).

TABELA 12. Redução do *deviance* (%) para equações ajustadas de amendoim-bravo, caruru-roxo e corda de viola.

Amendoim-bravo	$D^2 = \frac{(671.1 - 91.914)}{671.1} = 0.8630 = 86.30\%$
Caruru-roxo	$D^2 = \frac{(677.026 - 57.963)}{677.026} = 0.9144 = 91.44\%$
Corda de viola	$D^2 = \frac{(805.114 - 65.748)}{805.114} = 0.9183 = 91.83\%$

O próximo código plota os modelos de regressão ora ajustados. É necessário processar novamente o modelo binomial ao substituir `dose` por `DOSE` (assim, o sistema entende que dose é qualitativo), para extrair as médias de cada dose pela função `emmeans`.

```
38    bin_amendoim1<-glm(resp1~DOSE,family=binomial, data=dados)
39    media1 <- emmeans(bin_amendoim1,~ DOSE)
40    medfin1<-regrid(media1)
41    medfin1

42    bin_caruru1<-glm(resp2~DOSE,family=binomial, data=dados)
43    media2 <- emmeans(bin_caruru1,~ DOSE)
44    medfin2<-regrid(media2)
45    medfin2

46    bin_corda1<-glm(resp3~DOSE,family=binomial, data=dados)
47    media3 <- emmeans(bin_corda1,~ DOSE)
48    medfin3<-regrid(media3)
49    medfin3
```

Com as médias de cada espécie, é possível plotar os gráficos. As próximas funções plotam os gráficos de cada espécie (Figura 3) e podem ser alteradas para formatar tamanho, letra e forma do gráfico.

```
50    PLOT1=c(0,0.375,0.49375,0.63125,0.75,0.99375,1)
51    DOSE=c(0,5,10,15,20,25,30)
52    plot(DOSE,PLOT1,xlab="Dose(mg/L)",ylab="Probabilidade de
      controle")
53    curve(predict(bin_amendoim,data.frame(dose=x),type="resp"),
      add=TRUE)
54    points(dose,fitted(bin_amendoim),pch=20)

55    PLOT2=c(0,0.10,0.10625, 0.24375,0.43125,0.71875,0.99375)
```

```
56    plot(DOSE,PLOT2,xlab="Dose(mg/L)",ylab="Probabilidade de
      controle")
57    curve(predict(bin_caruru,data.frame(dose=x),type="resp"),
      add=TRUE)
58    points(dose,fitted(bin_caruru),pch=20)

59    PLOT3=c(0,0.3125,0.50625,0.7,0.99375,0.98125,0.99375)
60    plot(DOSE,PLOT3,xlab="Dose(mg/L)",ylab="Probabilidade de
      controle")
61    curve(predict(bin_corda,data.frame(dose=x),type="resp"),
      add=TRUE)
62    points(dose,fitted(bin_corda),pch=20)
```

Ao comparar as equações geradas, observou-se que os termos acompanhantes de *x* são parecidos, sendo que a corda de viola apresentou o menor valor (-0.2752); logo, quando se aumenta a dose do produto, a espécie é controlada de forma mais intensa que nas demais. Além disso, ao avaliar os termos independentes, caruru-roxo apresentou o maior valor (4.3767), o que indica maior dificuldade de controle, se comparado às demais espécies.

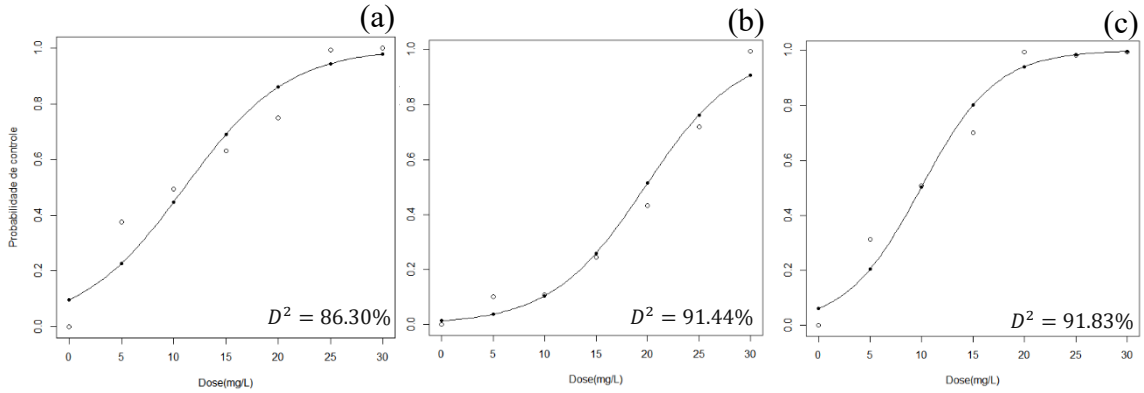


FIGURA 3. Regressões logísticas para a probabilidade de controle das plantas infestantes amendoim-bravo (a), caruru-roxo (b) e corda de viola (c), sendo *y* a probabilidade de controle e *x* a dosagem do produto em mg L^{-1}. Os círculos vazios indicam os valores observados.

2.9 Modelos Zero Inflacionados e Zero Truncados

É comum uma quantidade elevada de contagens nulas, ou seja, valores zero em avaliações de dados de contagem, principalmente no estudo do comportamento no ambiente. O excesso de zeros em um experimento é chamado de zeros inflacionados que,

ao serem ignorados, pode fazer com que os parâmetros estimados e os erros padrão se tornem tendenciosos, ao passo que o excesso de zeros pode levar à sobredispersão.

Zeros podem ser originários de diversas causas previamente relatadas na literatura (KUHNERT et al., 2005; MARTIN et al., 2005). Os que se originam do erro estrutural são obtidos quando não se encontrou o objeto de estudo, pois o hábitat não foi favorável. Os zeros que se originam do delineamento do experimento ocorrem se a época de coleta for incorreta, a área amostrada for pequena ou um tratamento próximo tenha afetado os resultados. Zeros originários do observador também surgem quando o avaliador não possui experiência em amostragem ou não enxerga o indivíduo contabilizado. Por fim, há o zero proveniente do próprio indivíduo, em que, apesar de o hábitat ser propício, ele não é encontrado.

Os zeros originários do delineamento ou observador são chamados de "zeros falsos", pois são indesejáveis na amostragem – zeros estruturais são chamados de "zeros positivos". Há duas possibilidades para que os zeros sejam modelados: como zeros inflacionados ou zeros truncados. Tais modelos corrigem variáveis de respostas que apresentam quantidades de zeros maiores que a esperada na distribuição – a diferença entre os dois modelos está na forma como os zeros são modelados.

Para o modelo zero truncado, também chamado de modelo Hurdle, a variável resposta não pode assumir valores nulos. Para o modelo de zeros inflacionados, a presença de zeros não é necessariamente um problema – ele é mais comum para dados agronômicos do que os modelos de zeros truncados. De fato, a variável resposta normalmente pode assumir valores nulos por conta dos efeitos do próprio tratamento. No modelo de zeros truncados, a variável resposta não pode assumir valor nulo, posto que os zeros presentes na amostragem são considerados falsos – cria-se uma *hurdle* ("barreira", em inglês) para os zeros presentes no banco de dados. A Figura 4 ilustra a possível origem dos zeros para o exemplo do controle biológico do pulgão-do-algodoeiro. A escolha entre os modelos de zeros truncados ou zeros inflacionados será em virtude da classificação dos zeros do banco de dados.

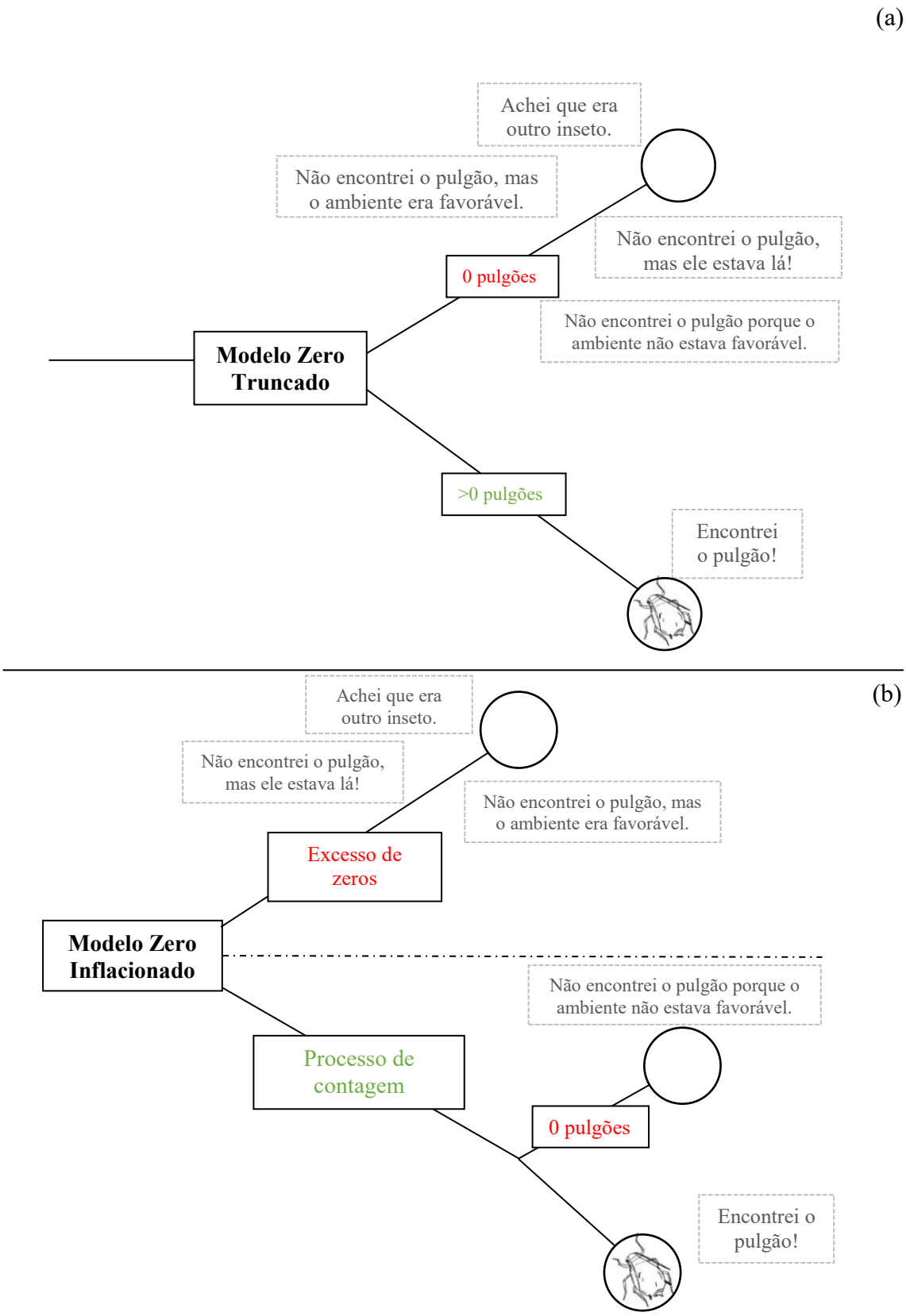


FIGURA 4. Modelagem do excesso de zeros para exemplo do controle biológico do pulgão. No modelo zero truncado (a) há uma barreira que modela separadamente os zeros encontrados (independente da origem deste zero). No modelo zero inflacionado (b) o modelo separa apenas os falsos zeros do modelo, sendo que os zeros positivos são inseridos no modelo do processo de contagem.

2.9.1 Modelos zero truncados (Modelos Hurdle)

Esses modelos foram originalmente propostos por Mullahy (1986) na área de econometria. Há dois componentes na modelagem Hurdle: os dados são considerados zeros *versus* não zeros, e um modelo de distribuição binomial modela a probabilidade, em que valores zero são observados; ou as observações não nulas são modeladas via distribuição de Poisson (ZTP) ou binomial negativa (ZTBN) truncada – como as distribuições são zero truncadas, elas não podem produzir zeros, e tal modelo não discrimina as diferenças entre os quatro tipos de zeros. As funções de probabilidade para as distribuições de Poisson e binomial negativa para o modelo excluem a probabilidade de obter valores nulos. Assim, os modelos Hurdle combinam um modelo de contagem ($f_{Poisson}$) e um modelo para os zeros ($f_{binomial}$):

$$f_{ZTP}(y;\ \beta,\gamma) = \begin{cases} f_{binomial}(y=0;\ \gamma) & \text{se y=0} \\ (1 - f_{binomial}(y=0;\gamma)) \times \dfrac{f_{Poisson}(y;\beta)}{1 - f_{Poisson}(y=0;\beta)} & \text{se y>0} \end{cases} \quad (15)$$

onde γ é o vetor dos parâmetros estimados ao modelo de zeros.

Com esse modelo, a especificação da verossimilhança pode ser maximizada de maneira separada. Modelos zero truncados serão descritos em razão da raridade de dados que se comportam em relação ao modelo nas Ciências Agrárias. No R, a modelagem Hurdle pode ser ajustada com a função `hurdle()` do pacote `pscl`, e ambos os modelos são estimados e calculados com a função dada por:

```
hurdle(formula, data=, ... , dist = "poisson", zero.dist = "binomial",
link = "logit", ...)
```

onde: `formula` é um objeto que anuncia a estrutura do modelo; `data` é o banco de dados; `dist` é a especificação da família (`poisson` ou `negbin`); `zero.dist` é a especificação da família para a modelagem dos zeros; `link` é a função de ligação a ser utilizada na modelagem dos zeros.

2.9.2 Modelos zero inflacionados (ZIP e ZIBN)

Modelos zeros inflacionados (MULLAHY, 1986; LAMBERT, 1992) também são capazes de lidar com excessos de zero da contagem. Com essa abordagem, os zeros advêm de dois processos: binomial e de contagem. Assim como nos modelos Hurdle, um MLG binomial é modela a probabilidade de medida de zeros; depois, o processo de contagem é modelado via Poisson (ZIP) e binomial negativa (ZIBN). A diferença fundamental com essa abordagem é que os processos de contagem podem produzir zeros (a distribuição não é zero truncada). Seja $P(Y_i)$ a probabilidade para ocorrer um evento *i*, então:

$$P(Y_i = 0) = P(falsos\ zeros) + [(1 - P(falsos\ zeros)) \times P(proceso\ de\ contagem\ dar\ zero)] \quad (16)$$

Assume-se que a probabilidade de Y_i ser um falso zero segue a distribuição binomial com probabilidade π_i. Assim, a probabilidade para que Y_i não seja um falso zero é igual a $1 - \pi_i$; logo, reescreve-se a Equação 16 como:

$$P(Y_i = 0) = \pi_i + [1 - \pi_i \times P(processo\ de\ contagem\ dar\ zero)] \quad (17)$$

sendo π_i a probabilidade de falsos zeros.

A probabilidade de o processo de contagem ser zero pode ser obtida ao assumir a distribuição de Poisson ou binomial negativa. Se o processo de contagem Y_i segue a distribuição a Poisson, a função de probabilidade é definida por:

$$\begin{cases} f(y_i = 0) = \pi_i + (1 - \pi_i) \times e^{-\mu_i} \\ f(y_i | y_i > 0) = (1 - \pi_i) \times \dfrac{\mu^{y_i} \times e^{-\mu_i}}{y_i!} \end{cases} \quad (18)$$

Há uma pequena modificação no cálculo da média e variância do modelo. No modelo ZIP, a esperança é dada por $E(Y_i) = \mu_i \times (1 - \pi_i)$, e a variância, como $var(Y_i) = (1 - \pi_i) \times (\mu_i + \pi_i \times \mu_i^2)$. Quando os dados sofrem sobredispersão, podem ser ajustados à distribuição binomial negativa, mas não à quasi-Poisson, pois os modelos de zeros inflacionados não permitem o uso das famílias quasi. Para a distribuição binomial negativa, as funções de probabilidade são definidas por:

$$\begin{cases} f(y_i = 0) = \pi_i + (1 - \pi_i) \times \left(\dfrac{k}{\mu_i + k}\right)^k \\ f(y_i | y_i > 0) = (1 - \pi_i) \times f_{BN}(y) \end{cases} \quad (19)$$

No modelo ZIBN, a esperança é dada por $E(Y_i) = \mu_i \times (1 - \pi_i)$, e a variância, como $var(Y_i) = (1 - \pi_i) \times \left(\mu_i + \frac{\mu_i^2}{k}\right) + \mu_i^2 \times (\pi_i^2 + \pi_i)$ – modelos ZIP e ZIBN podem ser modelados pelo pacote `pscl`. A função `zeroinfl` aplica um modelo de zeros inflacionados na forma:

```
zeroinfl(formula, data =, dist = "poisson",link = "logit",..., control
= zeroinfl.control(),...)
```

em que: `formula` é um objeto que anuncia a estrutura do modelo; `data` é o banco de dados; `dist` é a especificação da família (`poisson`, `negbin` ou `geometric`); `link` é a função de ligação a ser utilizada na modelagem dos zeros em um modelo binomial (`logit`, `probit`, `cloglog`, `cauchit` ou `log`); `control=zeroinfl.control(...)` é a lista com argumentos específicos para modelagem de zeros.

O parâmetro `formula` para o modelo de zeros inflacionados terá dupla entrada, sendo a primeira parte para descrever o modelo em si, e a segunda, para descrever o modelo para ocorrência de zeros. A descrição dos parâmetros dos dois modelos é separada por uma barra (|), e a primeira parte do modelo segue os mesmos princípios utilizados no parâmetro `formula` de outras funções. Nesse caso, a modelagem dos zeros pode ser especificada de três formas, em que a escolha dos parâmetros a serem utilizados é uma decisão puramente biológica sobre o que pode afetar a presença de zeros no banco de dados, e não estatística.

