UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Tiago Costa Rocha

Comparação de Modelos Lineares Generalizados e Modelos Generalizados Aditivos de Locação, Escala e Forma Aplicados a Micropropagação do Abacaxizeiro.

Campina Grande - PB

Dezembro 2019

Tiago Costa Rocha

Comparação de Modelos Lineares Generalizados e Modelos Generalizados Aditivos de Locação, Escala e Forma Aplicados a Micropropagação do Abacaxizeiro.

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira

Coorientador: Profa. Dra. Fabiane de Lima Silva.

Campina Grande - PB

Dezembro 2019

R672c Rocha, Tiago Costa.

Comparação de Modelos Lineares Generalizados e Modelos Generalizados Aditivos de locação, escala e forma aplicados a micropropagação do abacaxizeiro [manuscrito] / Tiago Costa Rocha. - 2019.

34 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2019.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira , Departamento de Estatística - CCT."

"Coorientação: Profa. Dra. Fabiane Lima Silva , UFMA - Universidade Federal do Maranhão"

1. Modelos Lineares Generalizados. 2. Modelos Generalizados Aditivos. 3. Análise de Variância. 4. Micro propagação. I. Título

21. ed. CDD 519.5

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

Tiago Costa Rocha

COMPARAÇÃO DE MODELOS LINEARES GENERALIZADOS E MODELOS GENERALIZADOS ADITIVOS DE LOCAÇÃO, ESCALA E FORMA APLICADOS A MICROPROPAGAÇÃO DO ABACAXIZEIRO.

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 05 de Dezembro de 2019.

BANCA EXAMINADORA

Prof. Dr. Tiago Almeida de Oliveira(Orientador)
Universidade Estadual da Paraíba

Profa. Dra. Érika Fialho Morais Xavier
Universidade Estadual da Paraíba

Prof. Dr. Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Dedicatória

"Dedico esse trabalho aos meus pais Josélio e Telma, que todos os dias me deram forças para superar as dificuldades e persistir no sonho de concluir uma graduação. A minha noiva Ceciliane, a minha irmã Tamires, e a meu irmão Heitor. Obrigado a todos que me apoiaram durante toda essa jornada, por fazerem o possível e o impossível por mim, com todo amor do mundo".

Epígrafe

"Tentar é uma opção,
enfrentar é um desafio,
concluir é uma superação"
(Tiago Costa)

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

GLM	Modelo Linear Generalizado.
GAM	Modelo Aditivo Generalizado.
GAMLSS	Modelo Linear Generalizado de Locação, Escala e Forma.
ML	Modelo Linear.
ANOVA	Análise de Variância.

Sumário

COMPARAÇÃO DE MODELOS LINEARES GENERALIZADOS E MODELOS GENERALIZADOS ADITIVOS DE LOCAÇÃO, ESCALA E FORMA APLICADOS A MICROPROPAGAÇÃO DO ABACAXIZEIRO.

Tiago Costa Rocha[1]
Fabiane Lima Silva[2]
Tiago Almeida de Oliveira [3]

RESUMO

O abacaxi é bastante comercial, sendo cultivado em vários locais no território nacional. O uso de técnicas convencionais, tais como a ANOVA(Análise de Variância) e teste de Tukey, tem demonstrado que para dados de micropropagação devido a super dispersão, não tem encontrado resultados satisfatórios estatisticamente. Os resíduos viabilizam extrair informações que permitem diagnosticar ou demonstrar a qualidade do ajuste de um modelo (estágio de diagnóstico), além de averiguar se as suposições foram satisfeitas. A partir dai utilizou-se plantas matrizes dos cultivares Imperial e *Smooth Cayenne*, sendo quatro plantas de cada variedade, obtidas no campo experimental da Embrapa Mandioca e Fruticultura. Foram estabelecidos três intervalos diferentes de subcultivos, 30, 45 e 60 dias para as duas variedades quando o número de brotos formados a cada repicagem foi contabilizado. Deste modo, como tratava-se de dados de contagem, para modelar o número de brotos de abacaxi, utilizamos as técnicas clássicas de modelagem (Modelo Linear e Linear Generalizado). No decorrer da análise, percebeu-se que estes modelos não foram adequados devido as técnicas diagnósticas utilizadas (*hnp plot*). Neste sentido, partiu-se para modelar o efeito de superdispersão pelos modelos Quasi-Poisson e Binomial Negativo, no entanto estes também não foram adequados. A partir daí, ajustou-se modelos mais flexíveis que não precisassem necessariamente pertencer a família exponencial. Utilizamos então a classe de modelos GAMLSS, que forneceu as melhores estimativas com o modelo Binomial do tipo I (GAMLSS) de acordo com as técnicas diagnósticas ao utilizar o subcultivo como efeito ajustado no parâmetro de dispersão. Por meio desta modelagem foi possível identificar pelo teste de Tukey que o subcultivo 6, localizado no cultivar 1 contendo a variedade Imperial, apresentou maior número de brotos.

Palavras-chaves: GAMLSS. ANOVA. Micro propagação. Subcultivo.

[1] Discente do Departamento de Estatística - thyagoestatistica@gmail.com
[2] Docente do UFRB - fabianezte@gmail.com
[3] Docente do Departamento de Estatística - tiagoestatistico@gmail.com

ABSTRACT

Pineapple is very commercial, being grown in several places in the national territory. The use of conventional techniques such as ANOVA and tukey test has shown that for micropopagation data, due to overdispersion, has not found good results. The residues make it possible to extract information that allows diagnosing or demonstrating the fit quality of a model (diagnostic stage), as well as verifying if the assumptions of the model were met. Parent plants of the cultivars Imperial and Smooth Cayenne were used, being four plants of each variety, obtained in the experimental field of Embrapa Cassava and Fruit. Three different subcultive intervals were established, 30 days, 45 days and 60 days for the two varieties when the number of shoots formed at each subcultive was accounted. Thus, as the database was counting data, to model the number of pineapple shoots, we used the classical modeling techniques (Generalized Linear and Linear Model), during the analysis, it was noticed that these models were not adequate due to the diagnostic techniques used (hnp plot), in this sense, it was started to model the effect of overdispersion by the Quasi-Poisson and Negative Binomial models, however these were not adequate either. From there, we adjusted more flexible models that did not necessarily need to belong to the exponential family, so we started to class GAMLSS models. Having the best estimates with the Binomial Type I (GAMLSS) model, knowing that it is suitable according to the diagnostic techniques, when using the subcultive as an adjusted effect on the dispersion parameter. Through this modeling it was possible to identify by the tukey test that the subculture 6 located in the cultivar Imperial presented higher number of shoots.

Key-words: GAMLSS. ANOVA. Micro propagation. Subcultive.

1 Introdução

O abacaxi é muito apreciado por apresentar características de sabor e aroma, sendo cultivado em vários locais no território nacional, com a produção global em torno dos 25,81 milhões de toneladas métricas em uma área de plantio com 1,04 milhão de hectares em 2014 (FAOSTAT, 2017). Segundo Cabral, Souza e Ferreira (1999), aqui no Brasil existe uma vasta diversidade de espécies do gênero Ananás, o que torna um importante ponto de origem nesta categoria, o território nacional abriga estas espécies tanto nas formas silvestresPlantio indireto, produzido pela própria natureza quanto na forma cultivadaPlantio direto, produzido através de mudas feitas a partir de uma planta matriz.

Com a procura pelo produto, a motivação para aumentar a produção se eleva, no entanto, para suprir essa demanda o investimento em tecnologia também cresce, assim as

técnicas de micropropagação (Consiste na produção de vários clones a partir de pequenos fragmentos extraídos de uma planta matriz) tentam buscar a melhoria na produção de variedades com uma taxa satisfatória de propagação para o desenvolvimento do fruto, mas isto requer uma série de cuidados, pois a propagação esta sujeita a doenças e pragas. O investimento em técnicas estatísticas que permitam o correto posicionamento para um melhor entendimento do processo produtivo é de extrema importância.