Na primeira modelagem, ao considerar um experimento com dois fatores X_1 e X_2 e a variável Y, a função `formula` é especificada em $Y \sim X_1 + X_2 | 1$. Nessa modelagem, a probabilidade de obter um falso zero está em função apenas do intercepto. A probabilidade é calculada por regressão logística como:

$$\pi_i = \frac{e^{\nu}}{1 + e^{\nu}} \tag{20}$$

em que ν é o intercepto.

Desse modo, a ocorrência dos falsos zeros é justificada apenas pelo ambiente, e não pelos fatores de estudo. A estimativa de π_i pode ainda incluir covariáveis: se os fatores de estudo podem provocar a presença de falsos zeros, na segunda modelagem, a fórmula é especificada em $Y \sim X_1 + X_2 | X_1 + X_2$, e a probabilidade de obter um falso zero é dada por:

$$\pi_i = \frac{e^{\nu+\gamma_1 X_1 \cdots \gamma_n X_n}}{1 + e^{\nu+\gamma_1 X_1 \cdots \gamma_n X_n}} \quad (21)$$

sendo γ o vetor dos coeficientes de regressão estimados para as variáveis do modelo binomial.

Caso existam covariáveis que provoquem a presença de falsos zeros (por exemplo, Z_1 e Z_2), na terceira modelagem, a fórmula é especificada em $Y \sim X_1 + X_2 | Z_1 + Z_2$, e a probabilidade de obter um falso zero é dada por:

$$\pi_i = \frac{e^{\nu+\gamma_1 Z_1 \cdots \gamma_n Z_n}}{1 + e^{\nu+\gamma_1 Z_1 \cdots \gamma_n Z_n}} \quad (22)$$

Nesse modelo, o processo de contagem é modelado por outras variáveis diferentes daquelas de estudo. Os resultados das funções `zeroinfl` e `hurdle` são visualizadas com a função `summary()`; a função `Anova()` do pacote `car` apresenta a significância dos fatores; e a função `AIC` (do pacote `stats`) indica o AIC do modelo.

2.9.2.1 Controle biológico do pulgão-do-algodoeiro

Por fim, os dados de contagem do pulgão-do-algodoeiro foram ajustados a um modelo zeros inflacionados, pois se observaram zeros provenientes do efeito dos tratamentos (zeros verdadeiros, resultado do controle do inseto-praga), além da presença de falsos zeros (repetições com valores nulos nas quais isso não era esperado) (Figura 4b). Os dados seguirão a distribuição binomial negativa, uma vez que foi constatada a sobredispersão. Como a presença de falsos zeros é função exclusiva do ambiente, não serão ajustadas covariáveis para o modelo binomial.

```
1     dados<-read.table("exemplo2.txt", header=T)
2     attach(dados)
3     str(dados)
4     library(car)
5     library(pscl)
6     zeroinfl<- zeroinfl(cont ~ trat + bloco|1, data = dados, dist =
      "negbin",link = "logit")
7     summary(zeroinfl)
8     Anova(zeroinfl, test="Chi")
9     AIC(zeroinfl)
```

```
10    media <- emmeans(zeroinfl, ~trat)
11    medfin<-regrid(media)
12    cld(medfin,alpha=0.05, Letters=letters, adjust="tukey")
```

Por meio da função `summary`, apresentam-se os coeficientes dos dois modelos criados. Nesse caso é modelada a probabilidade de falsos zeros do modelo, dado que o intercepto foi significativo.

```
Call:
zeroinfl(formula = cont ~ trat + bloco | 1, data = dados, dist = "negbin",
    link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.1084 -0.7446 -0.2213  0.3669  2.7574

Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.89283    0.50802   7.663 1.82e-14 ***
tratB       -2.26295    0.43545  -5.197 2.03e-07 ***
tratC       -2.22822    0.47706  -4.671 3.00e-06 ***
tratcont     2.57612    0.39868   6.462 1.04e-10 ***
tratD        0.46658    0.40094   1.164  0.24454
blocoII     -0.34209    0.53621  -0.638  0.52348
blocoIII    -1.15204    0.62583  -1.841  0.06565 .
blocoIV      0.25046    0.49462   0.506  0.61260
blocoIX     -1.42310    0.77243  -1.842  0.06542 .
blocoV       0.52634    0.52376   1.005  0.31494
blocoVI     -0.07424    0.56404  -0.132  0.89529
blocoVII     0.38837    0.52951   0.733  0.46328
blocoVIII    0.59446    0.54062   1.100  0.27151
blocoX       0.33675    0.51422   0.655  0.51255
Log(theta)   0.82583    0.27222   3.034  0.00242 **

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3578     0.3984  -3.408 0.000655 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 2.2838
Number of iterations in BFGS optimization: 24
Log-likelihood: -202.2 on 16 Df
```

Diante do modelo binomial com função logit, a probabilidade de obter um falso zero foi dada como:

$$\pi_i = \frac{e^{\nu}}{1+e^{\nu}} = \frac{e^{-1.3578}}{1+e^{-1.3578}} = \frac{0.257226}{1.257226} = 20.46\%$$

A probabilidade de obter um falso zero foi de 20.46%, em que o restante dos zeros foi proveniente dos tratamentos utilizados e foi inserido no modelo de contagem. A função `Anova` identifica diferenças significativas entre os tratamentos.

```
> Anova(zeroinfl, test="Chi")
Analysis of Deviance Table (Type II tests)
```

```
Response: cont
      Df   Chisq Pr(>Chisq)
trat   4 206.541     < 2e-16 ***
bloco  9  19.165     0.02383 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> AIC(zeroinfl)
[1] 436.336


> cld(medfin,alpha=0.05, Letters=letters, adjust="tukey")
 trat     emmean         SE  df   asymp.LCL  asymp.UCL .group
 B      4.408274   1.374663 Inf   0.8770889   7.939458  a
 C      4.564043   1.550815 Inf   0.5803668   8.547719  a
 A     42.369622  13.097535 Inf   8.7251561  76.014088   b
 D     67.560007  20.232517 Inf  15.5874634 119.532550   b
 cont 556.993030 141.129446 Inf 194.4649069 919.521153    c

Results are averaged over the levels of: bloco
Confidence level used: 0.95
Conf-level adjustment: sidak method for 5 estimates
P value adjustment: tukey method for comparing a family of 5 estimates
significance level used: alpha = 0.05
```

Como houve diferenças significativas para os tratamentos, as médias foram comparadas (linhas 10 a 12). Com esse modelo, consegue-se controlar tanto a sobredispersão dos dados como o excesso de zeros para a contagem de pulgões-do-algodoeiro, trazendo um modelo correto e com confiabilidade nos parâmetros estimados, pois as taxas de erro foram reduzidas. Diante dos testes de médias (Tabela 13), observou-se que, para o modelo Poisson sem ajuste ao excesso de zeros e com sobredispersão, tanto as médias quanto os intervalos de confiança foram subestimados, em comparação ao modelo ZIBN. Com os intervalos de confiança alterados, o modelo ZIBN não detectou diferenças entre os tratamentos A e D.

Para a análise gráfica dos resíduos, os resíduos de Pearson foram calculados para cada modelo. No modelo de zeros inflacionados, o resíduo de Pearson é obtido por:

$$\hat{\varepsilon}_i^P = \frac{y_i - (1 - \hat{\pi}_i) \times \hat{\mu}_i}{\sqrt{var(Y_i)}} \tag{23}$$

Como há um grande intervalo na estimativa do número de pulgões (algumas observações do modelo Poisson estimam mais que dois mil pulgões), também foi criado um gráfico na escala logarítmica dos valores preditos para verificar a dispersão dos resíduos.

TABELA 13. Número médio de pulgões-do-algodoeiro (*Aphis gossypii*) por planta após a aplicação de produtos em quatro modelos: MLG com distribuição de Poisson, MLG com distribuição quasi-Poisson, MLG com distribuição binomial negativa (BN) e MLG com distribuição binomial negativa e zero inflacionado (ZIBN).

	Modelo			
	Poisson	**Quasi-Poisson**	**BN**	**ZIBN**
Produto	*Pulgões por planta*			
A	23.36±3.28 b	23.36±25.70 a	43.65±46.05 a	42.37±33.64 b
B	2.47±0.94 a	2.47±7.38 a	3.29±3.84 a	4.41±3.53 a
C	2.37±0.92 a	2.37±7.22 a	3.03±3.56 a	4.56±3.98 a
D	32.72±4.07 c	32.72±31.89 a	58.12±61.13 a	67.56±51.97 b
Controle	380.27±28.88 d	380.27±226.33 b	406.57±424.11 a	556.99±362.53 c

Todos os modelos foram ajustados com a função de ligação logarítmica. Médias, acompanhadas pelos intervalos de confiança ajustados pelo método de Šidák e seguidas por letras distintas na coluna, diferem-se entre si pelo teste de Tukey a 0.05 de significância.

Os seguintes comandos são realizados para a plotagem dos resíduos de Pearson:

```
13  Resíduos<- resid(poi, type = "pearson")
14  Valores_preditos<-predict(poi, type = "response")
15  plot(x = Valores_preditos, y = Resíduos,xlim=c(0,2500),main =
    "Modelo Poisson", ylim=c(-15,20))
16  plot(x = log(Valores_preditos), y = Resíduos,main = "Modelo
    Poisson",xlim=c(-3,9), ylim=c(-15,20))

17  res<-resid(quasipoi, type = "response")
18  Valores_preditos <- predict(quasipoi, type = "response")
19  Resíduos <- res / sqrt(61.41471 * Valores_preditos)
20  plot(x = Valores_preditos, y = Resíduos, main = "Modelo
    QP",xlim=c(0,2500), ylim=c(-15,20))
21  plot(x = log(Valores_preditos), y = Resíduos,main = "Modelo
    QP",xlim=c(-3,9), ylim=c(-15,20))

22  Resíduos<- resid(bn, type = "pearson")
23  Valores_preditos<-predict(bn, type = "response")
24  plot(x = Valores_preditos, y = Resíduos,xlim=c(0,1000),main =
    "Modelo BN", ylim=c(-15,20))
25  plot(x = log(Valores_preditos), y = Resíduos,main = "Modelo
    BN",xlim=c(-3,9), ylim=c(-15,20))

26  Resíduos<- resid(zeroinfl, type = "pearson")
27  Valores_preditos<-predict(zeroinfl, type = "response")
28  plot(x = Valores_preditos, y = Resíduos,xlim=c(0,1000),main =
    "Modelo ZIBN", ylim=c(-15,20))
```

```
29    plot(x = log(Valores_preditos), y = Resíduos,main = "Modelo ZIBN
      ",xlim=c(-3,9), ylim=c(-15,20))
```

Nota-se que os modelos QP, BN e ZIBN foram capazes de reduzir a alta variação dos resíduos presente no modelo Poisson (Figura 5). Porém, houve grande concentração de resíduos nulos em todos os modelos – sabe-se que os valores nulos só foram ajustados no modelo ZIBN.

Outra medida comumente utilizada para comparar modelos é o Critério de Informação de Akaike (AIC) (AKAIKE, 1973), que representa a ausência de generalidade do modelo e penaliza tanto a falta de ajuste aos dados quanto a alta complexidade do modelo. Aqui há a preferência por menores valores de AIC, definido por:

$$AIC = -2logL + 2p \tag{24}$$

em que: p: número de parâmetros do modelo; L: valor da razão de máxima verossimilhança do modelo.

Entretanto, o AIC deve ser utilizado com cautela, pois não pode ser comparado a modelos com distribuições diferentes (como Poisson *versus* NB). Nesses termos, a mensuração da verossimilhança é diferente para cada distribuição, e isso também ocorre com as dimensões. O AIC pode comparar modelos aninhados, a exemplo dos zeros inflacionados (Poisson *versus* ZIP ou BN *versus* ZIBN) ou, ainda, comparar modelos com diferentes correções do excesso de zeros, fatores dentro do modelo ou diferentes funções de ligação.

Ademais, o AIC comparou os modelos de zeros inflacionados ajustados para a contagem do pulgão-do-algodoeiro. Concluiu-se que, apesar da inserção de novos parâmetros ao modelo de zeros inflacionados, houve redução na verossimilhança destes, mostrando que os modelos ZIP e ZIBN foram mais parcimoniosos que os modelos originais.

Reduziu-se o AIC de 1895.90 para 1165.53 com o modelo Poisson Zero Inflacionado (ZIP) (saídas não apresentadas para o modelo ZIP), mas ambos não corrigem a sobredispersão. Também foi diminuído o AIC de 456.24 para 436.34 com o modelo ZIBN, ratificando que o modelo ZIBN se tornou mais parcimonioso para o ajuste dos dados.

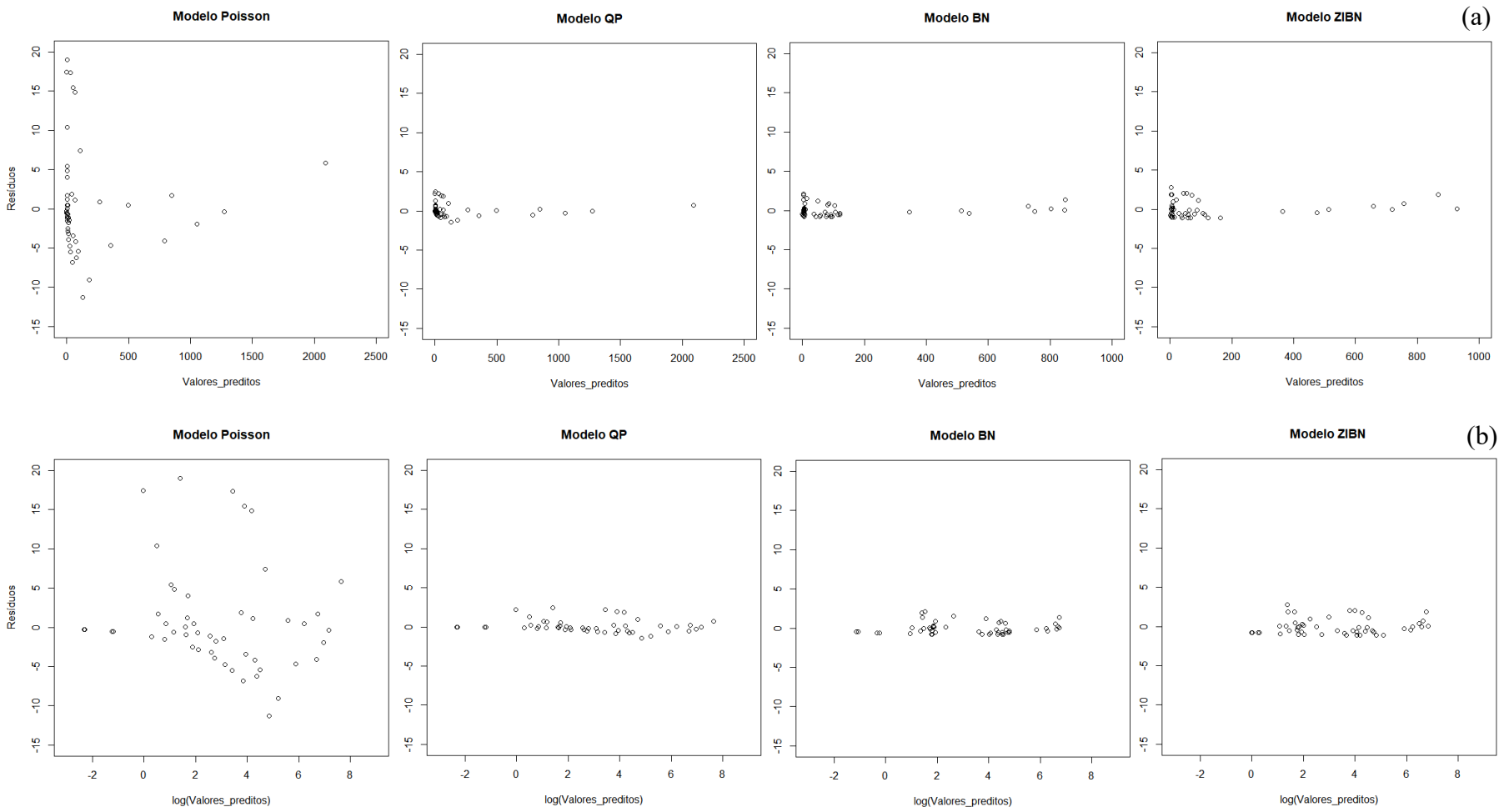


FIGURA 5. Resíduos de Pearson plotados *versus* os valores preditos (a) e o logaritmo dos valores preditos (b) para os modelos ajustados para a contagem de pulgão-do-algodoeiro. QP: quasi-Poisson; BN: binomial negativa; ZIBN: zero inflacionado binomial negativa.

3) CONSIDERAÇÕES FINAIS

Os MLGs oferecem qualidade de ajuste para dados binomiais, de contagem e com excesso de zeros, dado comumente encontrado nas Ciências Agrárias. Diante disso, o presente capítulo aproximou a teoria dos MLGs e a respectiva aplicação no ambiente computacional R de uma forma didática e prática, com técnicas mais adequadas para a análise de tais informações, devido à falta de materiais com esse objetivo. Com exemplos e saídas dos modelos, espera-se que o presente capítulo auxilie na análise e interpretação de experimentos agronômicos.

4) REFERÊNCIAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. *In*: PETROV, B. N.; CSAKI, F. (ed.). **Second International Symposium on Information Theory.** Budapest: Akademai Kiado, 1973. p. 267–281.

BLISS, C. I.; FISHER, R. A. Fitting the negative binomial distribution to biological data and a note on the efficient fitting of the negative binomial. **Biometrics**, [*s. l.*], v. 9, p. 176–200, 1953. DOI: https://doi.org/10.2307/3001850.

CALAMA, R.; MUTKE, S.; TOMÉ, J.; GORDO, J.; MONTERO, G.; TOMÉ, M. Modelling spatial and temporal variability in a zero-inflated variable: the case of stone pine (*Pinus pinea* L.) cone production. **Ecological Modelling**, [*s. l.*], v. 222, p. 606–618, 2011. DOI: https://doi.org/10.1016/j.ecolmodel.2010.09.020.

CAMERON, A. C.; TRIVEDI, P. K. Regression-based for overdispersion in the Poisson model. **Journal of Econometrics**, [*s. l.*], v. 46, p. 347–364, 1990. DOI: https://doi.org/10.1016/0304-4076(90)90014-K.

CHAMBERS J. M., HASTIE T. J. **Statistical Models in S**. Londres: Chapman and Hall, 1992. 624 p.

CHANDRA, N. K; ROY, D.; GHOSH, T. A Generalized Poisson Distribution. **Communications in Statistics - Theory and Methods**, [*s. l.*], v. 42, p. 2786–2797, 2013. DOI: https://doi.org/10.1080/03610926.2011.620207.

CRAWLEY, M. J. **The R book**. Inglaterra: Wiley, 2007. 942 p. DOI: https://doi.org/10.1002/9780470515075.

COHEN, A. C. An extension of a truncated Poisson distribution. **Biometrics**, [*s. l.*], v. 16, p. 446–450, 1960. DOI: https://doi.org/10.2307/2527694.

COHEN, A. C. Estimating the Poisson parameter from samples that are truncated on the right. **Technometrics**, [*s. l.*], v. 3, p. 433–438, 1961. DOI: https://doi.org/10.1080/00401706.1961.10489961.

CONSUL, P. C.; JAIN, G. C. A Generalization of the Poisson distribution. **Technometrics**, [*s. l.*], v. 15, p. 791–799, 1973. DOI: https://doi.org/10.1080/00401706.1973.10489112.

CONSUL, P. C. A simple urn model dependent upon predetermined strategy. **Sankhyā**, [*s. l.*], v. 36, p. 391–399, 1974.

CONSUL, P. C. **Generalized Poisson Distributions**: properties and applications. Nova Iorque: Marcel Dekker, 1989. 302 p.

CONSUL, P. C. On some properties and applications of quasi-binomial distribution. **Communications in Statistics - Theory and Methods**, [*s. l.*], v. 19, n. 2, p. 477–504, 1990. DOI: https://doi.org/10.1080/03610929008830214.

DENDWOOD, M. J.; STEAR, M. J.; MATTHEWS, L.; REID, S. W. J.; TOFT, N.; INNOCENT, G. T. The distribution of the pathogenic nematode *Nematodirus battus* in lambs is zero-inflated. **Parasitology**, Reino Unido, v. 135, p. 1225–1235, 2008. DOI: https://doi.org/10.1017/S0031182008004708.

DOBSON, A. J.; BARNETT, A. G. **An introduction to Generalized Linear Models**. 3. ed. Nova Iorque: Chapman & Hall, 2008. 307 p.

FISHER, R. A. The negative binomial distribution. **Annals of human genetics**, [*s. l.*], v. 11, n. 1, p. 182–187, 1941. DOI: https://doi.org/10.1111/j.1469-1809.1941.tb02284.x.

FOX, J; WEISBERG, S. **An R Companion to Applied Regression**. 2. ed. Thousand Oaks: Sage. 2011. 474 p.

GONZATTO JÚNIOR, O. A.; GUEDES, T. A.; GONÇALVES-ZULIANI, A. M. O.; NUNES, W. M. Zero-inflated beta regression model for leaf citrus canker incidence in orange genotypes grafted onto different rootstocks. **Acta Scientiarum**, [*s. l.*], v. 39, n. 2, p. 161–171, 2017. DOI: https://doi.org/10.4025/actascibiolsci.v39i2.33063.

HASS, S. E.; HOOTEN, M. B.; RIZZO, D. M.; MEENTEMEYER, R. K. Forest species diversity reduces disease risk in a generalist plant pathogen invasion. **Ecology Letters**, [*s. l.*], v. 14, p. 1108–1116, 2011. DOI: https://doi.org/10.1111/j.1461-0248.2011.01679.x.

HILBE, J. M. **Negative binomial regression**. 2. ed. New York: Cambridge University Press, 2011. 553 p. DOI: https://doi.org/10.1017/CBO9780511973420.

HOEF, J. M. V.; BOVENG, P. Quasi-Poisson vs. Negative Binomial regression: How should we model overdispersed count data? **Ecology**, [*s. l.*], v. 88, n. 11, p. 2766–2772, 2007. DOI: https://doi.org/10.1890/07-0043.1.

KERSTING, U.; SATAR, S.; UYGUN, N. Effect of temperature on development rate and fecundity ofapterous *Aphis gossypii* Glover (Hom., Aphididae) reared on *Gossypium hirsutum*. **Journal of Applied Entomology**, Berlin, v. 123, p. 23–27, 1999. DOI: https://doi.org/10.1046/j.1439-0418.1999.00309.x.

KLEIBER C.; ZEILEIS, A. **Applied Econometrics with R**. Nova Iorque: Springer-Verlag, 2008. 221 p. DOI: https://doi.org/10.1007/978-0-387-77318-6.

KUHNERT, P. M.; MARTIN, T. G.; MENGERSEN, K.; POSSINGHAM, H. P. Assessing the impacts of grazing levels on birds density in woodland habitats: a Bayesian approach using expert opinion. **Environmetrics**, [*s. l.*], v. 16, p.717–747, 2005. DOI: https://doi.org/10.1002/env.732.

JACKMAN, S. **pscl**: Classes and Methods for R Developed in the Political Science Computational Laboratory. United States Studies Centre, University of Sydney. Sydney, New South Wales, Australia, 2017. R package version 1.5.2. Disponível em: https://github.com/atahk/pscl/. Acesso em: 02 mar. 2018.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. **Generalized Linear Models with random effects**. Nova Iorque: Chapman & Hall, 2006. 380 p. DOI: https://doi.org/10.1201/9781420011340.

LENTH, R. **emmeans**: Estimated Marginal Means, aka Least-Squares Means. R package version 1.2.1, 2018. Disponível em: https://CRAN.R-project.org/package=emmeans. Acesso em: 02 mar. 2018.

MARTIN, T. G.; WINTLE, B. A.; RHODES, J. R; KUHNERT, P. M.; FIELD, S. A. LOW-CHOY, S. J.; TYRE, A. J.; POSSINGHAM, H. P. Zero tolerance ecology: improving ecological inference by modeling the source of zero observation. **Ecology Letters**, [*s. l.*], v. 8, p. 1235–1246, 2005. DOI: https://doi.org/10.1111/j.1461-0248.2005.00826.x.

MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**, 2. ed. Londres: Chapman & Hall, 1989. 493 p. DOI: https://doi.org/10.1007/978-1-4899-3242-6.

MILLER, R. G. **Beyond ANOVA**: basics of applied statistics. Florida**:** CRC Press, 1997. 336 p. DOI: https://doi.org/10.1201/b15236.

MULLAHY, J. Specification and Testing of Some Modified Count Data Models. **Journal of Econometrics**, [*s. l.*], v. 33, p. 341–365, 1986. DOI: https://doi.org/10.1016/0304-4076(86)90002-3.

LAMBERT, D. Zero-inflated Poisson regression with an application to defects in manufacturing. **Technometrics**, [*s. l.*], v. 34, p. 1–14, 1992. DOI: https://doi.org/10.2307/1269547.

MYERS, R. H.; MONTGOMERY, D. C.; VINING, G. G. **Generalized Linear Models, with applications in engineering and the sciences.** Nova Iorque: John Wiley and Sons Press, 2002. 342 p.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, [*s. l.*], v. 135, p. 370–384, 1972. DOI: https://doi.org/10.2307/2344614

POISSON, S. D. **Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés**. Paris: Bachelier, 1837. 415 p.

SAMPAIO, M. V.; KORNDÖRFER, A. P.; PUJADE-VILLAR, J.; HUBAIDE, J. E. A.; FERREIRA, S.E.; ARANTES, S. O.; BORTOLETTO, D. M.; GUIMARÃES, C. M.; SÁNCHEZ-ESPIGARES, J.A.; CABALLERO-LÓPEZ, B. Brassica aphid (Hemiptera: Aphididae) populations are conditioned by climatic variables and parasitism level: a study case of Triângulo Mineiro, Brazil. **Bulletin of Entomological Research**, Cambridge, v. 107, p. 410–418, 2017. DOI: https://doi.org/10.1017/S0007485317000220.

ŠIDÁK, Z. K. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. **Journal of the American Statistical Association**, [*s. l.*], v. 62, n. 318, p. 626–633, 1967. DOI: https://doi.org/10.1080/01621459.1967.10482935.

SILVA, M. F. Revisão taxonômica de gênero *Peltogyne* Vog. (Leguminosae-Caesalpinioideae). **Acta Amazônica**, [*s. l.*], v. 6, n. 1, p. 1–61, 1976. DOI: https://doi.org/10.1590/1809-43921976061s005.

GRAVES, S.; PIEPHO, H. P.; SELZER, L. DORAI-RAJ, S. **multcompView**: Visualizations of Paired Comparisons. R package version 0.1-7, 2015. Disponível em: https://CRAN.R-project.org/package=multcompView. Acesso em: 04 mar. 2018.

STROUP, W. W. **Generalized linear mixed models**: modern concepts, methods and applications. Florida: CRC Press, 2012. 555 p.

STROUP, W. W. Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. **Agronomy Journal**, Madison, v. 107, n. 2, p. 811–827, 2015. DOI: https://doi.org/10.2134/agronj2013.0342.

VENABLES, W. N.; RIPLEY, B. D. **Modern applied statistics with S**. 4. ed. Nova Iorque: Springer, 2002. 495 p. DOI: https://doi.org/10.1007/978-0-387-21706-2.

WEDDERBURN, R. W. M. Quasi-likelihood functions, Generalized Linear Models, and the Gauss-Newton method. **Biometrika**, [*s. l.*], v. 61, p.439–447, 1974. DOI: https://doi.org/10.2307/2334725.

YEŞILOVA, A.; KAYA, Y.; KAKI, B.; KASAP, I. Analysis of Plant Protection Studies with Excess Zeros Using Zero-Inflated and Negative Binomial Hurdle Models. **Gazi University Journal of Science**, [*s. l.*], v. 23, n. 2, p. 131–136, 2010.

ZUUR, A. F.; IENO, E. N.; WALKER, N. J.; SAVELIEV, A. A.; SMITH, G. M. **Mixed Effects Models and Extensions in Ecology with R.** Nova Iorque: Springer, 2009. 574 p. DOI: https://doi.org/10.1007/978-0-387-87458-6.

5) ANEXOS

ANEXO A. Dados da germinação de *Peltogyne confertiflora* em um experimento inteiramente casualizado com quatro lotes e 16 repetições (Exemplo 1).

lote	rep	germ	ngerm
L1	1	4	21
L1	2	9	16
L1	3	10	15
L1	4	8	17
L1	5	6	19
L1	6	8	17
L1	7	7	18
L1	8	8	17
L1	9	8	17
L1	10	10	15
L1	11	6	19
L1	12	4	21
L1	13	6	19
L1	14	8	17
L1	15	7	18
L1	16	12	13
L1	1	20	5
L2	2	20	5
L2	3	21	4
L2	4	22	3
L2	5	24	1
L2	6	23	2
L2	7	22	3
L2	8	23	2
L2	9	21	4
L2	10	17	8
L2	11	19	6
L2	12	22	3
L2	13	24	1
L2	14	24	1
L2	15	22	3
L2	16	21	4
L3	1	20	5
L3	2	15	10
L3	3	13	12
L3	4	15	10
L3	5	15	10
L3	6	17	8
L3	7	13	12
L3	8	17	8
L3	9	14	11
L3	10	17	8
L3	11	21	4
L3	12	12	13
L3	13	14	11
L3	14	12	13
L3	15	15	10
L3	16	21	4
L4	1	9	16
L4	2	18	7
L4	3	16	9
L4	4	11	14
L4	5	15	10
L4	6	14	11
L4	7	10	15
L4	8	11	14
L4	9	13	12
L4	10	12	13
L4	11	14	11
L4	12	14	11
L4	13	13	12
L4	14	13	12
L4	15	16	9
L4	16	17	8

ANEXO B. Dados da contagem de pulgão-do-algodoeiro em um delineamento de blocos casualizados com cinco tratamentos e dez blocos (Exemplo 2).

trat	bloco	cont
A	I	0
A	II	15
A	III	12
A	IV	27
A	V	0
A	VI	9
A	VII	23
A	VIII	184
A	IX	18
A	X	156
B	I	12
B	II	3
B	III	0
B	IV	15
B	V	2
B	VI	4
B	VII	0
B	VIII	8
B	IX	0
B	X	3
C	I	2
C	II	0
C	III	0
C	IV	8
C	V	9
C	VI	15
C	VII	6
C	VIII	0
C	IX	0
C	X	5
D	I	55
D	II	127
D	III	42
D	IV	37
D	V	58
D	VI	0
D	VII	187
D	VIII	39
D	IX	0
D	X	77
cont	I	510
cont	II	269
cont	III	0
cont	IV	895
cont	V	2356
cont	VI	278
cont	VII	1259
cont	VIII	987
cont	IX	0
cont	X	675

ANEXO C. Dados do número de plantas infestantes mortas em um experimento fatorial de blocos casualizados com três espécies de plantas infestantes e sete doses (Exemplo 3).

Especie	Dose	Bloco	morta	viva
Amend	0	I	0	40
Amend	5	I	13	27
Amend	10	I	18	22
Amend	15	I	25	15
Amend	20	I	29	11
Amend	25	I	40	0
Amend	30	I	40	0
Amend	0	II	0	40
Amend	5	II	14	26
Amend	10	II	19	21
Amend	15	II	25	15
Amend	20	II	32	8
Amend	25	II	40	0
Amend	30	II	40	0
Amend	0	III	0	40
Amend	5	III	17	23
Amend	10	III	20	20
Amend	15	III	26	14
Amend	20	III	30	10
Amend	25	III	40	0
Amend	30	III	40	0
Amend	0	IV	0	40
Amend	5	IV	16	24
Amend	10	IV	22	18
Amend	15	IV	25	15
Amend	20	IV	29	11
Amend	25	IV	39	1
Amend	30	IV	40	0
Caruru	0	I	0	40
Caruru	5	I	4	36
Caruru	10	I	3	37
Caruru	15	I	12	28
Caruru	20	I	17	23
Caruru	25	I	25	15
Caruru	30	I	40	0
Caruru	0	II	0	40
Caruru	5	II	3	37
Caruru	10	II	3	37
Caruru	15	II	7	33
Caruru	20	II	16	24
Caruru	25	II	29	11
Caruru	30	II	39	1
Caruru	0	III	0	40
Caruru	5	III	3	37
Caruru	10	III	6	34
Caruru	15	III	10	30
Caruru	20	III	19	21
Caruru	25	III	29	11
Caruru	30	III	40	0
Caruru	0	IV	0	40
Caruru	5	IV	6	34
Caruru	10	IV	5	35
Caruru	15	IV	10	30
Caruru	20	IV	17	23
Caruru	25	IV	32	8
Caruru	30	IV	40	0
Corda	0	I	0	40
Corda	5	I	10	30
Corda	10	I	20	20
Corda	15	I	25	15
Corda	20	I	39	1
Corda	25	I	40	0
Corda	30	I	39	1
Corda	0	II	0	40

Corda	5	II	13	27
Corda	10	II	21	19
Corda	15	II	29	11
Corda	20	II	40	0
Corda	25	II	38	2
Corda	30	II	40	0
Corda	0	III	0	40
Corda	5	III	13	27
Corda	10	III	20	20
Corda	15	III	28	12
Corda	20	III	40	0
Corda	25	III	39	1
Corda	30	III	40	0
Corda	0	IV	0	40
Corda	5	IV	14	26
Corda	10	IV	20	20
Corda	15	IV	30	10
Corda	20	IV	40	0
Corda	25	IV	40	0
Corda	30	IV	40	0

CAPÍTULO III

THE USE OF GENERALIZED ADDITIVE MODELS AND BAYESIAN STATISTICS HELP TO SOLVE THE OVERDISPERSION, AUTO-CORRELATION AND ZERO INFLATION PROBLEMS IN APHID POPULATION STUDY.

Fábio Janoni Carvalho

Abstract: Count variables are often positively skewed and include many zero observations, requiring specific statistical approaches. Abiotic conditions, such as temperature and precipitation, are known to affect the insect physiology and behavior. Interpreting the role of abiotic factors in changes in insect populations of crop pests can be difficult. For these data, the analysis becomes even more complicated because of possible temporal or spatial correlation, irregular spaced data, and heterogeneity over time. Generalized Additive Models (GAMs) are important tools to evaluate abiotic factors. Current software programs for generalized linear mixed modelling do not easily allow incorporating temporal correlation, and therefore Markov Chain Monte Carlo (MCMC) techniques can be used to fit a model that contains a temporal correlation structure, based on Bayesian statistics (BGAMs). In addition, when zero inflation is ignored in the analysis of insect incidence data, there are two consequences: the estimated parameters and standard errors may be biased and the excessive number of zeros can cause overdispersion. We compared methods of modelling the effects of temperature, precipitation and time for *Brevicoryne brassicae* (L.) population in Uberlândia, Minas Gerais. We applied the proposed BGAM to the data, comparing this to the frequentist model (GAM) with and without autocorrelation for time, using the software *R* as background for the model's construction. Analysis of *deviance* identified significant effects of the smoothers for precipitation and time on the frequentist models. For these models, residual analysis showed a clear pattern in the residuals over the fitted values because of the excess of zeros. With BGAM, the problem in variance estimations for precipitation and temperature from the previously models were resolved. The estimated smoothing curves showed a linear effect with an increase of precipitation, where lower precipitation indicated no presence of the aphid. With the precipitation of 150mm, it is expected the mean quantity of six aphids per plant, but also the highest variation, from one to 34 aphids. The average temperature did not affect the *B. brassicae* incidence. Extending the analysis to a binomial negative distribution helps to solve the overdispersion problem. Auto-correlation is solved with ARMA structures and the excess of zero was solved with zero inflation models. All these approaches are available in Bayesian models in *R*. The flexibility of Bayesian analysis comes at a cost in time-consuming simulation methods. Fortunately, freeware software is evolving and making the simulation even faster. The example of. *B. brassicae* incidence showed how well abiotic (and biotic) factors can be modeled and analyzed using BGAMs.