Existem algumas técnicas para chegar na escala comercial da micropropagação, o que gerou grande parte dos avanços biotecnológicos proporcionando impacto na agricultura, com aumento nos últimos 30 anos. No quesito propagação do genótipo que revolucione a produção, com o intuito de gerar o maior número de mudas em um espaço de tempo cada vez menor, sobrepondo a qualidade já existente da muda convencional, está técnica cresceu gradativamente no cultivo de várias espécies de reprodução vegetativa, como é o caso do abacaxizeiro (GUERRA et al., 1999; BE; DEBERGH, 2006). Uma questão que vale ressaltar sobre a micropropagação se refere aos subcultivos (Ou repicagens, se referem ao transplante de mudas provenientes das sementes), pelo fato da quantidade destes subcultivos não estabelecerem diferenças significativas, pois o ponto chave para gerar maior produção de mudanças ocorre devido o intervalo em dias entre ambos, onde podem interferir nos resultados obtidos no término do estudo (HAMAD; TAHA, 2008).

Segundo Souza (2018) uso de técnicas convencionais tais como a ANOVA (Análise de Variância) e teste de Tukey, tem demonstrado que para dados de micropropagação, devido a superdispersão, não tem encontrado resultados satisfatórios estatisticamente. Deste modo, a modelagem utilizada não foi adequada, por consequência tornou-se inadequada para predizer bem os resultados das taxas de multiplicação para plantas *in vitro* (Plantas que são isoladas em ambiente controlado e fechado de um laboratório que são feitos normalmente em recipientes de vidro), pelo simples fato de não traduzir com precisão o comportamento das relações biológicas existentes, o que acarreta dificuldade de compreender e interpretar de forma fidedigna os resultados obtidos.

Neste sentido, generalizações da modelagem permitem a inclusão de distribuições que pertençam a família exponencial, tais como modelos lineares generalizados GLM (NELDER; WEDDERBURN, 1972) e/ou de Técnicas de suavização que começaram a ficar populares nos anos 1980. Assim, Hastie e Tibshirani (1987) introduziram dentro da abordagem das classes de modelos GLM os parâmetros de suavização talhando o termo GAM (Generalized Additive Model). Nos anos 2000 o modelo GAM se popularizou com a implementação no *software R* por meio do pacote `mcg` (WOOD, 2006). Porém, nem todas as limitações foram resolvidas com advento dos modelos GLM e GAM. Pensando nisso, Rigby e Stasinopoulos (2005) propuseram a classe de modelos de regressão denominada GAMLSS - Modelos Aditivos Generalizados para Locação, Escala e Forma com a vantagem de modelar os parâmetros de assimetria e curtose, possibilitando uma maior flexibilidade

e uma expansão ao número de distribuições consideradas, pois agora as distribuições não precisam ser parte da família exponencial.

As técnicas estatísticas estão cada vez mais interligadas com as mais diversas áreas do conhecimento, não poderia ser diferente com agricultura, o ferramental estatístico adequado é de suma importância na tomada de decisões. A modelagem adequada ao dados possibilita a obtenção de conclusões fidedignas, no entanto, os critérios de seleção do modelo que se adeque aos dados deve obedecer todas as pré suposições para que seja correto.

Desta forma, esse trabalho teve por objetivo avaliar o efeito de três diferentes intervalos de sub cultivos (dias), tratando do potencial propagativo de variedades comerciais dos abacaxizeiros sendo elas Imperial e Smooth Cayenne, escolhendo a partir da classe de modelos linear generalizada, com o ajuste dos modelos gerados pela distribuição Binomial Negativa, Poisson, quasipoisson, Normal e por meio da classe GAMLSS via binomial negativa do tipo I e verificar qual entre as distribuições consiga modelar os efeitos de superdispersão presente nos dados.

Com base nesses preceitos, a classe de modelos GAMLSS apresentou resultados satisfatórios modelando bem os efeitos de superdispersão que evidenciaram a diferença na produção do número de brotos (Parte visível da planta em estágio inicial de desenvolvimento do caule e folhas) quando comparado duas variedades de abacaxizeiro, resultado que não foi obtido em estudos anteriores utilizando outro tipo de modelagem, contribuindo de forma significativa na otimização dos protocolos de micropropagação, podendo gerar maior produção por hectare, tornando viável o uso da micropropagação em escala comercial.

Poucos estudos foram feitos nessa área utilizando a modelagem GAMLSS, comumente como esse tipo de dados tem a característica de contagem a modelagem mais fácil de se trabalhar é a Poisson, que geralmente é utilizada, porém este tipo de modelagem não gera resultados satisfatórios quando se tem o efeito de superdispersão, a partir disso, optamos por uma classe de modelos que não pertencesse a família exponencial, podendo ter mais flexibilidade ao modelar o parâmetro de assimetria e curtose, diminuindo de forma significativa o erro, apresentando resultados que comprovam a qualidade superior da modelagem utilizada, quando comparada as técnicas clássicas.

2 Metodologia

Para alguns modelos básicos de regressão em dados de contagem, sendo eles, Poisson, Quasipoisson, Binomial Negativo e Binomial Negativo do tipo I é feita a descrição da modelagem, teoria e sua implementação em R Team (2018). Os modelos Poisson e binomiais são descritos em uma estrutura de modelo linear generalizado (GLM); eles são implementados em R pela função `glm` no pacote `stats` (CHAMBERS; HASTIE et al.,

1992) a função `glm.nb` é do pacote `MASS` para o modelo binomial negativo(VENABLES; RIPLEY, 2002). E a função `gamlss()` é do pacote `gamlss` (RIGBY; STASINOPOULOS, 2005).

2.1 Modelos Lineares Generalizados

Os modelos básicos de regressão aplicados a dados de contagem podem ser representados e entendidos usando o MLG, estrutura que surgiu na literatura estatística no início dos anos 70 (NELDER; WEDDERBURN, 1972). Temos em sequência alguns aspectos importantes relacionados à unificação de propriedades conceituais e sua implementação em R - para uma descrição teórica detalhada MLGs solicitar McCullagh e Nelder (1989). Os MLGs descrevem a dependência de uma variável escalar dada por: $Y_i(i = 1, ..., n)$, em um vetor de regressores x_i . A distribuição condicional de $y_i|x_i$ é uma família exponencial linear com densidade de probabilidade dada pela seguinte função:

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

Em que λ é o parâmetro canônico que depende dos regressores por meio de um preditor linear, e ϕ é um parâmetro de dispersão geralmente conhecido. As funções b $(\cdot)$ e c $(\ \cdot)$ são conhecidas e determinam qual membro da família é usado, por exemplo, a distribuição normal, binomial ou Poisson. Média condicional e variância de y_i são dadas por $E[y_i|X_i] = \mu_i = b"(\lambda_i)$ e $VAR[y_i|X_i] = \phi b"(\lambda_i)$. Definindo o parâmetro de dispersão ϕ, já a distribuição de y_i é determinada por sua média. já variância é proporcional a $V(\mu) = b"(\lambda(\mu))$, também chamada de função variação. A dependência da média condicional é $E[y_i|X_i] = \mu_i$ nos regressores x_i é especificado via função dada por:

$$g(\mu_i) = x_i^t\beta, \quad (2)$$

onde g $(\cdot)$ é uma função de ligação conhecida e β é o vetor de coeficientes de regressão que são normalmente estimados por máxima verossimilhança (ML) usando os algoritmos de mínimos quadrados iterativos (IWLS). Além de visualizar os MLGs como modelos para a probabilidade total (conforme determinado pela Equação 2.1), eles também podem ser considerados modelos de regressão apenas para a média (conforme especificado na Equação 2.2), em que as funções de estimativa usadas para ajustar o modelo são derivadas de uma família específica.

2.1.1 Modelo de Poisson

A distribuição mais simples usada para modelar dados de contagem é a distribuição de Poisson com função de densidade de probabilidade dada por:

$$f(y; \mu) = \left(\frac{\exp(-\mu).\mu^y}{y!}\right),$$

portanto, a regressão de Poisson é um caso especial da estrutura MLG. A ligação canônica é $g(\mu) = log(\mu)$ resultando em uma relação log-linear entre média e preditor linear. A variação no modelo de Poisson é idêntica à média, portanto a dispersão é fixada em $\phi = 1$ e a função de variação é $V(\mu) = \mu$.

No R, isso pode ser facilmente especificado na chamada `glm ()` apenas configurando $family = poisson$ (em que a ligação é, por padrão, a log e pode ser alterada na chamada da distribuição poisson $(\cdot)$). Na prática, o modelo de Poisson é útil para descrever a média μ_i, mas subestima a variação nos dados, tornando todos os testes baseados em modelos liberais (ZEILEIS; KLEIBER; JACKMAN, 2008).