Keywords: abiotic factors; ARMA structure; *Brevicoryne brassicae*; Markov Chain Monte Carlo simulation; regular time series event.

MODELOS ADITIVOS GENERALIZADOS E ANÁLISE BAYESIANA AUXILIAM NA CORREÇÃO DA SOBREDISPERSÃO, AUTOCORRELAÇÃO E ZEROS INFLACIONADOS, EM ESTUDO SOBRE INCIDÊNCIA DE PULGÃO.

Fábio Janoni Carvalho

Resumo: Variáveis de contagem tendem a ser positivamente enviesadas e incluir muitas observações nulas, necessitando abordagens estatísticas específicas. Fatores abióticos, como temperatura e precipitação, afetam a fisiologia e comportamento de insetos. Avaliar o efeito de fatores abióticos na flutuação populacional de pragas agrícolas pode ser difícil. Para estes dados, a análise se torna ainda mais complicada por causa de possíveis correlações (temporal ou espacial), coleta de dados no tempo com datas irregulares, e heterogeneidade ao longo do tempo. Modelos Aditivos Generalizados (MAGs) são uma importante ferramenta para avaliar fatores abióticos. Os programas computacionais estatísticos atuais para Modelos Lineares Generalizados Mistos não incorporam facilmente a correlação temporal. A técnica da Cadeia de Markov Monte Carlo (MCMC) pode ser utilizada para permitir uma estrutura de correlação temporal, baseado na técnica bayesiana (BMAGs). Ainda, quando zero inflacionados são ignorados da análise da incidência de insetos, há duas consequências: os parâmetros estimados e os erros padrão se tornam tendenciosos, e o excesso de zeros causa sobredispersão. Foi comparado métodos de modelagem dos efeitos da temperatura, precipitação e tempo para a população do afídeo *Brevicoryne brassicae* (L.) em Uberlândia, Minas Gerais. Foi aplicado o BMAG aos dados, comparando-o com outras estatísticas frequentistas (MAGs) com e sem autocorrelação temporal, utilizando o software R para a construção dos modelos. A análise de *deviance* identificou efeitos significativos para as curvas de precipitação e tempo nos modelos frequentistas. Nestes modelos, a plotagem dos resíduos com os valores ajustados mostrou um desvio provocado pelo excesso de zeros. Com BMAG, o problema na estimação das variâncias para a temperatura e precipitação dos modelos anteriores foi resolvida. As funções suavizadas das curvas mostraram efeito linear com o acréscimo da precipitação, onde menores precipitações não apresentaram o inseto. Com a precipitação de 150mm, é esperado a quantidade média de seis afídeos por planta, mas também a maior variabilidade, de um até 34 pulgões. A temperatura média não afetou a incidência de *B. brassicae*. A expansão da análise com a distribuição binomial negativa ajudou na resolução do problema de sobredispersão. A autocorrelação foi resolvida com as estruturas ARMA e o excessos de zeros com os modelos para zeros inflacionados. Todas estas metodologias estão disponíveis no R para a modelagem bayesiana. A flexibilidade da análise vem com o custo do tempo gasto com os métodos de simulação. Felizmente, os softwares estão evoluindo e realizando a simulação cada vez mais rápido. O exemplo da incidência de *B. brassicae* mostrou como fatores abióticos (e bióticos) podem ser bem modelados e analisados pelos BMAGs.

Palavras-chave: fatores abióticos; estrutura ARMA; *Brevicoryne brassicae*; simulação da Cadeia de Markov Monte Carlo; eventos de série temporal.

1) INTRODUCTION

Many researchers study the ecology of an organism counting its presence in a studied environment. However, count variables are often positively skewed and include many zero observations (ATKINS et al., 2013), requiring specific statistical approaches. Techniques such as Poisson, negative binomial regression, or zero-altered count models (e.g., zero-inflated or hurdle models) are much more appropriate for these type of data (ATKINS; GALLOP, 2007; NEAL; SIMONS, 2007; COXE et al., 2009; HILBE, 2011). In entomology, it is very common to evaluate how biotic and abiotic factors affect the physiology, behavior and population dynamics of the insect (PRICE et al., 2011), as an example, for wheat aphids (JAN et al., 2017) and *Drosophila suzukii* (HAMBY et al., 2016).

The population of an insect can be affected by the composition and configuration of the surrounding landscape. Abiotic conditions, such as temperature and precipitation, are known to affect the insect physiology and behavior (WHITNEY et al., 2016). Insects are poikilothermic organisms, being especially sensitive to variation in temperature (PARMESAN, 2007). Because of that, degree-day-based models have been used to predict the emergence and growth of insect pests (WILSON; BARNETT, 1983). Variation in precipitation and moisture availability may have both direct and indirect effects on herbivore insects. Rainfall can impede aphid dispersal (THACKRAY et al., 2004) and dislodge aphids from host plants, potentially leading to mortality from impact, predation, or starvation (WINDER, 1990).

Interpreting the role of abiotic factors in changes in insect populations of crop pests can be difficult. For these data, the analysis becomes even more complicated because of possible temporal or spatial correlation, irregular spaced data, and heterogeneity over time (ZUUR et al., 2009). The Generalized Linear Mixed Models (GLMMs) arise as an alternative to include this information in the model. GLMMs are an extension of the Generalized Linear Models (GLMs), where the linear predictor contains random effects in addition to the usual fixed effects (STROUP, 2012).

Another important analysis to evaluate abiotic factors, like precipitation or temperature, are the Generalized Additive Models (GAMs), which are a type of GLM where the linear predictor is given by a user specified sum of smooth functions of the covariates plus a conventional parametric component of the linear predictor (WOOD,

2017). GLMs and GAMs are appropriate for discrete and continuous variables, like the count process. These models are already applied for agronomic data. However, their application is scarce and with few publications that help the researchers to use this type of analysis.

Current software programs for generalized linear mixed modelling do not easily allow incorporating temporal correlation, and therefore Markov Chain Monte Carlo (MCMC) techniques can be used to fit a model that contains a temporal correlation structure, based in Bayesian statistics. MCMC was originated with the classic paper of Metropolis et al. (1953), where it was used to simulate the distribution of states for a system of idealized molecules. In 1987, a landmark paper united the MCMC and molecular dynamics approaches, calling their method "Hamiltonian Monte Carlo" which abbreviates to "HMC" (DUANE et al., 1987). Statistical applications of HMC began with neural network models (NEAL, 1996).

An explosive growth and spread of the Bayesian approach has occurred only recently because of the arrival of MCMC, which in turn became popular because of advances in computation (GHOSH et al., 2006). The development of a freeware software implementing such simulation tools has also helped greatly to popularize Bayesian approaches.

In a frequentist statistic, we formulate a hypothesis for the regression parameters, apply the model and estimate the parameters, standard errors, 95% confidence intervals, and *p*-values. The parameters, such as mean, variance, and regression coefficients, are fixed, but unknown. Based on observed data, the unknown parameters are estimated in such a way that the observed data agrees well with the statistical model. With that, frequentist approaches are objective and only the information contained in the current dataset is used to estimate the parameters (ZUUR et al., 2009).

As in classical or frequentist approach to inference, the Bayesian method is developed in the presence of observation x whose values are initially uncertain and described through a probability distribution with probability function defined as $f(x|\Theta)$. The quantity Θ serves as an index of the possible family distributions for the observations. In Bayesian analysis prior knowledge of index Θ can be incorporated in the analysis diverging from frequentist approaches. Frequentist analysis do not admit previous information because it has not been observed and therefore not subject to empiric verification (GAMERMAN; LOPES, 2006). Incorporating the GAMs structure, the Bayesian Generalized Additive Models (BGAMs) became an extremely good option to

analyze incidence data. The BGAMs allow for a smooth flexible trend to be fitted (ISHWARAN; RAO, 2005).

Two main packages were implanted in *R* to analyze GAMs and BGAMs: `mgcv` and `brms`, respectively. For GAMs, the package `mgcv` fits the model with the function `gam`, including any quadratically penalized GLM and a variety of other models estimated by a quadratically penalized likelihood type approach. The smooth terms are represented using penalized regression splines with smoothing parameters selected by GCV, UBRE, AIC, REML or by regression splines with fixed degrees of freedom (WOOD, 2019).

For BGAMs, the package `brms` supports a wide range of distributions and link functions, allows for multiple grouping factors, each with multiple group-level effects, autocorrelation of the response variable, defined covariance structures, as well as flexible and explicit prior specifications, using `stan` on the back-end. Accordingly, all samplers implemented in `stan` can be used to fit BGAMs. In addition, `brms` allows drawing samples from the posterior predictive distribution as well as from the pointwise log-likelihood. Both can be used to assess model fit, allowing a comparison between the actual response y and the response $\hat{y}$ predicted by the model (BÜRKNER, 2017).

In addition, it is very common for experiments with insect counts to deal with an excess of null counts. When zero inflation is ignored, there are two consequences: the estimated parameters and standard errors may be biased and the excessive number of zeros can cause overdispersion (ZUUR et al., 2009). Approaches to analyze data with an excessive proportion of zeros in the response variable generally involve the creation of two datasets from the original data (FALK et al., 2015). One dataset is created with a binary response variable for whether or not the presence was detected, and logistic regression is used to analyze it. Another dataset for the abundance at present locations for count data is created and analyzed with ordinary regression (FLETCHER et al., 2005).

This chapter compared methods of modelling the effects of temperature, precipitation and time for *Brevicoryne brassicae* (L.) population in the Triângulo Mineiro region. We applied the proposed Bayesian Generalized Additive Model (BGAM) to the data, comparing this to the frequentist model (GAM) with and without autocorrelation for time, using the software *R* as background for the model's construction.

2) MATERIAL AND METHODS

2.1) Experimental data

The data used for the analysis were from an experiment conducted by Sampaio et al. (2017). The authors' study was carried out in Uberlândia, Triângulo Mineiro region of Minas Gerais - Brazil, with two enclosed areas in the period of July 2005 to March 2006 (1st area) and September 2006 to January 2008 (2nd area) to evaluate the influence of abiotic and biotic variables on Brassica aphids (*Lipaphis pseudobrassicae* (Davis), *Myzus persicae* (Sulzer), and *Brevicoryne brassicae*) under field conditions. Here, only the data of the aphid *B. brassicae* in the second area was modeled.

Collard greens (*Brassica oleracea* var. *acephala* (L.)) was cultivated as the aphid host plant. The experimental field had three rows, each one with 25 plants, giving a total of 75 plants. The spacing between plants was one meter between rows and 0.5 m between plants. To quantify aphid population dynamics, samples were taken on a weekly basis (69 samples). Each sample consisted of three randomly selected plants, one from each row in each plot. One leaf from each upper, middle, and lower positions per plant was removed and examined, totalizing three leaves per plant and nine leaves per data sample. The variable of the analysis was the sum of the identified *B. brassicae* species on the three leaves per plant (Figure 1). The sampling design included a restriction that the same plant would not be sampled again for another four weeks.

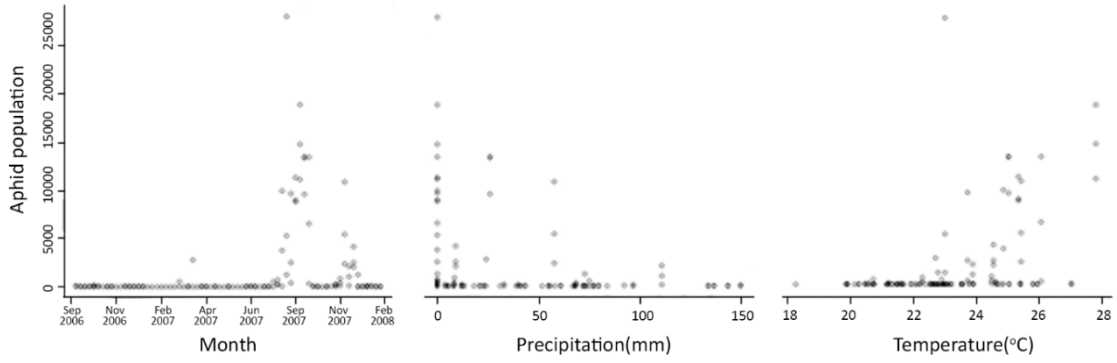


FIGURE 1. Number of *Brevicoryne brassicae* aphids over weeks, precipitation and average temperature in in Uberlândia- Minas Gerais, Brazil.

The aim of the Sampaio et al. (2017) paper was to examine the influence of abiotic and biotic variables on Brassica aphids under field conditions. The abiotic variables were precipitation (estimated as 7-day accumulated values), and minimum, average and

maximum temperature (Figure 2). Pearson's correlation showed a significant correlation ($p<0.01$) of maximum and minimum temperature with average temperature and precipitation (Table 1), so these variables (maximum and minimum temperatures) were dropped from the analysis, because of the importance to avoid using collinear explanatory variables in GAM (ZUUR et al., 2009).

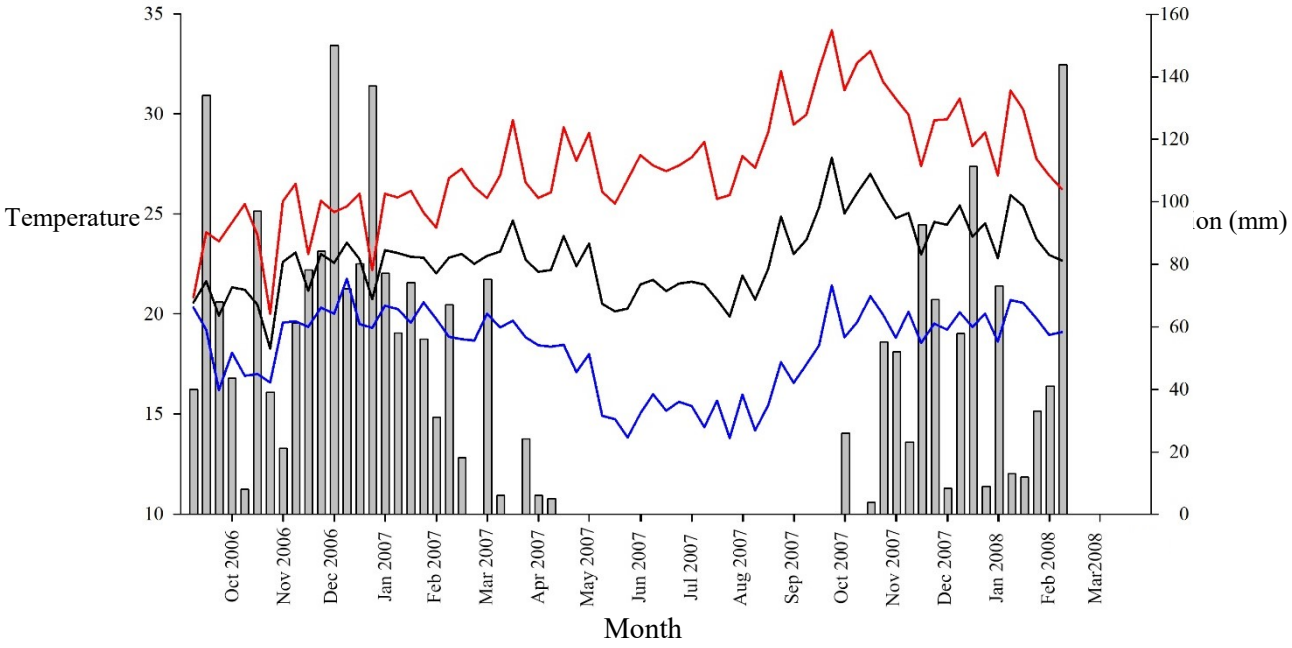


FIGURE 2. Temperature and precipitation conditions during the experiment of Sampaio et al. (2017) in the Triângulo Mineiro region. Maximum Temperature; Average Temperature; Minimum temperature; Precipitation.

TABLE 1. Pearson's correlation for abiotic variables measured in the experiment of Brassica aphids in the Triângulo Mineiro region (data collected from September 2006 to January 2008).

	Precipitation	*Max. temperature*	*Ave. temperature*	*Min. temperature*
Precipitation	1	-0.407**	-0.079	0.419**
Max. temperature	-0.407**	1	0.838**	0.141
Ave. temperature	-0.079	0.838**	1	0.659**
Min. temperature	0.419**	0.141	0.659**	1

Max.: Maximum; Ave.: Average; Min.: Minimum. **Correlation significant at the 0.01 level (2-tailed).

The interaction between precipitation and temperature was also dropped from the model adjustment because the precipitation measurements did not fully combine with the temperature measurements and the aim of the study was the isolated effect of each variable on the aphid's population. The factors were considered fixed because the error

determining the explanatory variable was small compared to the range of the explanatory variable (FARAWAY, 2005).