2.2 Modelo de Quasi-Poisson

Outra maneira de lidar com a dispersão excessiva é usar a função de regressão média e a função de variância do modelo Poisson. Portanto, ϕ não é assumido como fixo em 1, mas é estimado a partir dos dados. Essa estratégia leva com as mesmas estimativas de coeficiente que o modelo padrão de Poisson, mas a inferência é ajustada para excesso de dispersão. Consequentemente, ambos modelos Quasi-Poisson e Poisson adotam a visão da função de estimativa do modelo de Poisson e não correspondem aos modelos com probabilidades totalmente especificadas. Em R, o modelo Quasi-Poisson com parâmetro de dispersão estimado também pode ser ajustado com a função `glm()`, simplesmente configurando *family = quasipoisson* (ZEILEIS; KLEIBER; JACKMAN, 2008).

2.3 Modelo Binomial Negativo

Uma terceira maneira de modelar dados de contagem com alta variabilidade é assumir a distribuição Binomial Negativa para $y_i|x_i$ que pode surgir como uma mistura de distribuições de Poisson. Uma parametrização de sua função de densidade de probabilidade é dada por:

$$f(y; \mu, \theta) = \frac{(\Gamma(y + \theta))}{(\Gamma(\theta)y!)} \frac{(\mu^y \theta^\theta)}{(\mu + \theta)^{(y+\theta)}},$$

com μ médio e parâmetro de forma θ, já γ $(\cdot)$ é a função gama. Cada θ fixo portanto, é um caso especial da estrutura GLM. Este modelo também possui $\phi = 1$, mas com função de variação dada por:

$$V(\mu) = \mu + \frac{\mu^2}{\theta}.$$

O pacote MASS (VENABLES; RIPLEY, 2002) fornece a função de família binomial negativa que pode ser conectadas diretamente ao glm, desde que o argumento θ seja especificado. Uma aplicação seria o modelo geométrico, o caso especial em que $\theta = 1$, que pode consequentemente ser ajustado no R definindo a família = binomial negativa ($\theta = 1$)

na função glm. Se θ não é conhecido, deve ser estimado a partir dos dados, o modelo binomial negativo não é um caso especial do GLM geral. No entanto, um ajuste de Modelo linear (ML) pode ser facilmente calculado reutilizando a metodologia GLM pela estimativa iterativa de β dado θ e vice-versa. Isso leva a estimativas de ML para β e θ que podem ser calculadas usando a função glm.nb do pacote MASS (ZEILEIS; KLEIBER; JACKMAN, 2008).

2.4 Modelos Aditivos Generalizados para Locação, Escala e Forma (GAMLSS)

Admitindo no ajuste de modelos que a distribuição da variável resposta não necessariamente segue uma distribuição pertencente à família exponencial, propôs-se a substituição por uma família de distribuição geral. No GAMLSS, a parte sistemática do modelo é expandida para permitir modelar não apenas a média, mas todos os parâmetros de locação, escala e forma da distribuição da variável resposta.

Desta forma, que todos esses parâmetros permitem ser modelados em função das variáveis explicativas e, ainda, os preditores também podem inserir funções não-paramétricas de suavização, efeitos aleatórios, ou outros termos aditivos, ou seja, os GAMLSS são modelos de regressão semi-paramétricos. Paramétricos, por requererem a suposição de distribuição paramétrica para a variável resposta, e semi-paramétricos no sentido de que permitem a modelagem dos parâmetros da distribuição como funções de variáveis explicativas por meio de funções de suavização não-paramétricas. Esta nova categoria de modelos, os GAMLSS, dispõe dos modelos MRL, GLM e GAM como casos especiais. Entretanto, estes modelos ainda adotam que as observações da variável resposta como independentes entre si.

Um modelo GAMLSS é descrito por,

$$Y \overset{ind}{\sim} D\left(\mu, \sigma, \nu, \tau\right)$$

$$\begin{aligned}
\eta_1 &= g_1\left(\mu\right) = X_1\beta_1 + s_{11}\left(x_{11}\right) + ... + s_{1J_1}\left(x_{1J_1}\right) \\
\eta_2 &= g_2\left(\sigma\right) = X_2\beta_2 + s_{21}\left(x_{21}\right) + ... + s_{2J_2}\left(x_{2J_2}\right) \\
\eta_3 &= g_3\left(\nu\right) = X_3\beta_3 + s_{31}\left(x_{31}\right) + ... + s_{3J_3}\left(x_{3J_3}\right) \\
\eta_4 &= g_4\left(\tau\right) = X_4\beta_4 + s_{41}\left(x_{41}\right) + ... + s_{4J_4}\left(x_{4J_4}\right),
\end{aligned} \tag{3}$$

em que, D $\left(\mu, \sigma, \nu, \tau\right)$ é uma distribuição geral de quatro parâmetros, onde o parâmetro de locação é μ, escala σ, ν e τ são os parâmetros de forma da distribuição, comumente associados à assimetria e curtose, respectivamente. X_1, X_2, X_3 e X_4 são as matrizes que podem ou não coincidir, ou melhor, o preditor de cada parâmetro da distribuição recebe diferentes variáveis explicativas (RIGBY; STASINOPOULOS, 2005).

Na função (3) que descreve o modelo GAMLSS com a máxima quantidade de parâmetros implementados no software R (PINHEIRO et al., 2018), que de modo teórico,

ainda é possível admitir uma distribuição de probabilidade para a variável resposta com mais de quatro parâmetros, inserindo mais preditores no modelo. Além disso, nos permite definir modelos intermediários para distribuições com dois ou três parâmetros e ajustar σ e/ou ν em função de covariáveis, considerando diminuir a quantidade de preditores. Se no modelo (2.3) não existem os efeitos aleatórios, logo o modelo resume-se à

$$Y \overset{ind}{\sim} D(\mu, \sigma, \nu, \tau)$$

$$\begin{aligned} \eta_1 &= g_1(\mu) = X_1\beta_1 \\ \eta_2 &= g_2(\sigma) = X_2\beta_2 \\ \eta_3 &= g_3(\nu) = X_3\beta_3 \\ \eta_4 &= g_4(\tau) = X_4\beta_4, \end{aligned} \tag{4}$$

Stasinopoulos et al. (2017) apontam as equações (4) como modelo GAMLSS paramétrico e as equações (3) como modelo GAMLSS com efeito aleatório. Quando um GAMLSS é paramétrico, isto é, que não existem funções de suavizações (efeitos aleatórios determinados anteriormente), exige só a estimação de β. Tratando-se de um modelo GAMLSS com efeitos aleatórios, exige-se não apenas o β, mas γ e λ determinado por,

$$\lambda = (\lambda_{11}^T, ..., \lambda_{1J_1}^T, \lambda_{21}^T, ..., \lambda_{4J_4}^T)^T.$$

O GAMLSS paramétrico descritos nas equações (4) pode ser entendido também no contexto dos modelos de delineamentos experimentais (LIU et al., 2015; FERRARA; VIDOLI, 2017), no qual modela-se o parâmetro de locação, porém com o ganho de poder modelar o parâmetro de escala, assimetria e curtose. Os modelos GAMLSS paramétricos são ajustados pelas estimativas de máxima verossimilhança em relação à estimação de β. Expressa que $Y \sim D(\mu, \sigma, \nu, \tau)$ implica que o logaritmo da função de verossimilhança, determinado pela verossimilhança observada da amostra é,

$$l = \sum_{i=1}^{n} ln[f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)], \tag{5}$$

de acordo com o pressuposto de independência das observações. Maiores informações, bem como uma tabela com os nomes das distribuições, espaço paramétrico e domínio das funções, respectivas funções de ligação e nome da distribuição na implementação do R, podem ser vistos em (STASINOPOULOS et al., 2017) páginas 148 - 151.

2.5 Modelo Binomial Negativo do tipo I

A definição para a distribuição Binomial Negativa do Tipo I, no pacote gamlss é dada por,

$$P(Y=y|\mu,\sigma)=\frac{\Gamma\left(y+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)}\Gamma(y+1)\left(\frac{(\sigma\mu)}{(1+\sigma\mu)}\right)^{y}\left(\frac{1}{1+\sigma\mu}\right)^{\left(\frac{1}{\sigma}\right)},$$

para $y=0,1,2,...,\infty, \mu>\sigma$. Essa parametrização é equivalente à usada por Anscombe (1950), com $\alpha=1/\sigma$ em vez de σ. O pacote `gamlss` contém a função `gamlss()` que pode ser usada para ajustar modelos de regressão. O pacote `gamlss.dist` contém duas funções as quais podem ser implementadas as seguintes distribuições.