2.2) Data distribution

The incidence of *B. brassicae* aphid (AP) is considered count data, being suitable to follow a Poisson distribution. However, because the aphid population tends to exponentially grow over time (HUGHES; GILBERT, 1968), the variance was expressively higher than the mean, which made it impossible to use this distribution. To correct the high variance problem, the number of aphids followed a negative binomial distribution, described as:

$$E(AP_{ijk}) = \mu_{ijk} \text{ and } Var(AP_{ijk}) = \mu_{ijk} + \frac{\mu_{ijk}^2}{\theta}$$

$$\mu_i = e^{\alpha + f(PRECIPITATION_i) + f(TEMPERATURE_j) + f(TIME_k)} \quad \text{with } AP_{ijk} \sim NB(\mu_{ijk}, \theta)$$

where AP_{ijk} is the number of aphids in precipitation *i*, in temperature *j* and at time *k*. The notation $f(X)$ represents the smoothing function of the explanatory variable *X*, and *NB* is a negative binomial distribution with mean μ_{ijk} and dispersion parameter θ. The relationship between the mean μ_{ijk} and systematic component was specified by the logarithmic link function: $\log(\mu_{ijk}) = \eta_{ijk}$, which can also be written as $\mu_{ijk} = \exp(\eta_{ijk})$.

2.3) Temporal correlation

Due to the independence assumption, residuals from different time points are not allowed to covariate. However, in time series events the violation of this independence is very common. This correlation needs to be modeled, introducing a correlation function (*h(.)*) that takes values from -1 and 1. Because the data was a regularly spaced time series, we assumed a correlation structure called autoregressive model (AR1) (BOX et al., 1994).

Referring ε_t as an observation taken at time *t* and ε_s at time *s*, the distance (*lag*) between two observations ε_t and ε_s is given by $|t - s|$, assuming that the correlation between both residuals only depends on their time difference. Hence, the correlation between ε_t and ε_s is the same that between ε_{t+1} and ε_{s+1}, and between ε_{t+n} and ε_{s+n} (ZUUR et al., 2009). An Autoregressive model (AR1) expresses the current observation as a linear function of the previous observations and adds a homoscedastic noise term, a_t, centered at zero and assumed independently from the previous observations (PINHEIRO;

BATES, 2000), in the form of: $\varepsilon_s = \rho\varepsilon_{s-1} + a_t$, where ρ is the correlation parameter. This cryptic notation stands for an auto-regressive model of order 1. Hence the correlation between ε_t and ε_s is equals to $\rho^{|t-s|}$. The further away two residuals are separated in time, the lower is their correlation (ZUUR et al., 2009).

2.4) Zero inflation

Another problem with biological data is the large number of zeros in the dataset. Zero-inflation models describe the probability of observing an extra zero that is not generated by the conditional model, where the overall distribution is a mixture of the conditional model and zero-inflation model. The zero-inflation probability is bounded between zero and one by using a logit link on the zero-inflation model (RHODES, 2015). We used a zero-inflation negative binomial (ZINB) model. In this model, the zeros were modelled as coming from two different processes: the binomial and the count process. The probability functions were:

$$f\left(AP_{ijk} = 0\right) = \pi_i + (1 - \pi_i) \times \left(\frac{\theta}{\mu_{ijk} + \theta}\right)^{\theta}$$

$$f\left(AP_{ijk} | AP_{ijk} > 0\right) = (1 - \pi_i) \times \left[\frac{\Gamma\left(AP_{ijk} + \theta\right)}{\Gamma(\theta) \times \Gamma\left(AP_{ijk} + 1\right)} \times \left(\frac{\theta}{\mu_{ijk} + \theta}\right)^{\theta} \times \left(1 - \frac{\theta}{\mu_{ijk} + \theta}\right)^{AP_{ijk}}\right]$$

where π_i is the probability of having a false zero.

2.5) Computer modeling

All models were performed in *R* software version 3.5.0 because of the model complexity, where *R* is the only available software capable of running all the following models. Moreover, some of the models have a certain complexity level that even *R* is not yet capable of running, for example a GAM with negative binomial distribution with zero inflation, auto-correlation of observations and a new variance structure. The summary of the three proposed models in *R* are presented in the Supplementary Material section.

2.6) First model: GAM (mgcv package)

The first model considered a smooth term for each of the factors adjusting a Generalized Additive Model (GAM) to the data with a negative binomial distribution

with a log link function. The model does not take into account the temporal correlation or zero inflation. The smoothing parameter was estimated by the Restricted Maximum Likelihood Method (REML) with a thin plate regression spline smooth (WOOD, 2003). The method views the smooth components as random effects. The variance component for each smooth random effect was given by the scale parameter divided by the smoothing parameter. REML is less prone to local minima than the other criteria, and may therefore be preferable (WOOD, 2017).

The basis estimation (*be*) is the dimensionality of the spline basis expansion of one or possibly more covariates. The value of *be* was estimated using the function `gam.check` which produces some diagnostic information about the fitting procedure and results. We adjusted $\gamma = 1.4$ to avoid overfitting without compromising the model (KIM; GU, 2004). The R code for the following model was: `gam(aphid ~ s(precipitation, k=9) + s(temperature, k=9) + s(time,k=15),family = nb, gamma=1.4)` with the package `mgcv`.

Significance of the smooths were tested with Analysis of *deviance* by Chi-Squared (χ^2) test. *Deviance* residuals were plotted against fitted values, precipitation, time and temperature to check model fitting. To detect any patterns over time, the auto-correlation function (ACF) was performed with the *deviance* residuals, where the value of the ACF at different time lags gives an indication whether there is any auto-correlation (ZUUR et al., 2009).

2.7) Second model: GAM with autocorrelation for time (mgcv package)

The second model followed the same structure of the first model with the addition of the Autoregressive model (AR1) for correlation measurement of the observations over time (PINHEIRO; BATES, 2000). The R code was: `gam(aphid ~ s(precipitation, k=9) + s(temperature, k=9) + s(time,k=15), correlation=corAR1(form=~time), family=negbin(th), gamma=1.4)` where `th` is the θ value for the negative binomial distribution, estimated as 0.237. Akaike Information Criterion – AIC (AKAIKE, 1974) was used to compare the first to the second model.

2.8) Third model: BGAM with zero inflation (brms package)

We assumed the number of *B. brassicae* aphids as a stochastic variable, with density function $f(AP|\Theta)$, where $AP = (y_1, \ldots, y_{207})$ with 207 observations and Θ is a vector containing unknown parameters estimated from the observed data. The major difference in Bayesian statistics is that instead of assuming that Θ is an unknown parameter vector, like the previous models, we now assume that Θ is stochastic. The prior distribution of Θ was obtained and denoted by $\gamma(\Theta)$. The prior information was combined with information from the data to give the posterior distribution of $\gamma(\Theta|y)$, which represents the information about Θ after observing the data of *B. brassicae* aphids. In contrast to maximum likelihood, where a point estimate for Θ is obtained, with Bayesian statistics a density of Θ was the final result. This density averages the prior information with information from the data, giving a posterior distribution (ZUUR et al., 2009).

Sampling of the posterior distribution was performed with the extension of Hamiltonian Monte Carlo algorithm (HMC) called no-U-turn sampler – NUTS (HOFFMAN, GELMAN; 2014), an MCMC method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult (NEAL, 2003). NUTS was used because it produces samples that are much less autocorrelated than those of other samplers and allows setting HMC parameters automatically (BÜRKNER, 2017). The behavior of the NUTS sampler was adjusted with the `control` argument to decrease the number of divergent transitions (`adapt_delta = 0.99`) and not exceed the depth of the tree evaluated in each iteration (`max_treedepth = 15`).

The `brms` package provides the functions to fit GAMs using *Stan* for full Bayesian inference. The model followed a zero-inflation negative binomial (ZINB) model with a log link function and a binomial model with logistic link for the model of false zeros versus the non-false zeros (the true zeros plus the positive counts). The R code was:

```
brm(aphid ~s(precipitation, k=9)+ s(temperature, k=9)+ s(time, k=15),
data=data, control = list(adapt_delta = 0.99, max_treedepth = 15),
family=zero_inflated_negbinomial(link = "log", link_shape = "log",
link_zi = "logit"), autocor = cor_bsts(formula = ~time)).
```

The probability of having a false zero was modeled as $\pi_i = \frac{e^{\upsilon}}{1+e^{\upsilon}}$, being υ the intercept. In other words, the probability to obtain a false zero is only related with the intercept and no other effects. To check the goodness of fit, empirical cumulative distribution function (CDF) of the observations and random draws from the posterior model for the population of *B. brassicae* were plotted, using the function `pp_check`, with the argument `type = "ecdf_overlay"`.

3) RESULTS

3.1) First model: GAM

Analysis of *deviance* identified significant effects of the smoothers for precipitation ($p<0.05$) and time ($p<0.01$) by Chi-Squared test, with an AIC of 1868.561. The smoothers explained only 37.4% of the variance in the data and 60.0% of the null deviance; 1.00, 1.00 and 12.32 degrees of freedom were used for precipitation, average temperature and time respectively.

The model suggested that precipitation has a linear effect on the aphid population (Figure 3), with a lack of adjustment on the edges, especially with higher precipitation (above 100 mm), because few observations were taken between these intervals, penalizing the estimation and rising the confidence interval. The incidence of *B. brassicae* rose from zero, at 0 mm, to 5.5 aphids, at 150 mm of precipitation. For time, with 12.32 degrees of freedom, the population fluctuation was higher over time, varying from zero at the end of January 2007 to 403 aphids at the middle of September 2007. The confidence interval was smaller, because the measurements over time followed a fixed interval.

Auto-correlation function (ACF) determines what type of time series model might be required for the residuals. A possible correlation was detected for small lags by ACF, but a part of this correlation comes from the excess of zeros in the dataset and not from time correlation (Figure 4). Residual analysis showed a clear pattern in the residuals over the fitted values, again because of the excess of zeros (Figure 5). This pattern is also seen when the residuals were plotted against precipitation, because of the large presence of days without rain (29.7% of data). When the *deviance* residuals were plotted against time or temperature, no pattern was verified. Residual analysis showed the need for the use of a zero inflated model (ZINB).

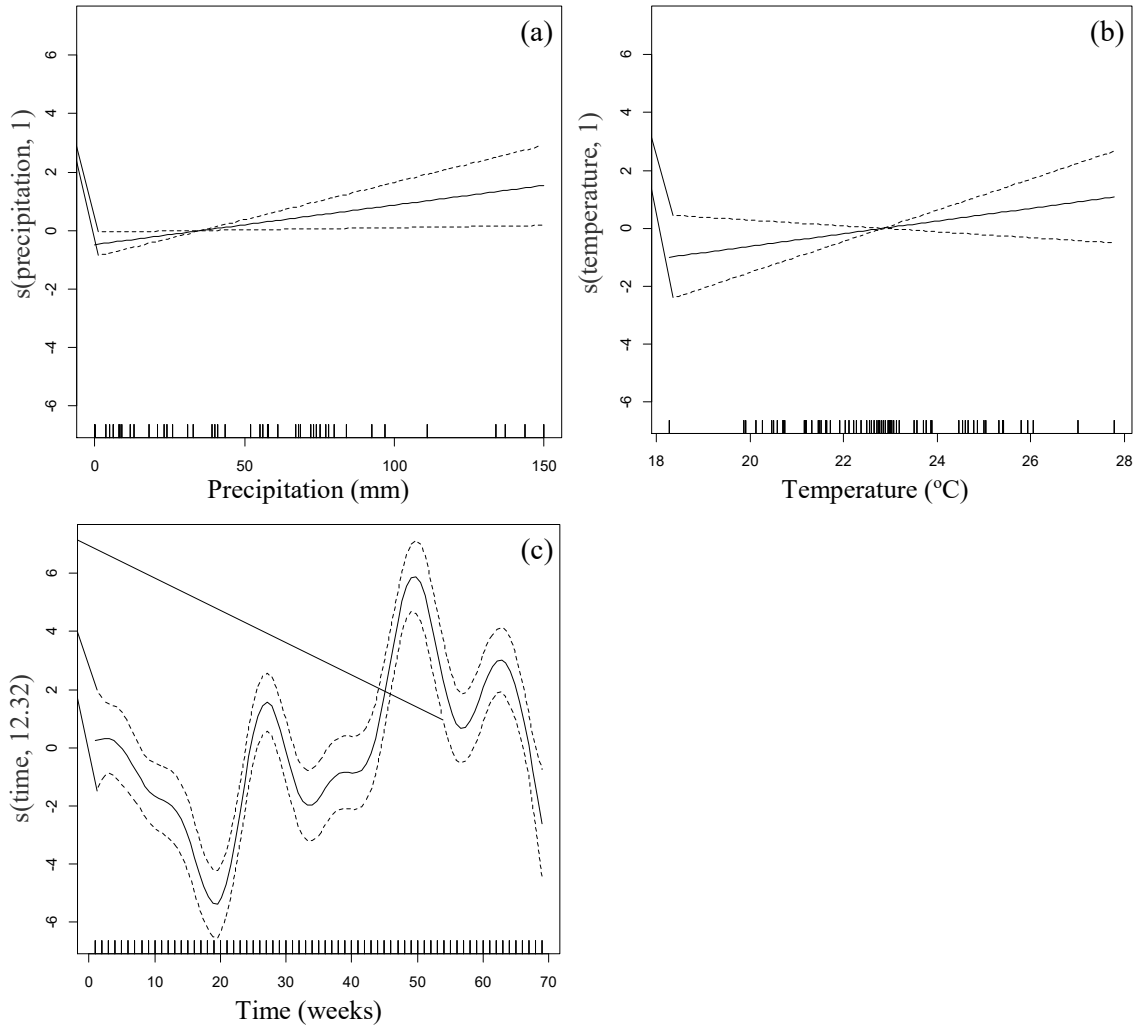


FIGURE 3. Estimated smoothing curves for precipitation (a), average temperature – not significant (b), and time (c) for the log of the population of *B. brassicae* for the first model. The solid line is the smoother and the dotted lines are 95% point-wise confidence bands.

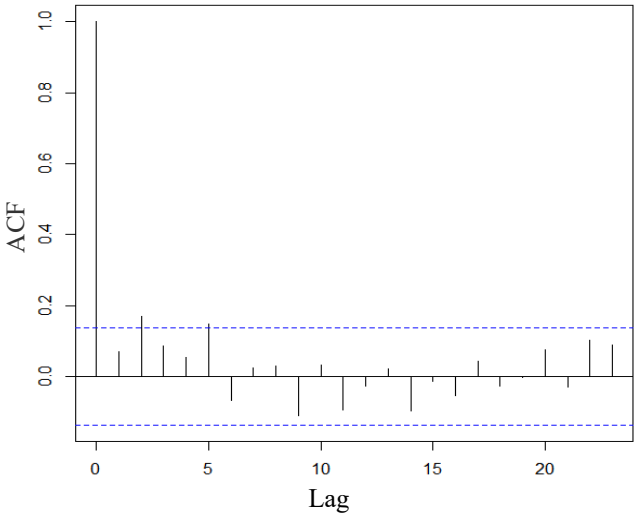


FIGURE 4. Auto-correlation plot for the residuals (ACF) obtained by applying linear regression on the *B. brassicae* time series (Lag).

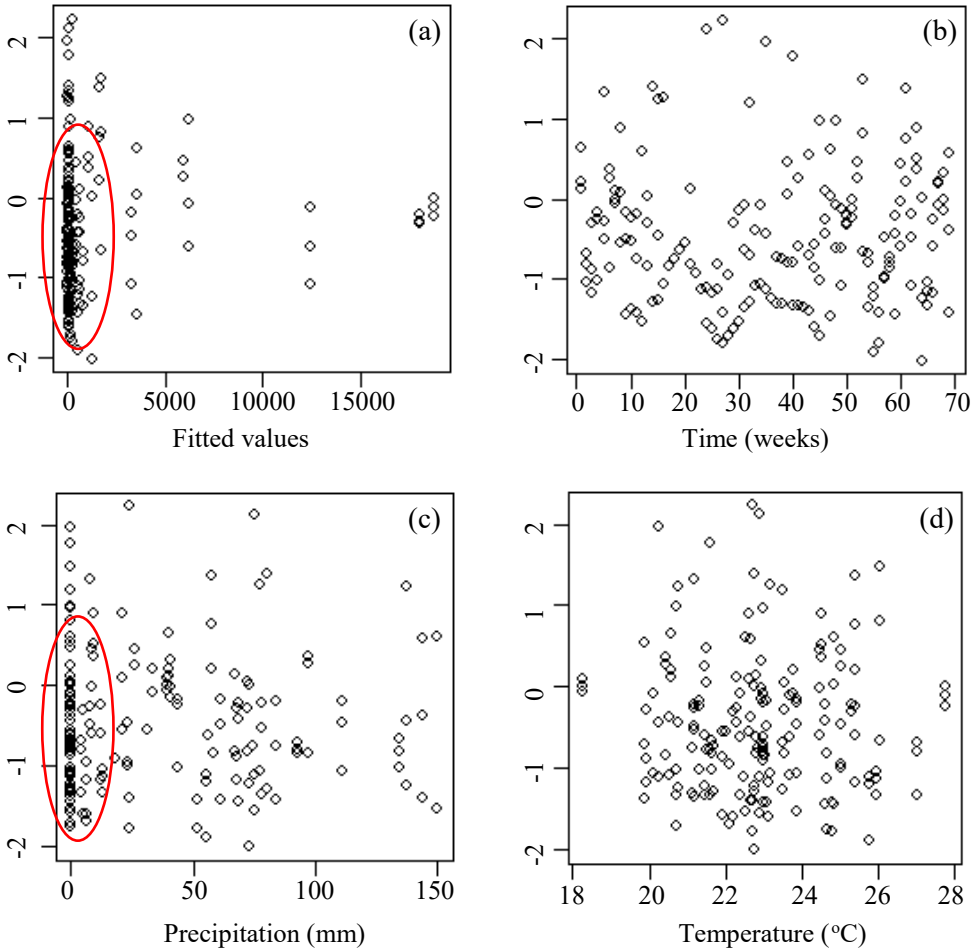


FIGURE 5. *Deviance* residuals (*y* axis) plotted versus fitted values (a), time (b), precipitation (c) and temperature (d) for the number of *B. brassicae* in Triângulo Mineiro, Minas Gerais (September 2006 to January 2008) from the first model. The red circles show a residual pattern.