- NBI() implementa a distribuição Binomial Negativa do tipo I, as parametrizações tais como a média e a variância de uma variável Binomial Negativa são dadas por:
- NBI(): média=μ e variância=$\mu+\sigma\mu^2$.
- Em contraste, a função glm.nb do pacote MASS usa a parametrização na qual uma variável que segue distribuição Binomial Negativa tem:
- Média = μ e variância = $\mu+\mu^2/\theta$.
- Se nós denotamos $1/\theta$ por σ, a parametrização da variância usada por `glm.nb()`,pode ser recuperada para a Binomial Negativa do tipo I do `gamlss()` pela seguinte igualdade: $\mu+\sigma\mu^2=\mu+\mu^2/\theta$.
- Por *default* na função `glm.nb()`, quando se modela $log(\mu)$ se estima o valor de θ. Por outro lado, usando `gamlss()` com a opção da familia (`family=NBI()`), modela-se $log(\mu)$ e reporta a estimativa do valor de $log(\sigma)$. Se computar a estimativa do valor de $log(1/\theta)$, baseado no resumo do modelo obtido por usar `gamlss()` com a opção `family=NBI`. Um exemplo usando dados que mostram a relação entre `glm.nb` e `gamlss` com a parametrização para NBI pode ser vista no link [4].

Uma questão importante é que ao implementar modelos no pacote `gamlss`, a Binomial Negativa do Tipo I, pode-se modelar o parâmetro σ, o que permite acomodar melhor efeitos de superdispersão.

2.6 Critério de seleção dos modelos

O GAIC é um critério de informação que leva em consideração o número de parâmetros e de graus de liberdade utilizados no modelo para penalizar os modelos mais complexos e evitar sobreajustes aos dados em amostras de grandes tamanhos (PAIVA;

[4] https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/

FREIRE; CECATTI, 2008). Ele é utilizado no mesmo contexto do critério de informação de Akaike (AKAIKE, 1998). O GAIC é definido por Voudouris et al. (2012), como:

$$GAIC(K) = -2l(\hat{\theta}) + (k \times gl),$$

em que l(θ) é o logaritmo da função verossimilhança e gl são os graus efetivos de liberdade do modelo ajustado. k é constante e torna-se a penalidade para cada grau de liberdade utilizado. Refere-se a $-2l(\hat{\theta})$ como desvio global, pois o GAIC (k) é a estatística obtida pela adição do desvio global.

O desvio global (GDEV) é uma importante medida para a seleção de modelos em GAMLSS, define-se como,

$$GDEV = -2l(\hat{\theta}),$$

em que, $l(\hat{\theta})$ é o logaritmo da função de verossimilhança ajustada, mostrado na Equação (5).

Para selecionar o melhor modelo será adotado o critério GAIC (k), ou seja, o modelo que apresentar menor valor de GAIC k escolhido, este será indicado como melhor modelo para o ajuste, pois o GAIC(k) penaliza modelos com muitos parâmetros. O critério de Akaike (AIC) (AKAIKE, 1998) e o critério de informação bayesiano (BIC) (SCHWARZ et al., 1978) são casos especiais do GAIC (k). Um caso se $k = 2$ o GAIC torna-se o (AIC) e um outro caso $k = ln(n)$ o GAIC torna-se o (BIC) (VOUDOURIS et al., 2012).

A seleção da distribuição apropriada é feita em dois estágios que são: o estágio de ajuste e o estágio de diagnóstico. O estágio de ajuste envolve a comparação de diferentes modelos GAMLSS ajustados usando um critério generalizado de informação de Akaike (GAIC) por Voudouris et al. (2012). O estágio de diagnóstico é feito via análise gráfica do resíduos e pelo gráfico de *worm plot*.

Na prática a seleção de variáveis explicativas é de extrema importância no ajuste do modelo estatístico. Quando X_k é uma coleção de variáveis explicativas acessíveis para consideração na modelagem do parâmetro θ_k de um modelo GAMLSS, em que $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) = (\mu, \sigma, \nu, \tau)$. Geralmente, X_k incluirá fatores e variáveis quantitativas que podem vir a fazer parte do modelo de forma linear ou por intermédio de funções de suavização (STASINOPOULOS et al., 2017).

2.7 Análise de resíduos

Os resíduos viabilizam a extração de informações que permitem diagnosticar ou demonstrar a qualidade do ajuste de um modelo (estágio de diagnóstico), além de averiguar se as suposições do modelo foram satisfeitas. Muitas ferramentas gráficas são utilizadas para identificar incongruências entre o modelo ajustado e as observações coletadas, um

equilíbrio entre essas duas fontes de informação é necessário para que o modelo possa ser analisado com algum sentido prático.

Este tipo de análise leva em consideração os resíduos dos quantis aleatórios normalizados que podem ser utilizados para explorar exclusivamente a distribuição utilizada para ajustar o modelo. Introduzidos por (DUNN; SMYTH, 1996), estes resíduos podem ser definidos por,

$$r_i = \phi^{-1}\{F(y_i; \hat{\theta})\},$$

em que ϕ^{-1} denota a inversa da função de distribuição acumulada de uma normal padrão, F$(\cdot)$ é a função de distribuição acumulada adequada aos dados e $\hat{\theta}$ é o vetor de parâmetros. Repare que, um modelo adequado tem seus resíduos r_i seguindo a distribuição Normal Padrão, para o modelo ser considerado correto.

Uma técnica primordial aplicada para analisar os resíduos de um GAMLSS são os gráficos de minhoca (*worm plot*), o estágio de diagnóstico envolve o uso deste gráfico bem específico, que permite a detecção de inadequações no modelo globalmente ou dentro de um intervalo de uma variável explicativa (BUUREN; FREDRIKS, 2001; BUUREN, 2007). O *worm plot* funciona como o gráfico normal quantil-quantil (Q-Q) sem tendência a fim de evidenciar alterações locais no domínio de uma dada covariável.

Uma importante propriedade do *worm plot* está no estabelecimento de intervalos com $(1-\alpha)\%$ de confiança para os quantis normais teóricos. Para tal, considerando um quantil z associado a uma probabilidade p e um tamanho amostral n, possibilitando determinar os limites do intervalo baseado na expressão $\pm z_{\frac{\alpha}{2}} f(z)^{-1}\sqrt{p(1-p)/n}$, em que $f(z)$ é a função de densidade de probabilidade da distribuição normal. Um diferencial do GAMLSS são os gráficos *worm plots* por serem uma ferramenta de diagnóstico para análise de resíduos que por ser expresso como um único gráfico, mas que abrange todo o intervalo da variável explicativa ou em diferentes regiões (intervalos).

Os gráficos *worm plots* permitem identificar regiões em que o modelo não é bem ajustado aos dados. O eixo vertical do *worm plots* descreve para cada observação, a diferença entre a sua localização nas distribuições teórica e empírica. Os pontos observados em conjunto, representam uma curva que se assemelha a uma minhoca, na Tabela 1 tem se a interpretação de diferentes *worm plot*.

3 RESULTADOS E DISCUSSÃO

3.1 Caracterização do experimento

O banco de dados foi obtido do trabalho de Souza (2018), no qual foram utilizadas plantas matrizes das cultivares Imperial, Pérola[5] e Smooth Cayenne, sendo quatro plantas de cada variedade, obtidas no Campo Experimental da Embrapa Mandioca e Fruticultura. Foram estabelecidos três diferentes intervalos de subcultivos, 30 dias, 45 dias e 60 dias para as três variedades quando o número de brotos formados a cada repicagem foi contabilizado. Na etapa de estabelecimento foram contabilizados o número de gemas intumescidas, contaminações fungicas (%), contaminações bacterianas (%) e gemas oxidadas. Para a realização do experimento, considerando a etapa de multiplicação, o número de gemas iniciais variou a depender da variedade, já que houve perda de gemas na etapa de estabelecimento. Para cada variedade, as gemas intumescidas foram divididas de maneira proporcional para os três intervalos de subcultivo adotados e foram realizadas seis repicagens, retirando os brotos dos frascos a cada repicagem e cortando-os longitudinalmente em partes iguais, para em seguida serem repicados em frascos com meio fresco de multiplicação. O número de brotos foi contabilizado ao final de cada repicagem, assim como as contaminações/frasco em cada repicagem foram verificadas por fora, anotadas e os frascos posteriormente descartados.

O delineamento experimental foi o inteiramente ao acaso em esquema fatorial 2 x 3 x 6 (dois cultivares, três intervalos de subcultivo e seis repicagens), com um número desbalanceado de repetições por tratamento, onde cada repetição se constituiu em uma planta. Foi utilizado o teste de Tukey a 5% de significância para compará-los, no caso da ANODEV e teste de Wald significativo. O experimento teve como variável dependente o número de brotos de cada intervalo de subcultivo. Considerando que os dados da fase de multiplicação são de contagem e possuem a *priori* uma distribuição de Poisson da família exponencial. Considerando existia superdispersão foi ajustado ainda os modelos de QuasiPoisson, Binomial Negativa e Binomial Negativa do tipo I, sob metodologia de modelos lineares generalizados para os três primeiros e Modelo Lineares Generalizados de Locação, Escala e Forma (GAMLSS) para o último, respectivamente. A equação do modelo experimental é dada por:

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + \tau\beta_{ij} + \tau\gamma_{ik} + \beta\gamma_{ik} + \tau\beta\gamma_{ijk} + \epsilon_{ijk},$$

em que, i é índice de cultivar (1 Imperial, 2 Smooth Cayenne). k é índice de subcultivo (1...6) e j representa o indice do tempo (30, 45, 60). E os demais termos no modelo seguem descrição padrão de um modelo fatorial 2 x 3 x 6 (fatorial triplo) maiores detalhes em (MONTGOMERY, 2017). Todas as análises foram feitas pelo *software* R (TEAM, 2018) e principalmente pelos pacotes MASS (RIPLEY et al., 2013), `gamlss` (STASINOPOULOS

[5] Pérola não foi utilizado por perda de parcelas elevada.

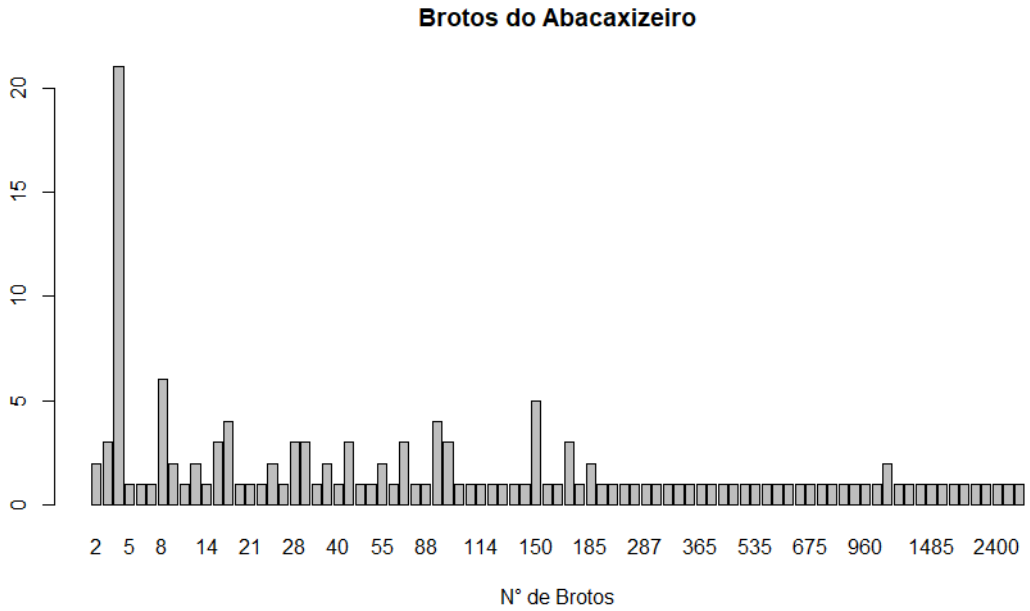
et al., 2015), Bsagri (SCHAARSCHMIDT; SCHAARSCHMIDT, 2018), Multicomp (VIHINEN et al., 1992). Nesta seção, vamos abordar os resultados obtidos com a modelagem estatística aplicada aos dados de abacaxizeiros. Na Tabela 1 compõe os resultados das estatísticas descritivas para as unidades das variáveis Subcultivos, Repicagens, Brotos.

Tabela 1 – Resultado das estatísticas descritivas dos dados do abacaxizeiro.

Variável	**Mín**	**Máx**	**Média**	**Mediana**	**Var**	**Assimetria**	**Curtose**	**Dp**
Sub	1	6	3,5	3,5	2,9	0,0	-1,3	1,7
Rep	1	4	2,5	2,5	1,3	0,0	-1,4	3,1
Brotos	2	3470	306,7	67,5	351229,1	3,1	11,1	592,6

Pelos resultados, é possível perceber que o número mínimo de brotos encontrados nos cultivares foi igual a dois, e o máximo foi de 3470, provavelmente pelo fato de existir grandes quantidades em alguns cultivares e pequenas em outros. A média ficou bastante distante da mediana e teve seu valor de 306,7. A variação foi de 351229,1. Como a variância não é igual a média, isso implica dizer que a distribuição de Poisson pode não ter um ajuste adequado para este tipo de dado, surgindo um indicativo para distribuição Binomial Negativa. Porém, esta variável brotos possui assimetria á esquerda, com coeficiente cerca de 3,1, já a curtose da curva é alta chegando a 11,1, e continuando com o desvio padrão também foi alto cerca de 592,6.

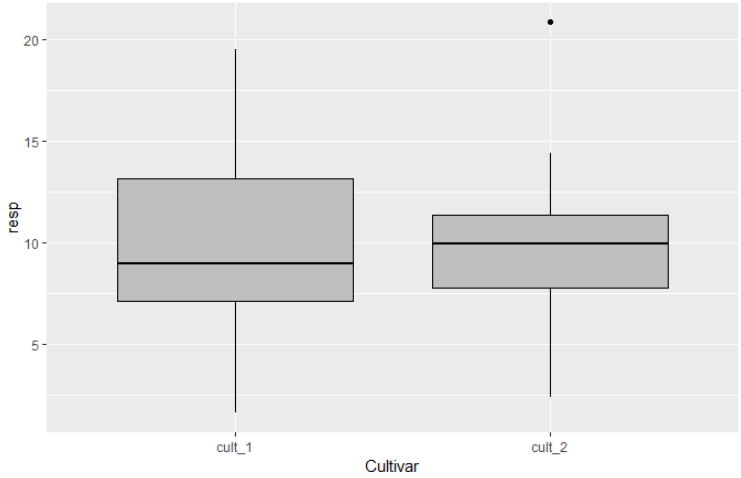
Figura 1 – Gráfico de Barras da variável Número de Brotos.



O gráfico de barras da Figura 1, reforça o que foi "visto"; com números na Tabela 1, em que a variável brotos tem assimetria à esquerda, com uma variação alta no início e uma ligeira queda ao decorrer do tempo. Seguindo com a análise, temos os gráficos de *boxplot* que é uma ferramenta gráfica para representar a variação dos dados observados. Em

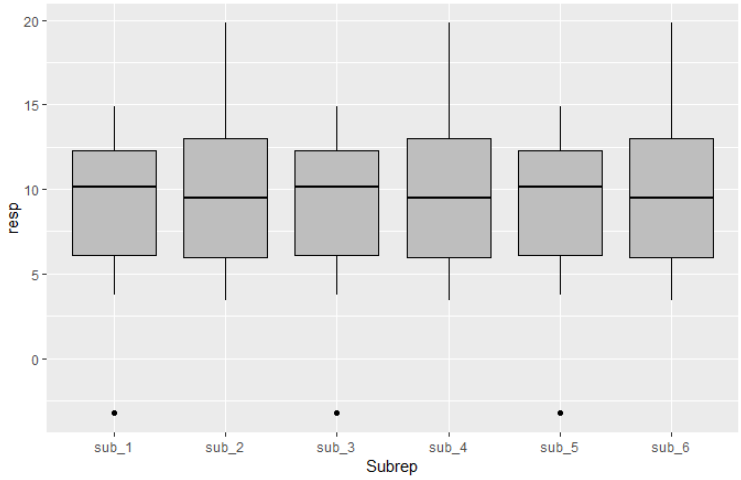
primeira instância temos o *boxplot* da variável Cultivar (Cult), no qual será apresentado na Figura 2. Pode-se observar que no cultivar 2 - Smooth Cayenne, apresenta-se um ponto