3.2) Second model: GAM with autocorrelation for time

The model achieving the autocorrelation in time did not differ from the first model. Analysis of *deviance* also identified significant effects of the smoothers for precipitation ($p<0.05$) and time ($p<0.01$) by Chi-Squared test, with an AIC of 1865.428. The smoothers explained the same amount of null deviance and the graphs presented the same behavior (Figure 6), showing that the autoregressive model (AR1) was not helpful to explain the correlation and variance of the model. Residual analysis also detected some lack of adjustment from the model (Figure 7). The second model had a reduction of AIC of only 3.133, corroborating the results that AR1 was not helpful for the model.

Even with the belief that temporal data is autocorrelated, comparisons with the first and second models did not indicate differences between them. These models actually report some divergent variation structure for zero precipitation (Figure 5 and 7) that could

not be resolved with autocorrelation models, but could be solved with a new variance structure for precipitation.

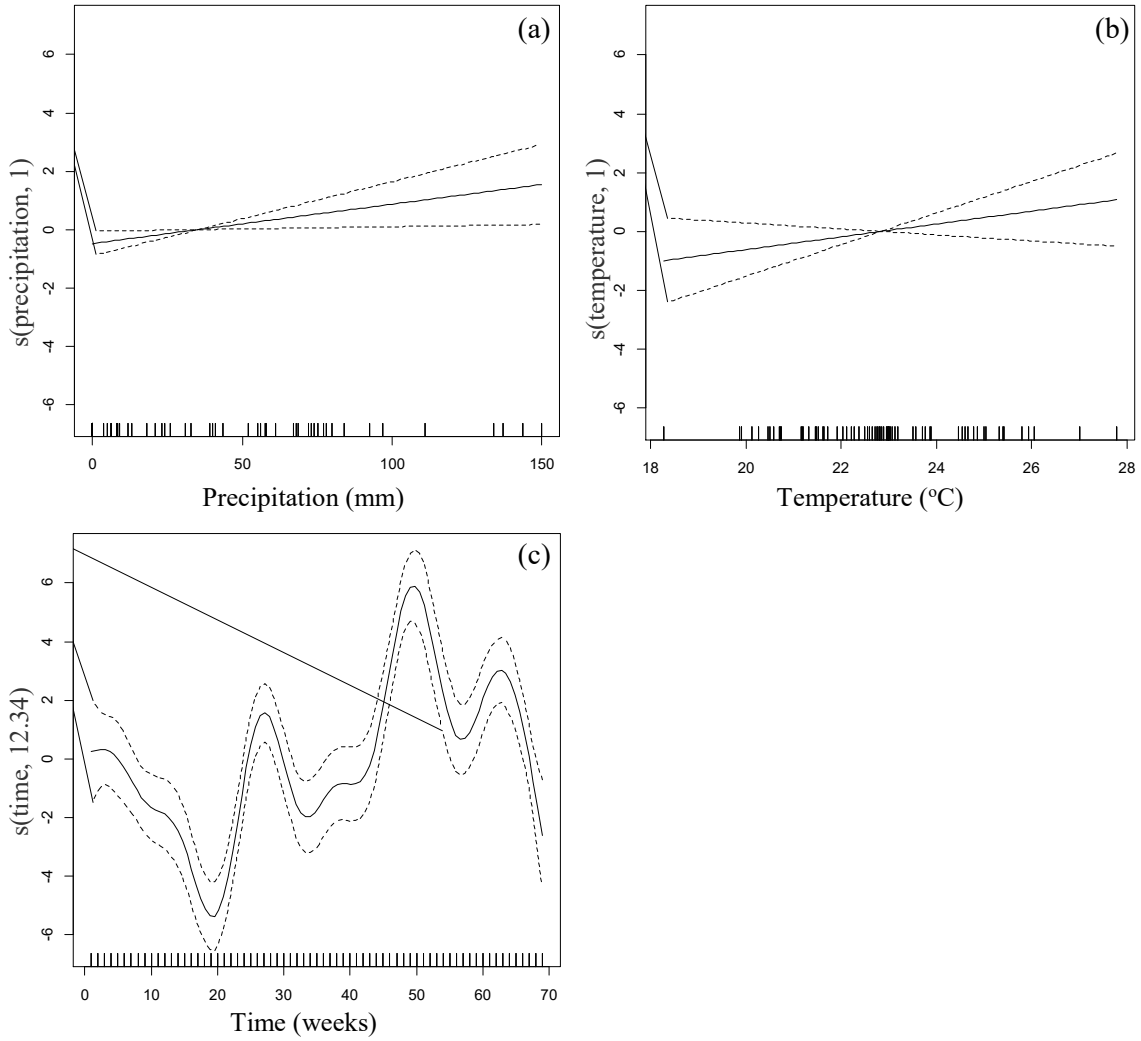


FIGURE 6. Estimated smoothing curves for precipitation (a), average temperature – not significant (b), and time (c) for the log of the population of *B. brassicae* in second model. The solid line is the smoother and the dotted lines are 95% point-wise confidence bands.

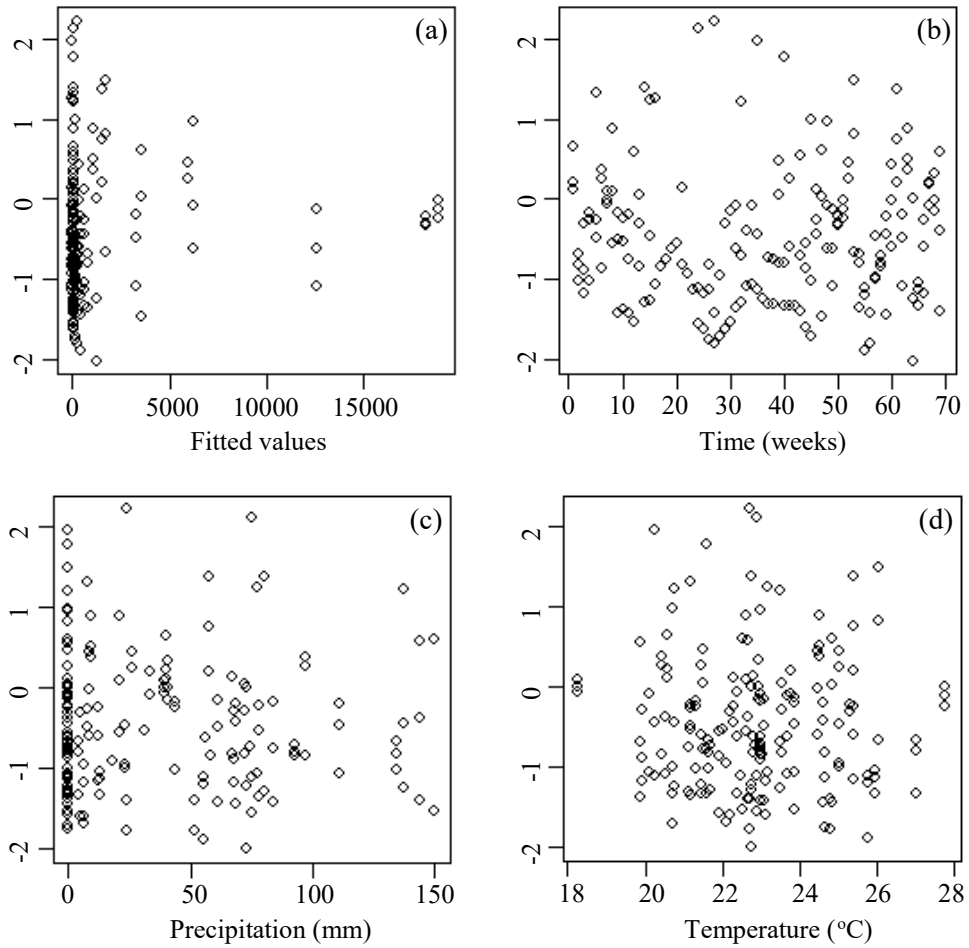


FIGURE 7. *Deviance* residuals (*y* axis) plotted versus fitted values (a), time (b), precipitation (c) and temperature (d) for the number of *B. brassicae* in Triângulo Mineiro, Minas Gerais (September 2006 to January 2008) from the second model.

3.3) Third model: BGAM with zero inflation

With Bayesian methods, the effective sample size of the posterior samples was 1725 for precipitation, 1726 for temperature and 880 for time in the smooth terms. These estimations revealed the number of independent samples from the posterior distribution that would be expected to produce the same standard error of the posterior mean as it was obtained from the dependent samples returned by the MCMC algorithm (BÜRKNER, 2017). Gelman-Rubin diagnostic (GELMAN; RUBIN, 1992) indicated that the chains had converged well and were equal to 1.

The model was fitted using four chains, each with 2000 iterations, where the first 1000 were warmup to calibrate the sampler, leading to a total of 4000 posterior samples. The estimate variance parameters for precipitation and temperature were 2.72 and 1.79, respectively (Appendix C). These lower values indicated a linear effect from temperature

and precipitation. For time, with a variance parameter of 87.14, the value reveals that the smooth was wigglier than precipitation or temperature. Even if the data presented 32.85% of zero data, the probability of having a false zero in the ZIBN model was estimated in 3.00%. That percentage implies that nearly 1% of the zeros presented in data were false zeros.

Comparing the variance component representation of the smooth with the previous models, the first model estimated 0.0158 for precipitation, 0.0125 for temperature and 75.10 for time. The second model estimated 0.0200 for precipitation, 0.0306 for temperature and 128.99 for time. The range of the variance component of the smooth in the Bayesian model was estimated from 0.07 to 11.52 for precipitation, 0.04 to 7.05 for temperature and 47.55 to 142.79 for time, in a 95% credible interval. The range estimation for the first model was way too large for precipitation and temperature, and for time it was from 41.68 to 135.33, being similar to the Bayesian model. Bayesian statistics resolved the problem in variance estimations for precipitation and temperature of the previous models.

The estimated smoothing curves showed a linear effect with an increase of precipitation, where lower precipitation indicated no presence of the aphid. With the precipitation of 150mm, a mean quantity of 6 aphids per plant is expected, and also the highest variation, from 1 to 34 aphids, in a credible interval of 95% (Figure 8a). The increase in temperature, elevated the incidence from 0.4 (18.3ºC) to 2.8 aphids (27ºC) per plant, with higher variations of the aphid in temperatures lower than 20ºC or higher than 25ºC (Figure 8b). The average temperature of the Cerrado Mineiro during September 2006 to January 2008 did not seem to affect the *B. brassicae* incidence. The smoothing curve for time over the seasons of the year (Figure 9) suggested that the aphid's population tends to decrease in the Spring. In the middle of the Summer, with higher precipitation, the population increased. In the Autumn, it seems that the population stabilizes and, in the Winter, the population tends to expand.

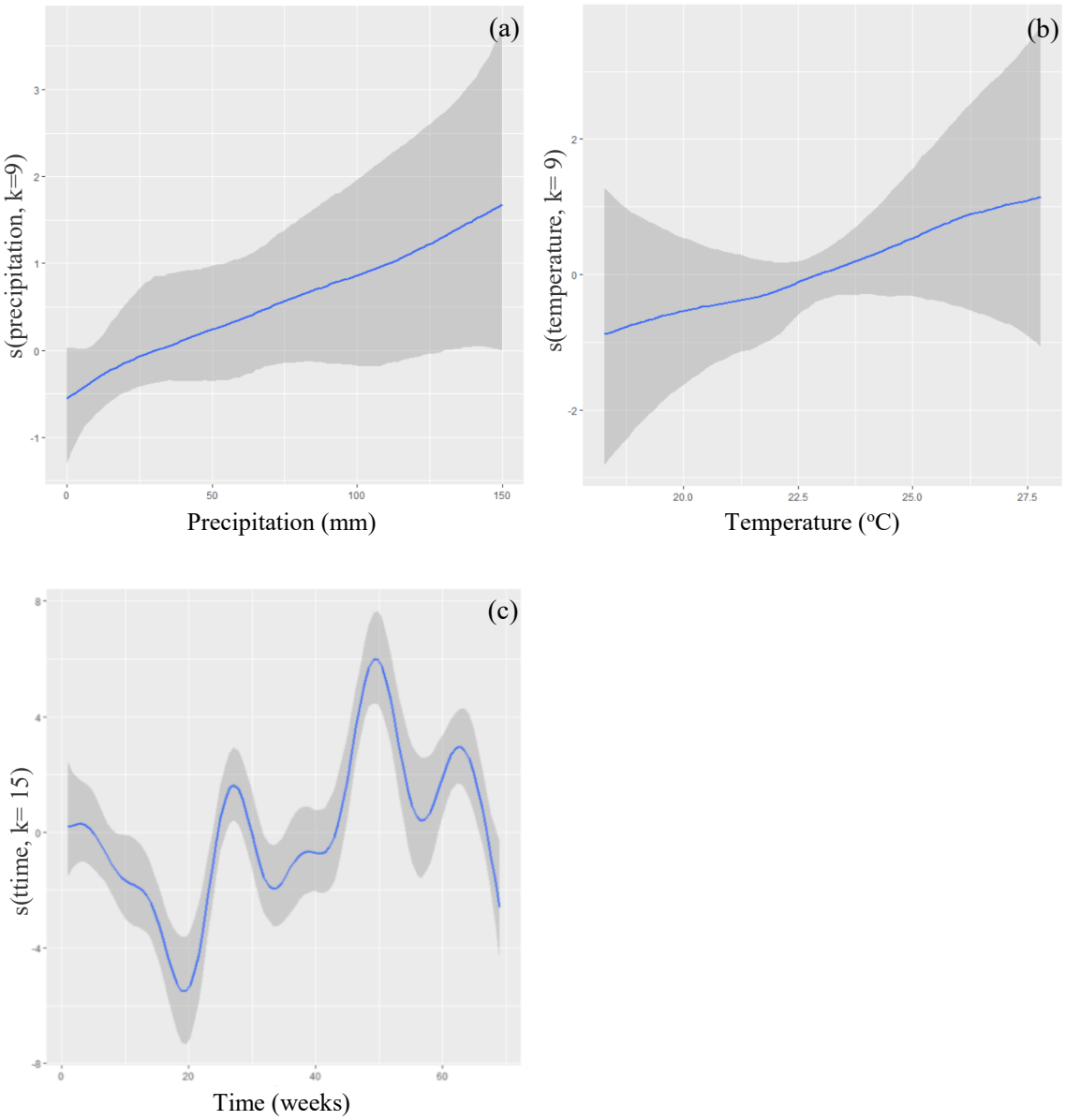


FIGURE 8. Estimated smoothing curves for precipitation (a), average temperature (b), and time (c) for the population of *B. brassicae* in Bayesian model. The blue line is the smoother and the dark grey shows 95% point-wise confidence bands.

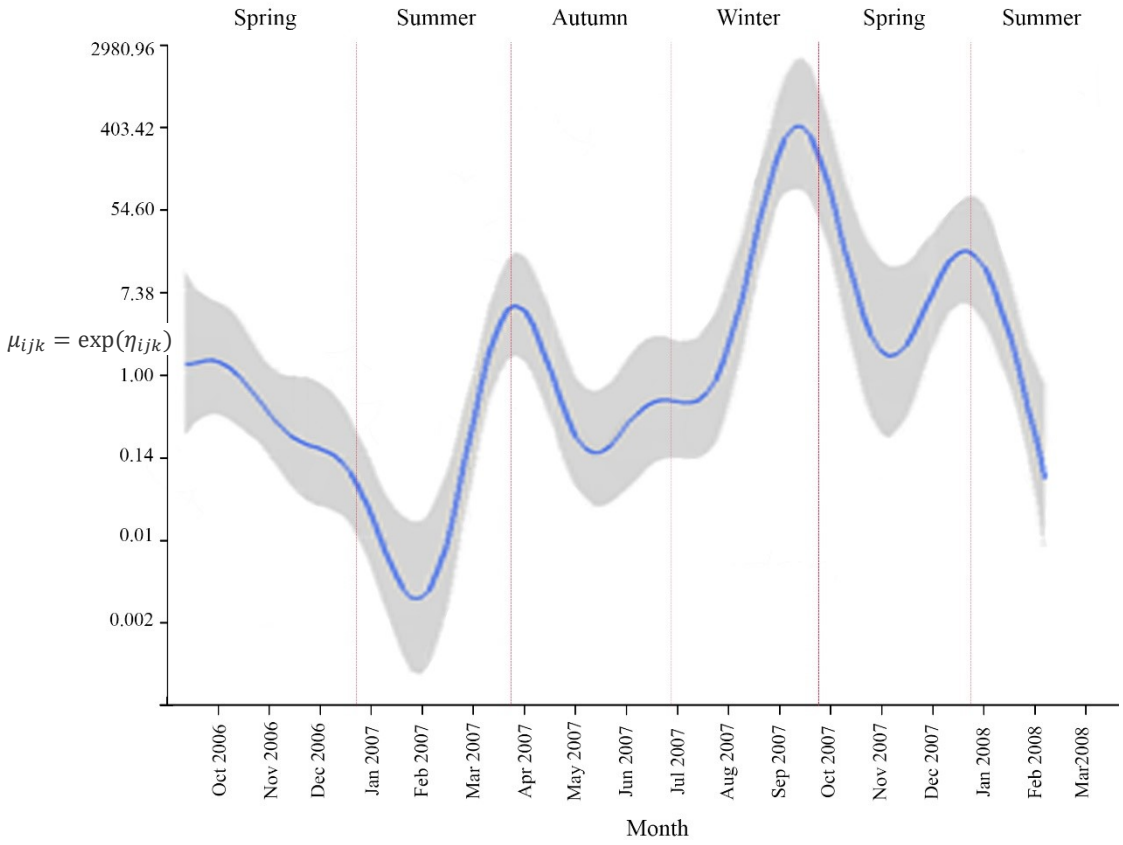


FIGURE 9. Time estimated smoothing curve for the population of *B. brassicae* over the seasons of the year, in the Triângulo Mineiro region from September 2006 to January 2008). The blue line is the smoother and the dark grey shows 95% point-wise confidence bands.