Figura 2 – *Boxplot* da variável Cultivar para o Número de Brotos.



fora dos limites do *boxplot*, o que pode ter influenciado na característica do gráfico, e ter deixado o mesmo não centralizado em torno da média. já com a variável Subcultivo,

Figura 3 – *Boxplot* da variável Subcultivo para o Número de Brotos.



o *boxplot* (Figura 3) é observado que os sub cultivos 1 ,3 e 5 apresentaram pontos fora da caixa, o que pode ter influenciado para decentralização da média destes subcultivos. Observa-se na (Figura 4) que os tempos 30, 45 e 60 não apresentam pontos fora da caixa, porem os três apontam crescimento da mediana em relação ao tempo. de acordo com a Figura 5, o índice (1.00) refere-se ao cultivar 1 com a variedade Imperial, o índice (2.00) refere-se ao cultivar 2 com a variedade Smooth Cayenne. Índices variando de (1 a 6) refere-se aos subcultivos. Percebe-se que os subcultivos de 1 à 3, não apresenta diferença significativa na produção do número de brotos entre o cultivar 1 e 2. Já nos subcultivos 4 à 6, percebe-se que houve decréscimo na produção do número de brotos, quando se compara cultivar 1 com o cultivar 2. Este resultado é sugestivo de efeito de interação entre

Figura 4 – *Boxplot* da variável Inter para o Número de Brotos.

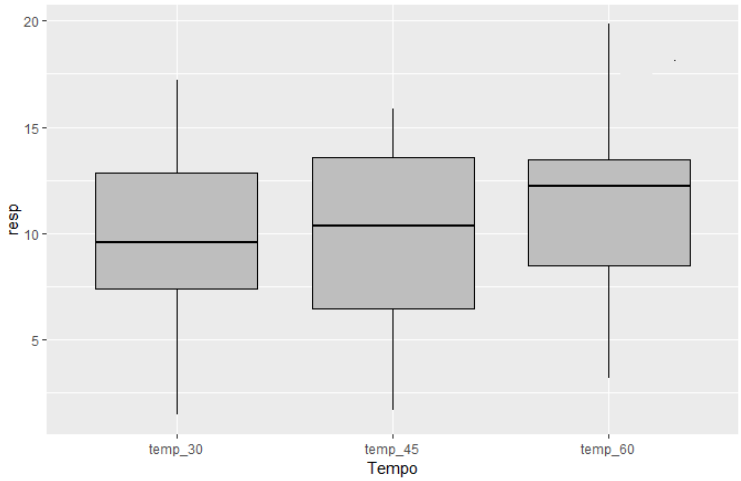
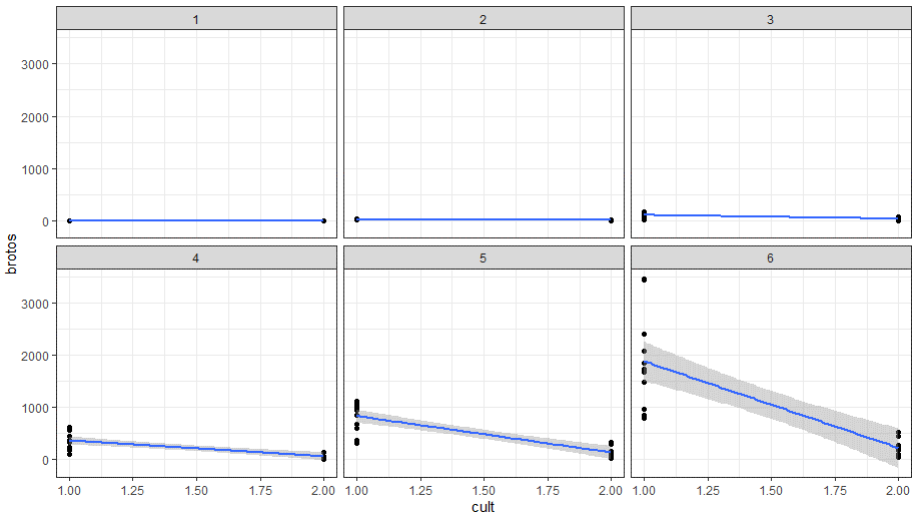


Figura 5 – Gráfico da interação de Cultivo e Subcultivo referentes ao Número de Brotos.



cultivar e subcultivo. Neste sentido, é importante salientar que não foi observado igual comportamento entre cultivar e Inter, subcultivo e Inter, e a interação tripla (figuras não apresentadas).

Foi construída uma tabela contendo os modelos propostos Normal, Poisson, QuasiPoisson, Binomial Negativo e Binomial Negativo do Tipo I para comparação pela distribuição que apresentar menor valor de Akaike Generalizado (GAIC), para ser indicada como o melhor modelo para os dados. Na Tabela 2, estão contidos os valores obtidos na comparação dos modelos. com base no que vimos descrito na Tabela 2, o modelo que se mostra mais adequado para este tipo de dado é o modelo Binomial Negativo do tipo I (gamlss), segundo o critério de seleção do modelo com menor GAIC. Para confirmar de forma gráfica que o modelo Binomial Negativo é adequado, vamos conferir as Figuras 6 à Figura 11, que mostra os gráficos *hnp* (Envelope simulado) dos modelos.

Tendo em vista que outros trabalhos já foram realizados nesta área, evidenciando

Tabela 2 – Comparação entre os modelos ajustados.

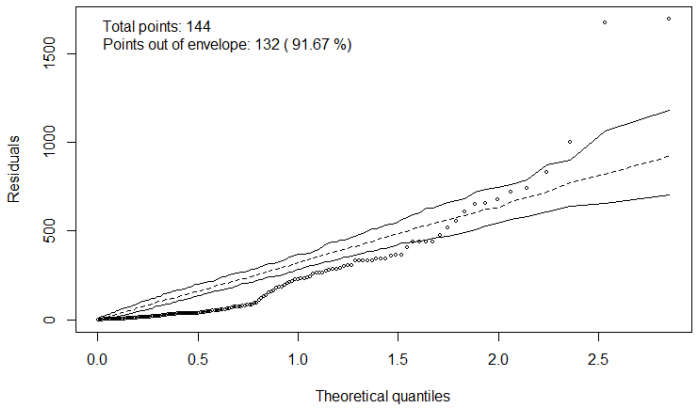
Modelo	Classe	GAIC	Deviance Residual	GL
Normal	ML	2034,65	1960,61	107
Poisson	GLM	7864,84	7792,84	108
Quasi-Poisson**	GLM	*	8094,7	136
Binomial Negativo	GLM	1608,65	1534,65	107
Binomial Negativo tipo I***	GAMLSS	1461,83	1377,31	102

* Parâmetro de dispersão para a familia quasipoisson dada por 59,82727; ** Package MASS (glm function); *** parâmetro σ modelado com o efeito de Subcultivo ($\sigma_k = \beta_0 + \gamma_k$).

um desses que realizou com mais de 60 genótipos de abacaxizeiros, no qual foram extraídos do banco de Germoplasma da Embrapa Mandioca e Fruticultura, SILVA et al. (2016), mostraram que apesar de uma distribuição normal nos dados da micro propagação destes acessos por quatro sub cultivos, a DMS (Diferença Mínima Significativa) gerada pela Anova foi muito elevada e não permitiu a expressão de visíveis diferenças significativas entre as variedades.

Para selecionar o melhor modelo que se ajustasse aos dados utilizamos também os Gráficos *Half Normal Plot* (hnp) como critério de confirmação de seleção do modelo que se ajustasse bem aos dados.

Figura 6 – Gráfico *Half Normal Plot (hnp)*|para a distribuição Normal.