Trace and density plots for population-level effects for all parameters converged well, with the four chains overlapping (Appendix D). The parameter represented the fixed effect part of the splines, which is the linear function of the smooth, and is not a useful result for GAMs. For the variance parameters (Figure 10), the first two chains from the trace started from different values and then gradually converged to the last chain, especially for π_i, temperature and time parameters. The initial part, where the chains do not overlap, was the burn-in period and reflects that the chains have not converged to the same stationary distribution yet. However, the burn-in period was very brief. The first chain was more sinuous than the others, which revealed a larger burn-in period for that chain.

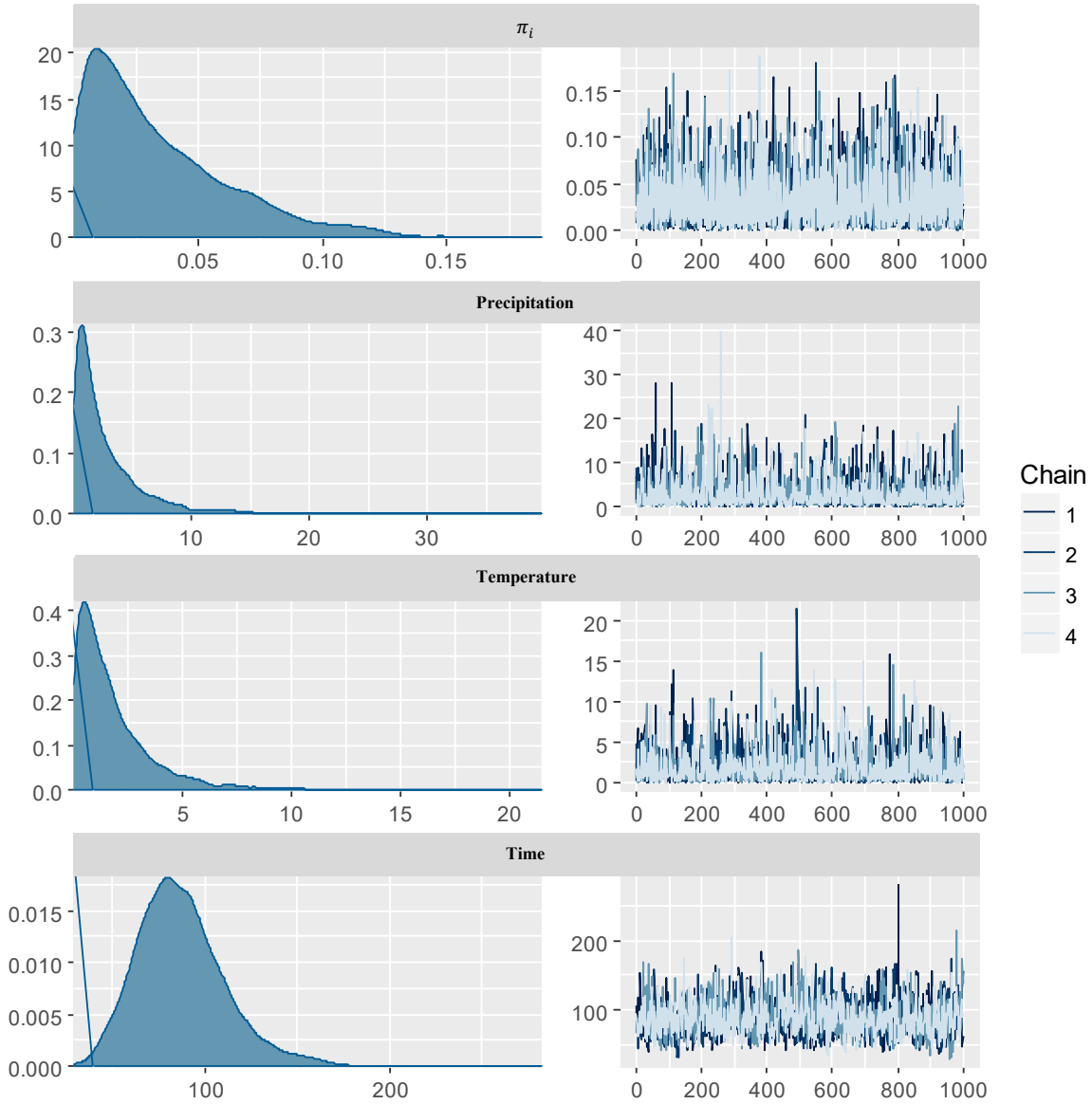


FIGURE 10. Zero Inflation parameter and smooth terms of precipitation, temperature and time from the Bayesian model.

A type of posterior predictive check plot is the empirical cumulative distribution function (CDF) of the observations and random draws from the posterior model (Figure 11). The CDF from the chains and the observations showed a good adjustment for small *B. brassicae* populations. With higher estimations, the chains tended to overestimate the population up to the number of 12500 aphids. Then, the chains tended to underestimate the population. CDF showed that some observations of the aphid appear to have different variances, which need to be modeled.

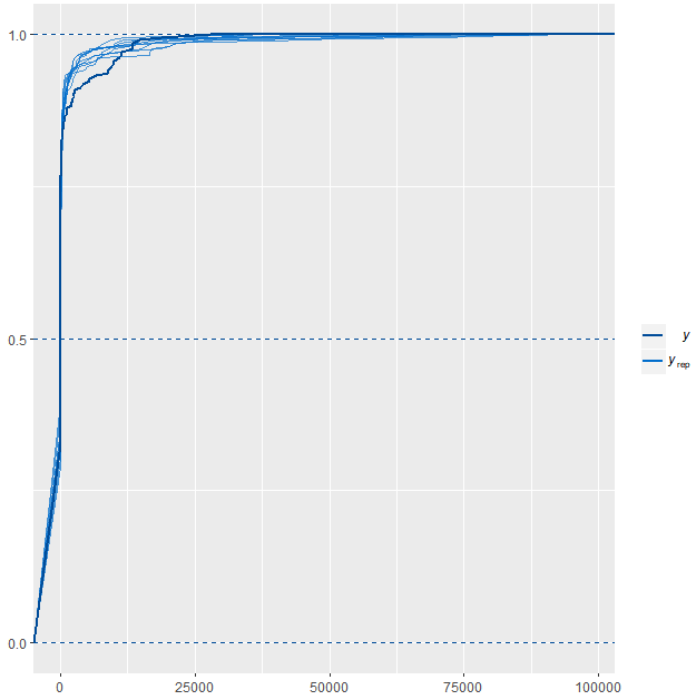


FIGURE 11. Empirical cumulative distribution function of the observations (y) and random draws from the posterior model (y_{rep}) for the population of *B. brassicae* in Bayesian model.

4) DISCUSSION

For count data in entomology, the assumption of normal distribution for the variable will be hardly met. Count data are discrete, non-negative and integer, describing a Poisson distribution. However, an insect population normally deals with a high reproductive potential and zero excess in data, impacting directly the data variance and violating the assumption of variance and mean equality from the Poisson distribution. To solve this problem, the negative binomial distribution adds an additional scale term, providing a more realistic distribution for biological data (STROUP, 2015). Negative binomial distribution is often the best fit for abundance (GUISAN; ZIMMERMAN, 2000; WELSH et al., 2000; WARTON, 2005).

Our study contributes to fill the gap of how ecological and agronomical researches can explore the effect of abiotic measurements in the biology of a living being. The major problem is that we need to take into account the real variation of this abiotic factor in the field, where it cannot be controlled. Abiotic factors can be considered fixed because these explanatory variables are deterministic. However, some authors consider if we take at a random sample and then measure it, for example the temperature in the field, then it is a random effect (ZUUR et al., 2009). If the error in determining the explanatory variable is

small compared to the range of the explanatory variable, we avoid unbiased regression parameters and can consider them as fixed terms. For example, if 30 samples have the temperature varying between 15 and 25ºC and the error of the thermometer is 0.1ºC, then we can consider the temperature as a fixed factor (FARAWAY, 2005).

The use of GAMs to explain the variability of quantitative abiotic factors presented a good fit in other biological researches, where GAMs were used to model water temperature, chlorophyll *a* concentration, and hydrophobic compounds on marine zooplankton (EVERAERT et al., 2018); to model water temperature and depth on the fish delta smelt (*Hypomesus transpacificus*) population (FEYRER et al., 2011); and to model depth gradient on pelagic bioluminescence in the northeast Atlantic Ocean (ZUUR et al., 2007). Nevertheless, there are no publications using GAMs in agronomic researches. The advantage of smoothing functions rather than detailed parametric relationships is the possibility to avoid the sort of cumbersome and unwieldy models. However, this flexibility and convenience come at the cost of two new theoretical problems: it is necessary to represent both the smooth functions in some way and choose how smooth the curve should be (WOOD, 2017).

Thin plate regression spline smooth was used to estimate the smoothing parameters in all models. This spline produces knot free bases, for smooths of any number of predictors and are a very general solution to the problem of estimating a smooth function of multiple predictor variables, from noisy observations of the function at particular values of those predictors (DUCHON, 1977). Moreover, this regression avoids the problem of knot placement, is relatively easy to compute, and can be constructed for smooths of any number of predictor variables (WOOD, 2017). The problem with thin plate splines is the computational cost, because these smoothers have as many unknown parameters as there are data (strictly, number of unique predictor combinations), and the computational cost of model estimation is proportional to the cube of the number of parameters (WOOD, 2013).

Correlation structures are used to model dependence among observations. For mixed-effect models and extended linear models, they are used to model dependence among the within-group errors. Correlation structures have been developed for two main classes of data: time-series and spatial data. The former is generally associated with observations indexed by an integer-valued time variable, while the latter refers primarily to observations indexed by a two-dimensional spatial location vector, taking values in the

real plane (PINHEIRO; BATES, 2000). The AR1 model is one of the few serial correlation structures that can be generalized for continuous time measurements.

Even if the samples presented a regular spaced time series, the data did not present a strong correlation over the weeks, and the use of the autoregressive model for correlation (AR1) was not necessary. The low changes at AIC values help to confirm this statement. The AIC statistics is recommended to compare models with the same family and log link function and gives precious information to the user, showing if the addition of a parameter or a variable helps the model or not. This criterion is widely used in GLMs and GAMs (LEE et al., 2006; FARAWAY, 2006). The addition of AR1 can even make the model worst, because of the addition of unnecessary parameters (BECK; KATZ, 1995; FRISTON et al., 2000). However, we strongly recommend that the researcher adjust the model with and without the correlation structure to then decide which is the best model.

For data with many zeros clustered together in the covariate space, it is easy to set up GAMs that suffer from identifiability problems, particularly when using Poisson or binomial families. The problem is that with log or logit links, mean value zero corresponds to an infinite range on the linear predictor scale (BÜRKNER, 2017). This problem was observed in the first and second model. Even with the low probability to obtain a false zero observed in the Bayesian model, these models have a problem with the variance estimation of the variables.

Although the package `mgcv` already has the Poisson zero inflated distribution added, the function still does not run zero inflated negative binomial data. There are five main packages in R available for modeling zero-inflated data: `pscl`, `INLA`, `MCMCglmm`, `glmmADMB`, and `brms`. The only package capable of running a GAM for zero inflated negative binomial data is `brms`, revealing the importance of using a Bayesian approach.

There are few literatures comparing the goodness of fit from abundance data with zero-inflated distributions. Warton (2005) found that the negative binomial was a good model for the number of zeros in counted abundance datasets, suggesting that a good approach to analyze such data will often be to use negative binomial log-linear models. In his study, he even concluded no significant differences between ZIBN and BN models, recommending that most zeros could be attributed to the systematic component of the model, rather than taking the more complicated route and incorporating them into the random component of the model. The conclusions of Yau et al. (2003) are quite different and detected how important the use of ZIBN is for overdispersed count data. The

divergences alerted that zero inflation should be used with caution and also compared with the original model to check goodness of fit.

The application of Bayesian statistics brought the use of ZIBN in the GAMs. The Bayesian approach uses cross validation, where part of the data is used to make inference and the other part to validate them, evaluating the performance of the decision (GHOSH et al., 2006). To simulate the data, Markov chain Monte Carlo (MCMC) methods offered schemes for drawing a series of correlated samples that will converge in a distribution closest to the target distribution (NEAL, 1993). When model parameters are continuous rather than discrete, Hamiltonian Monte Carlo algorithm (HMC) is able to suppress random walk behavior by means of a clever auxiliary variable scheme transforming the problem of sampling from a target distribution into the problem of simulating Hamiltonian dynamics (NEAL, 2011).

The problem with the use of HMC is that it is necessary to calculate the gradient of the log-posterior, which can be automated using algorithmic differentiation but is still a time-consuming process for more complex models. Thus, using HMC leads to higher quality samples but takes more time per sample than other algorithms typically applied, but this requirement can be made less onerous by using automatic differentiation (GRIEWANK; WALTHER, 2008).

Another problem is that HMC also requires that the user specify at least two parameters: a step size ε and a number of steps L to run a simulated Hamiltonian system. A poor choice of either of these parameters will result in a drop in HMC's efficiency. To avoid this problem, using an HMC extension, called no-U-turn sampler (NUTS), eliminates the need to choose the number of steps L and automatically tune the step size parameter ε (HOFFMAN; GELMAN, 2014).

Bayesian statistics is based on a different approach in statistics, because it assumes that prior information is available for the parameters, which is then combined with information contained in the data to get the posterior distribution. This is intrinsically different from frequentist statistics, where the data are analyzed in a stand-alone manner, independent from any other sources of information (ZUUR et al., 2009).

Although the use of prior information may seem a drawback, it can also be used as an advantage. For example, if the collected data has an annual basis, then the models can be updated annually where the results from previous years can form the prior distribution, which is then combined with the data from the current year. Furthermore, if a Bayesian analysis without prior information is required, then this can be achieved by

choosing non-informative prior distributions, which should give results similar to those obtained from maximum likelihood estimation (GELMAN et al., 2003).

The use of Bayesian methods has several advantages over frequentist approaches, including easy interpretation, incorporation of prior information, practical estimation of any function of parameters or predictive values and reduced small-sample bias compared to maximum likelihood procedures (GHOSH et al., 2006). However, their practical use was limited for a long time because more complex models could not be well fitted analytically. In the last few decades, this has changed with the development of new algorithms and the rapid increase of general computing power (BÜRKNER, 2017).

For some researchers, checking the goodness of fit from a Bayesian model could be a little difficult, because the statistics involved are also different from frequentist statistics. However, many techniques are available. The CDF, for example, is a great tool for goodness of fit, and is visually easy to interpret. The trace and density plots for the parameters, smooth terms and population-level effects also show how well the chains converged and the variation in the parameters' estimation. For comparing two Bayesian models with same distribution, leave-one-out cross-validation - LOO-CV (VEHTARI et al., 2017) and Watanabe-Akaike information criterion - WAIC (WATANABE, 2010) are methods for estimating pointwise out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values.

Also, Gelman-Rubin diagnostic is a good way to check goodness of fit. Gelman-Rubin diagnostic test compares the variation between chains to the variation within a chain. Initially, the value of the Gelman-Rubin statistic will be large, but when convergence has been reached, it will have decreased to a value close to one. Gelman (1996) suggests that a value lower than 1.1 or 1.2 is acceptable and should be applied to each of the model parameters. This diagnostic indicated that all model parameters were within an acceptable variance.

It was noted a high variance for the data close to zero precipitation, indicating that a variance structure should be used to model this high variance. However, package `mgcv` uses the function `glmmPQL` to fit other distributions than normal and `glmmPQL` does not fit models with different variance structures. Even with the advantages of the Bayesian approach, package `brms` is also not able to adjust a different variance structure for negative binomial distribution. Despite the statistical advances reached by *R* software, there is still some models that need to be developed in this environment.

5) CONCLUSIONS

Fitting a Generalized Additive Model to the aphid abundance data results in overdispersion, auto-correlation and zero inflation, making any inferences questionable. Extending the analysis to a negative binomial distribution helps solve the overdispersion problem. Auto-correlation is solved with ARMA structures and excess of zeros with zero inflation models. All these approaches are available in Bayesian models in *R*, also capable of inserting random effects and incorporating missing values. However, different variance structures are not yet available for ZINB models and need to be implemented in *R*.

The flexibility of Bayesian analysis comes at a cost in time-consuming simulation methods. Fortunately, freeware software is evolving and making the simulation even faster. The example of. *B. brassicae* incidence showed how well abiotic (and biotic) factors can be modeled and analyzed using GAMs.

6) REFERENCES

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. *In*: PETROV, B. N.; CSAKI, F. (ed.). **Second International Symposium on Information Theory**. Budapest: Akademai Kiado, 1973. p. 267–281.

ATKINS, D. C.; BALDWIN, S. A.; ZHENG, C.; GALLOP, R. J.; NEIGHBORS C. A tutorial on count regression and zero-altered count models for longitudinal substance use data. **Psychol Addict Behaviour**, [*s. l.*], v. 27, n. 1, p. 166–177, 2013. DOI: https://doi.org/10.1037/a0029508.

ATKINS, D. C.; GALLOP, R. J. Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. **Journal of Family Psychology**, [*s. l.*], v. 21, p. 726–735, 2007. DOI: https://doi.org/10.1037/0893-3200.21.4.726.

BECK, N.; KATZ, J. N. What to do (and not to do) with Time-Series Cross-Section Data. **American Political Science Review**, [*s. l.*], v. 89, n. 3, p. 634–647, 1995. DOI: https://doi.org/10.2307/2082979.