Observamos na Figura 6 que quase todos os pontos estão fora do envelope, o que nos indica que a modelagem Normal não se adéqua a esses dados apresentando 91,67% dos pontos fora das bandas de confiança. De acordo com a Figura 7,percebe-se que todos os pontos estão fora do envelope, o que nos confirma a informação obtida nas estatísticas descritivas de que a modelagem Poisson não se ajusta bem estes dados pelo fato da média ser diferente da variância, pré suposto básico do modelo apresentando 100% dos pontos fora das bandas de confiança. Visualizando a Figura 8 constatou-se que embora a modelagem Quasi-Poisson estime o parâmetro de dispersão a partir dos dados, mesmo

Figura 7 – Gráfico *Half Normal Plot (hnp)*|para a distribuição Poisson.

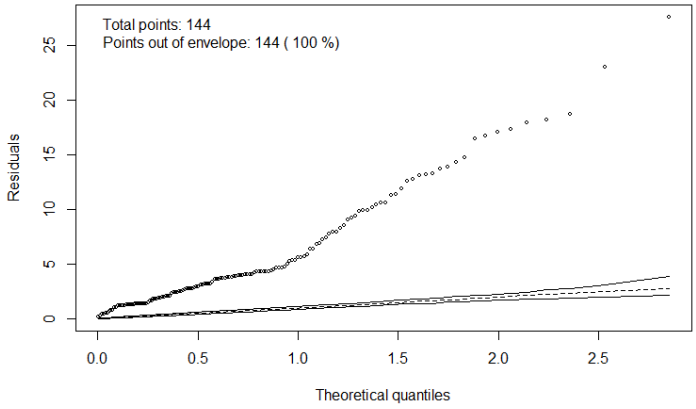
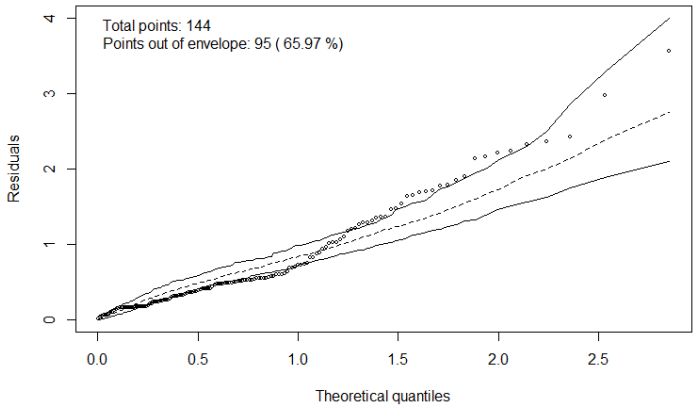


Figura 8 – Gráfico *Half Normal Plot (hnp)*|para a distribuição Quasi-Poisson.



assim não apresentou um bom ajuste para esses dados apresentando $65,97$ dos pontos fora das bandas de confiança. A respeito da Figura 9 apresenta a modelagem Binomial Negativa, que embora seja para dados de contagem e que apresentem superdisperssão também não apresentou bons resultados apresentando mais de 5% dos pontos fora das bandas de confiança.

Figura 9 – Gráfico *Half Normal Plot (hnp)*|para a distribuição Binomial Negativo.

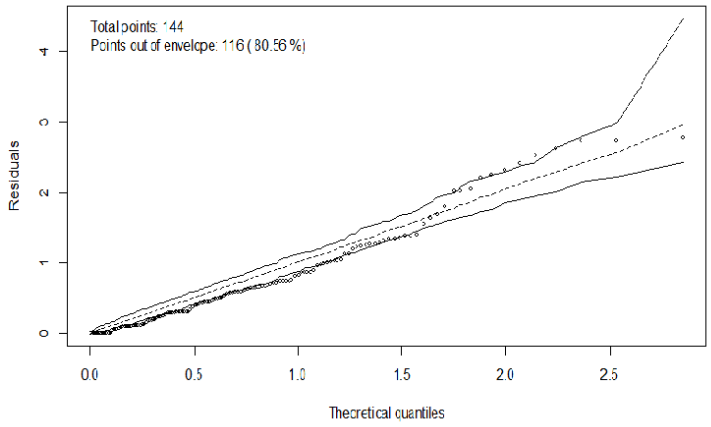
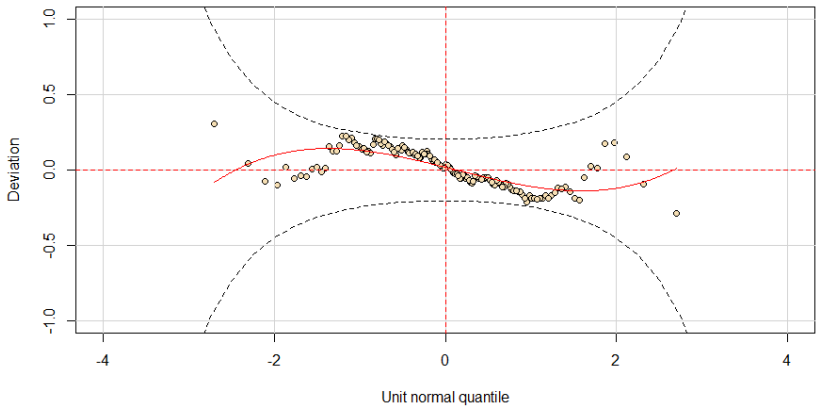


Figura 10 – *worm plot* para o número de brotos.



Com o resultado apresentado na Figura 10 todos os pontos estão dentro das bandas de confiança, o que nos indica que a modelagem Binomial Negativa do tipo I se ajusta bem aos dados.

Confirmando a ajuste na Figura 11 quase todos os pontos estão dentro do envelope, com a margem de erro menor que 5% cerca de 2,78%, indicando que a modelagem Binomial Negativa do tipo I se ajusta bem aos dados. O gráfico da Figura 12 corrobora a qualidade do ajuste com o *plot* dos resíduos via quantil dos resíduos versus valores ajustados, e dos quantis dos resíduos ordenados (índice). A densidade tem aparência sinuosa centrada em zero com pequenas variações nas caudas.

Figura 11 – Half Normal Plot (b) para o número de brotos.

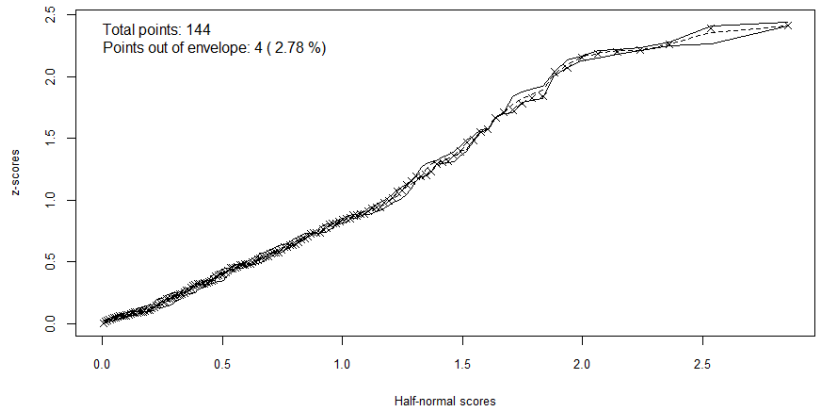


Figura 12 – Gráfico de resíduos do modelo GAMLSS para o número de brotos.

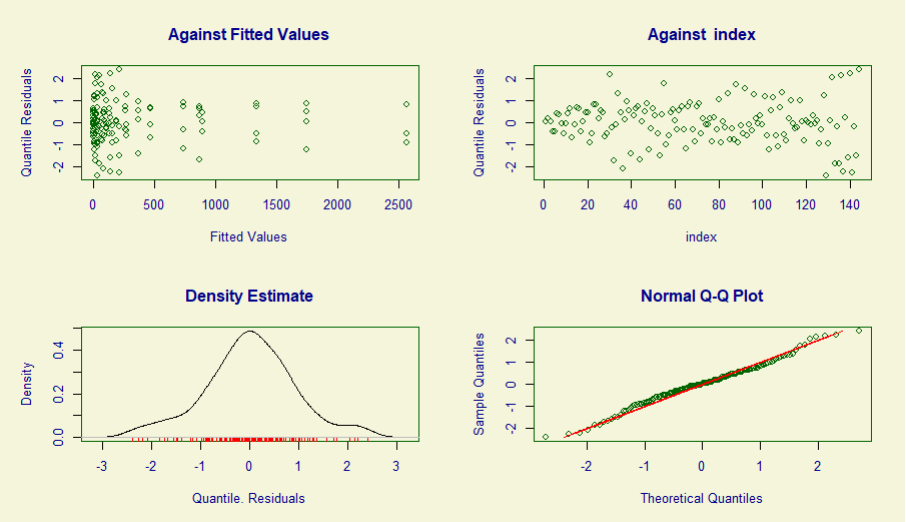


Tabela 3 – Resumo das estatísticas dos momentos da distribuição do Número de Brotos no modelo Binomial Negativo do tipo I

Estatística	**Momentos**
Média	0,0134
Variância	0,8621
Coeficiente de Assimetria	-0,0587
Coeficiente de Curtose	3,3138
Coeficiente de Correlação de Filliben	0,9938

Pelos momentos percebe-se que não houve grandes desvios de assimetria, ou curtose, ao se realizar o teste de Shapiro-Wilks para os resíduos percebe-se que para o modelo Binomial Negativo tipo I, que o *valor-P* foi de 0,1765 (com Estatística de teste de W = 0,98658).