BÜRKNER, P. C. brms: An R package for Bayesian multilevel models using Stan**. Journal of Statistical Software**, [*s. l.*], v. 80, p. 1–28, 2017. DOI: https://doi.org/10.18637/jss.v080.i01.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time Series Analysis:** forecasting and control. 3. ed. San Francisco: Holden-Day, 1994. 500 p.

COXE, S.; WEST, S. G.; AIKEN, L. S. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. **Journal of Personality Assessment**, [*s. l.*], v. 91, p. 121–136, 2009. DOI: https://doi.org/10.1080/00223890802634175.

DUANE, S.; KENNEDY, A. D.; PENDLETON, B. J.; ROWETH, D. Hybrid Monte Carlo. **Physics Letters B**, [*s. l.*], v. 195, n. 2, p. 216–222, 1987. DOI: https://doi.org/10.1016/0370-2693(87)91197-X.

DUCHON, J. Splines minimizing rotation-invariant semi-norms in sobolev spaces. *In*: SCHEMPP, W.; ZELLER, K. (ed). **Construction theory of functions of several variables**. Berlin: Springer, 1977. p. 85–100. DOI: https://doi.org/10.1007/BFb0086566.

EVERAERT, G.; ESCHUTTER, Y.; TROCH M.; COLIN, R. J.; SCHAMPHELAERE K. Multimodel inference to quantify the relative importance of abiotic factors in the population dynamics of marine zooplankton. **Journal of Marine Systems**, [*s. l.*], v. 181, p. 91-98, 2018. DOI: https://doi.org/10.1016/j.jmarsys.2018.02.009.

FALK, M. G.; O'LEARY, R.; NAYAK, M.; COLLINS, P.; CHOY, S. L. A Bayesian hurdle model for analysis of an insect resistance monitoring database. **Environmental**

and Ecological Statistics, [*s. l.*], v. 22, p. 207–226, 2015. DOI: https://doi.org/10.1007/s10651-014-0294-3.

FARAWAY, J. J. **Linear models with R**. Florida: Chapman and Hall/CRC, 2005. 225 p.

FARAWAY, J. J. **Extending the linear model with R:** generalized linear, mixed effects and nonparametric regression models. Florida: Chapman and Hall, 2006. 301 p.

FEYRER, F.; NEWMAN, K.; NOBRIGA, M.; SOMMER, T. Modeling the Effects of Future Outflow on the Abiotic Habitat of an Imperiled Estuarine Fish. **Estuaries and Coasts**, [*s. l.*], v. 34, p. 120-128, 2011. DOI: https://doi.org/10.1007/s12237-010-9343-9.

FLETCHER, D.; MACKENZIE, D.; VILLOUTA, E. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. **Environmental and Ecological Statistics**, [*s. l.*], v. 12, p. 45–54, 2005. DOI: https://doi.org/10.1007/s10651-005-6817-1.

FRISON, K. J.; JOSEPHS, O.; ZARAHN, E.; HOLMES, A. P.; ROUQUETTE, S.; POLINE, J. B. To smooth or not to smooth? **NeuroImage**, [*s. l.*], v. 12, p. 196–208, 2000. DOI: https://doi.org/10.1006/nimg.2000.0609.

GAMERMAN, D.; LOPES, H. F. **Markov Chain Monte Carlo**: Stochastic Simulation for Bayesian Inference. 2. ed. New York: Chapman and Hall, 2006. 342 p.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian Data Analysis**. 2. ed. New York: Chapman and Hall, 2003. 668 p.

GELMAN, A. Inference and monitoring convergence. *In*: WILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. J. (ed.). **Markov chain Monte Carlo in practice**. London: Chapman and Hall, 1996. p.131–143.

GELMAN, A.; RUBIN, D. B. Inference from Iterative Simulation Using Multiple Sequences. **Statistical Science**, [*s. l.*], v. 7, p. 457–511, 1992. DOI: https://doi.org/10.1214/ss/1177011136.

GHOSH, S. K.; MUKHOPADHYAY, P.; LU, J. C. Bayesian analysis of zero-inflated regression models. **Journal of Statistical Planning and Inference**, [*s. l.*], v. 136, p. 1360–1375, 2006. DOI: https://doi.org/10.1016/j.jspi.2004.10.008.

GRIEWANK, A.; WALTHER, A. **Evaluating Derivatives**: principles and techniques of algorithmic differentiation. Society for Industrial and Applied Mathematics (SIAM). 2. ed. Philadelphia: SIAM, 2008. 438 p. DOI: https://doi.org/10.1137/1.9780898717761.

GUISAN, A.; ZIMMERMAN, N. E. Predictive habitat distribution models in ecology. **Ecological Modelling**, [*s. l.*], v. 135, p. 147–186, 2000. DOI: https://doi.org/10.1016/S0304-3800(00)00354-9.

HAMBY, K. A.; BELLAMY, D. A.; CHIU, J. C.; LEE, J. C.; WALTON, V. M.; WIMAN, N. G.; YORK, R. M.; BIONDI, A. Biotic and abiotic factors impacting development, behavior, phenology, and reproductive biology of *Drosophila suzukii*. **Journal of Pest Science**, [*s. l.*], v. 89, p. 605-619, 2016. DOI: https://doi.org/10.1007/s10340-016-0756-5.

HILBE, J. M. **Negative binomial regression**. 2. ed. New York: Cambridge University Press, 2011. 553 p. DOI: http://dx.doi.org/10.1017/CBO9780511973420.

HOFFMAN, M.D.; GELMAN, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. **Journal of Machine Learning Research**, London, v. 15, n. 1, p. 1593–1623, 2014.

HUGHES, R. D.; GILBERT, N. A Model of an Aphid Population-A General Statement. **Journal of Animal Ecology**, [*s. l.*], v. 37, n. 3, p. 553–563, 1968. DOI: https://doi.org/10.2307/3074.

ISHWARAN, H.; RAO, J. S. Spike and slab variable selection: frequentist and Bayesian strategies. **Annals of Statistics**, [*s. l.*], v. 33, n. 2, p. 730–773, 2005. DOI: https://doi.org/10.1214/009053604000001147.

JAN, H.; AKHTAR, M. N.; AKHTAR, Z. R.; NAVEED, W. A.; LATIF, M.; SHAH, S. Z. A. Effect of biotic and abiotic factors on the population dynamics of wheat aphids. **Journal of Entomology and Zoology studies**, [*s. l.*], v. 5, n. 6, p. 2349–2353, 2017.

KIM, Y. J.; GU, C. Smoothing spline gaussian regression: more scalable computation via efficient approximation. **Journal of the Royal Statistical Society**, London, v. 66, p. 337–356, 2004. DOI: https://doi.org/10.1046/j.1369-7412.2003.05316.x.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. **Generalized Linear Models with Random Effects**. New York: Chapman and Hall, 2006. 380 p. DOI: https://doi.org/10.1201/9781420011340.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. **Journal of Chemical Physics**, [*s. l.*], v. 21, p. 1087–1092, 1953. DOI: https://doi.org/10.1063/1.1699114.

NEAL, D. J.; SIMONS, J. S. Inference in regression models of heavily skewed alcohol use data: A comparison of ordinary least squares, generalized linear models, and bootstrap resampling. **Psychology of Addictive Behaviors**, Washington, v. 21, p. 441–452, 2007. DOI: https://doi.org/10.1037/0893-164X.21.4.441.

NEAL, R. M. **Bayesian Learning for Neural Networks**. New York: Springer, 1996. 187 p. DOI: https://doi.org/10.1007/978-1-4612-0745-0.

NEAL, R. M. MCMC Using Hamiltonian Dynamics. *In*: BROOKS, S.; GELMAN, A.; JONES, G. L.; MENG, X. (ed.). **Handbook of Markov Chain Monte Carlo**. Boston: CRC Press, 2011. p. 113-162. DOI: https://doi.org/10.1201/b10905-6.

NEAL, R. M. **Probabilistic inference using Markov chain Monte Carlo methods**. Technical Report CRG-TR-93-1, Toronto: University of Toronto, 1993. 140 p.

NEAL, R. M. Slice Sampling. **The Annals of Statistics**, [*s. l.*], v. 31, n. 3, p. 705–741, 2003. DOI: http://dx.doi.org/10.1214/aos/1056562461.

PARMESAN, C. Influences of species, latitudes, and methodologies on estimates of phonological response to global warming. **Global Change Biology**, [*s. l.*], v. 13, p. 1860–1872, 2007. DOI: https://doi.org/10.1111/j.1365-2486.2007.01404.x.

PINHEIRO, J. C.; BATES, D. M. **Mixed-Effects Models in S and S-PLUS**. New York: Springer, 2000. DOI: https://doi.org/10.1007/b98882. 530p.

PRICE, P. W.; DENNO, R. F.; EUBANKS, M. D.; FINKE, D. L.; KAPLAN, I. **Insect ecology behavior, populations and communities**. United Kingdom: Cambridge, 2011. 791 p. DOI: https://doi.org/10.1017/CBO9780511975387.

RHODES, J. R. Mixture models for overdispersed data. *In:* FOX, G. A.; NEGRETE-YANKELEVICH, S.; SOSA, V. J. (ed.). **Ecological Statistics**. Oxford: Oxford University Press, 2015. 414 p. DOI: https://doi.org/10.1093/acprof:oso/9780199672547.003.0013.

SAMPAIO, M. V.; KORNDÖRFER, A. P.; PUJADE-VILLAR, J.; HUBAIDE, J. E. A.; FERREIRA, S.E.; ARANTES, S. O.; BORTOLETTO, D. M.; GUIMARÃES, C. M.; SÁNCHEZ-ESPIGARES, J.A.; CABALLERO-LÓPEZ, B. Brassica aphid (Hemiptera: Aphididae) populations are conditioned by climatic variables and parasitism level: a study case of Triângulo Mineiro, Brazil. **Bulletin of Entomological Research**, Cambridge, v. 107, p. 410–418, 2017. DOI: https://doi.org/10.1017/S0007485317000220.

STROUP, W. W. **Generalized Linear Mixed Models:** Modern Concepts, Methods and Applications. 1. ed. Boca Raton: CRC Press, 2012. 555 p.

STROUP, W. W. Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. **Agronomy Journal**, Madison, v. 107, n. 2, p. 811-827, 2015. DOI: https://doi.org/10.2134/agronj2013.0342.

THACKRAY, D. J.; DIGGLE, A. J.; BERLANDIER, F. A.; JONES, R. A. C. Forecasting aphid outbreaks and epidemics of Cucumber mosaic virus in lupin crops in a Mediterranean type environment. **Virus Research**, [*s. l.*], v. 100, p. 67–82, 2004. DOI: https://doi.org/10.1016/j.virusres.2003.12.015.

VEHTARI, A.; GELMAN, A.; GABRY, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. **Statistics and Computing**, Switzerland, v. 27, n. 5, p. 1413–1432. DOI: https://doi.org/10.1007/s11222-016-9696-4.

WARTON, D. I. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. **Environmetrics**, [*s. l.*], v. 16, p. 275–289, 2005. DOI: https://doi.org/10.1002/env.702

WATANABE, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of Machine Learning Research**, [*s. l.*], v. 11, p. 3571–3594, 2010.

WELSH, A. H.; CUNNINGHAM R. B.; CHAMBERS, R. L. Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay. **Biometrics**, [*s. l.*], v. 56, p. 22–30, 2000. DOI: https://doi.org/10.1111/j.0006-341X.2000.00022.x.

WHITNEY, S. K.; MEEHAN, T. D.; KUCHARIK, C. J.; ZHU, J.; TOWNSEND, P. A.; HAMILTON, K.; GRATTON, C. Explicit modeling of abiotic and landscape factors reveals precipitation and forests associated with aphid abundance. **Ecological Applications**, [*s. l.*], v. 26, n. 8, p. 2600-2610, 2016. DOI: https://doi.org/10.1002/eap.1418.

WILSON, L. T.; BARNETT, W. W. Degree days: an aid in crop and pest management. **California Agriculture**, California, v. 37, p. 4-7, 1983.

WINDER, L. Predation of the cereal aphid *Sitobion avenae* by polyphagous predators on the ground. **Ecological Entomology**, [*s. l.*], v. 15, p. 105–110, 1990. DOI: https://doi.org/10.1111/j.1365-2311.1990.tb00789.x.

WOOD, S. Package 'mgcv'. **R package version**, v. 1.8-27, [*s. l.*], 6 feb. 2019. Available at: https://cran.r-project.org/web/packages/mgcv/mgcv.pdf. Accessed in: 18 mar. 2019.

WOOD, S. N. **Generalized Additive Models**: An Introduction with R. 2. ed. New York: Chapman and Hall/CRC, 2017. 476 p. DOI: https://doi.org/10.1201/9781315370279.

WOOD, S. N. Thin plate regression splines. **Journal of the Royal Statistical Society**, [*s. l.*], v. 65, n. 1, p. 95–114, 2003. DOI: https://doi.org/10.1111/1467-9868.00374.

YAU, K. K. W.; WANG, K.; LEE, A. H. Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. **Biometric Journal**, Weinheim, v. 45, n. 4, p. 437–452, 2003. DOI: https://doi.org/10.1002/bimj.200390024.

ZUUR, A. F.; IENO, E. N.; SMITH, G. M. **Analysing Ecological Data**. New York: Springer. 2007. 680 p. DOI: https://doi.org/10.1007/978-0-387-45972-1.

ZUUR, A. F.; IENO, E. N.; WALKER, N. J.; SAVELIEV, A. A.; SMITH, G. M. **Mixed Effects Models and Extensions in Ecology with R.** New York: Springer, 2009. 574 p. DOI: https://doi.org/10.1007/978-0-387-87458-6.

7) SUPPLEMENTARY MATERIAL

Appendix A. Summary of a Generalized Additive Model (GAM) to the number of *B. brassicae* aphids, with a negative binomial distribution with a log link function, using the function `gam` from the package `mgcv`.

```
Family: Negative Binomial(0.237)
Link function: log

Formula:
aphid ~ s(precipitation, k=9) + s(temperature, k=9) + s(time,k=15),family = nb,
gamma=1.4

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.8973     0.1466   26.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df  Chi.sq p-value
s(prec)  1.00   1.00   5.096   0.024 *
s(medt)  1.00   1.00   1.871   0.171
s(time) 12.32  13.53 216.912  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.374   Deviance explained =   60%
REML = 686.58  Scale est. = 1         n = 207

AIC 1868.561

> gam.check(BM4)

Method: REML   Optimizer: outer newton
full convergence after 9 iterations.
Gradient range [-9.731991e-05,0.0001344015]
(score 686.5787 & scale 1).
Hessian positive definite, eigenvalue range [7.404829e-05,61.86547].
Model rank =  31 / 31

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

          k'  edf k-index p-value
s(prec)  8.0  1.0    0.67    0.25
s(medt)  8.0  1.0    0.70    0.54
s(time) 14.0 12.3    0.65    0.13

> gam.vcomp(BM4, rescale = FALSE)

Standard deviations and 0.95 confidence intervals:

            std.dev        lower        upper
s(prec)  0.01582724 1.131424e-45 2.214037e+41
s(medt)  0.01247137 4.334833e-52 3.588029e+47
s(time) 75.10447732 4.167963e+01 1.353343e+02

Rank: 3/3
```

Appendix B. Summary of a Generalized Additive Model (GAM) to the number of *B. brassicae* aphids, with a negative binomial distribution with a log link function, with an autocorrelation structure for time, using the function `gam` from the package `mgcv`.

```
Family: Negative Binomial(0.237)
Link function: log

Formula:
aphid ~ s(precipitation, k=9) + s(temperature, k=9) + s(time,k=15),family = nb,
gamma=1.4

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.8957     0.1467   26.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df  Chi.sq p-value
s(prec)  1.00   1.00   5.095   0.024 *
s(medt)  1.00   1.00   1.860   0.173
s(time) 12.34  13.54 217.178  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.368   Deviance explained = 60.1%
REML =  687.1  Scale est. = 1         n = 207

> AIC  1865.428

Method: REML   Optimizer: outer newton
step failed after 10 iterations.
Gradient range [-0.0002898318,-9.688577e-05]
(score 687.1004 & scale 1).
Hessian positive definite, eigenvalue range [9.687382e-05,3.768404].
Model rank =  31 / 31

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

          k'  edf k-index p-value
s(prec)  8.0  1.0    0.67    0.22
s(medt)  8.0  1.0    0.70    0.51
s(time) 14.0 12.3    0.66    0.15

> gam.vcomp(MOD2, rescale = FALSE)
     s(prec)      s(medt)      s(time)
  0.02003406   0.03064407 128.99340685
```

Appendix C. Summary of a Bayesian Generalized Additive Model (BGAM) to the number of *B. brassicae* aphids, with a zero-inflation negative binomial (ZINB) model with a log link function and a binomial model with logistic link for the model of false zeros versus the non-false zeros, with an autocorrelation structure for time, using the function `brms` from the package `mgcv`.

```
Family: zero_inflated_negbinomial
  Links: mu = log; shape = identity; zi = identity
Formula: bb ~ s(prec, k = 9) + s(medt, k = 9) + s(time, k = 15)
   Data: data (Number of observations: 207)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000


Smooth Terms:
             Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sds(sprec_1)     2.72      3.11     0.07    11.52       1725 1.00
sds(smedt_1)     1.79      1.88     0.04     7.05       1726 1.00
sds(stime_1)    87.14     24.04    47.55   142.79        880 1.00

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept     4.08      0.18     3.75     4.44       2175 1.00
sprec_1       0.92      0.95    -0.46     3.42       1417 1.00
smedt_1       0.33      0.66    -0.99     1.63       2313 1.00
stime_1     -12.26      5.08   -22.31    -2.74       2376 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
shape     0.26      0.03     0.20     0.33       4000 1.00
zi        0.03      0.03     0.00     0.11       4000 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
>
```

Appendix D. Trace and density plots for Population-Level Effects of intercept, precipitation, temperature and time from the Bayesian Generalized Additive Model (BGAM) to the number of *B. brassicae* aphids.