Para efeitos de comparação temos que o (Sim) representa significativo a 5% e o (Não) representa não significativo a 5%, percebe-se que ao escolher o modelo Poisson, haveriam interpretações equivocadas com a possibilidade de fazer o desdobramento da

interação tripla de Cultivo, Subcultivo e Tempo. Sendo que pela tabela dos efeitos das distribuições Binomial Negativa percebe-se que apenas que a interação dupla Cultivo e Subcultivo foi significativa. Levando ao desdobramento dessa interação e teste pos hoc de Tukey para Cultivo dentro de Subcultivo e de Subcultivo dentro de Cultivo, trazendo resultados confiáveis quando comparado aos modelos Poisson, Quasi-Poisson, a Anova tradicional e o Binomial negativo do tipo I (Tabela 4). Este tipo de problema é comum,

Tabela 4 – Comparação entre os modelos ML, GLM e GAMLSS.

Distribuição	**Normal (ML)**	**Poisson (GLM)**	**Quasi-Poisson (GLM)**	**NB (GLM)**	**Nb tipo I (GAMLSS)**
(Intercepto)	Sim	Sim	Não	Não	Sim
Cult	Sim	Sim	Não	Não	Não
Sub	Sim	Sim	Sim	Sim	Sim
Inter	Não	Não	Não	Não	Não
Cult:Sub	Sim	Sim	Não	Sim	Sim
Cult:Inter	Não	Não	Não	Não	Não
Sub:Inter	Sim	Sim	Não	Não	Não
Cult:Sub:Inter	Sim	Sim	Não	Não	Não

de modo que já foi tratado em um trabalho por Mendes et al. (1999), o qual propôs o uso da técnica da regressão de Poisson, afim de examinar a ocorrência das taxas de multiplicação efetuadas na cultura da bananeira, o que após os estudos mostrou ser uma técnica que melhor se ajustou ao tipo de dados que estavam sendo trabalhados, que eram dados quantitativos.

Tabela 5 – Desdobramento da interação para o modelo NB tipo I GAMLSS para o número de brotos.

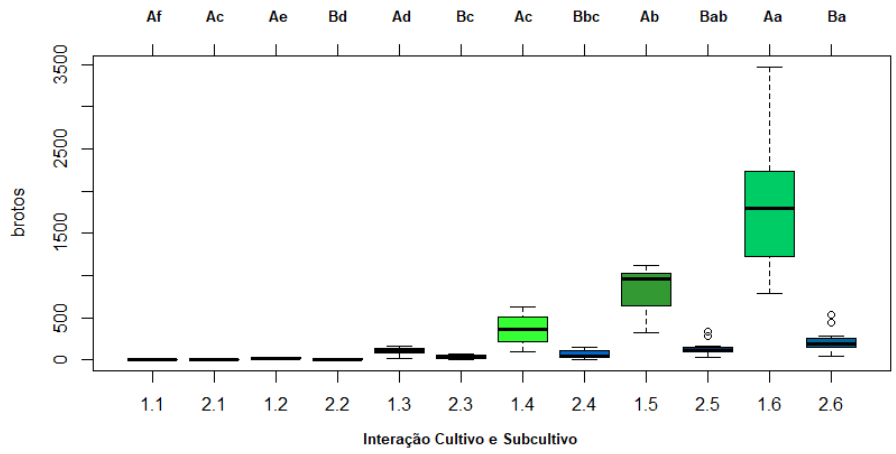
Cult\Sub	**Médias**					
	1	**2**	**3**	**4**	**5**	**6**
1	3,41 Af	19,41 Ae	103,25 Ad	365 Ac	833,91 Ab	1885 Aa
2	4 Ac	9,58 Bd	31,75 Bc	63,83 Bbc	136,58 Bab	224,08 Ba

Médias seguidas da mesma letra maiúscula nas linhas e minusculas nas colunas não diferem entre si pelo teste de Tukey ao nivel de 5%.

Com base na Tabela 5, observamos para o desdobramento da interação que no subcultivo 1 as médias não diferem estatisticamente entre si nos cultivares 1 e 2 contendo as variedades Imperial e Smooth Cayenne respectivamente. Já a partir do subcultivo 2 até o subcultivo 6, a média de Brotos do cultivar com a variedade Imperial, foi maior do que no cultivar 2 com a variedade Smooth Cayenne, constituindo a maior produção de brotos da variedade. Observando dentro do cultivar 1 com a variedade Imperial, o subcultivo 6 teve maior média de brotos em relação aos demais. Já dentro do cultivar 2 com a variedade Smooth Cayenne, os subcultivos 6 e 5, tiveram maior média de brotos relacionados aos demais, mesmo esta variedade tendo maior índice de brotos relacionados aos demais subcultivos não se sobressaiu a variedade Imperial. Souza (2018), ao modelar os brotos de abacaxi, obteve resultados semelhantes no que tange ao efeito de interação, porém a denominação das letras ao nível de 5% de significância do teste, diferiram nos subcultivos,

o que nos traz a conclusão que não utilizar o ferramental estatístico correto para os dados pode acarretar diferenças errôneas. Os resultados da Tabela 5 foram postos no gráfico da Figura 7 para uma melhor visualização em que as cores em tom de verde representam o cultivar 1 com a variedade Imperial e as em tom de azul o cultivar 2 com a variedade Smooth Cayenne, com graduação da coloração nos subcultivos. Isto posto, percebe-se que o tratamento 1.6 que é a combinação do cultivar 1 (Imperial) com o subcultivo 6, teve o maior valor médio de brotos de abacaxi da variedade Imperial, com a diferença pelo teste tukey representada pelas letras na parte superior do gráfico. Percebe-se também que os tratamentos formados pelo cultivar 1 e subcultivos 6 e 5, depois do tratamento 1.6, foram os que apresentaram maior número de brotos. Na bibliografia, existem algumas

Figura 13 – *Boxplot* da interação entre cultivo e subcultivo e diferenças de médias pelo teste de Tukey para o número de brotos no modelo Binomial do tipo I (GAMLSS).



informações acerca do efeito dos intervalos de subcultivos, pelo fato de exercerem sobre os valores de multiplicação. De forma geral o número de brotos ao longo dos subcultivos costuma não tender para uma distribuição normal, sendo questionável por isso o uso da ANOVA para esse tipo de avaliação. Mesmo com o uso de transformações nos dados para atingir esse objetivo, pode-se não obter bons resultados.Em estudos atuais (SOUZA, 2018) trata dessa questão da micropopagação do abacaxizeiro, que já havia sido estudada por Kofi e Adachi (1993) as quais analisaram a aplicação de dois ensaios de subcultivos de 30 e 60 dias na cultura do abacaxizeiro, os autores citados anteriormente ratificaram que o ensaio mais extenso desenvolveu maior número de brotos para um mesmo número de repicagens. Hamad e Taha (2008) mantiveram esse desempenho ao obter maior número de brotos para a variedade Smooth Cayenne no intervalo de 75 dias. Contudo Hamad e Taha (2008) acentuam que ao investigar o número de brotos em tempo similar (número de meses) em diversos períodos de subcultivos, o menor ensaio produziu maior número de brotos devido ao maior índice de repicagens efetuadas. Conforme essas literaturas, as

repicagens levam mais tempo na maturidade do crescimento dos brotos, precisando de um tempo maior no intervalo de subcultivos para nivelar essa aplicação e gerar brotos levemente desenvolvidos. Desta forma, alguns materiais dos autores comprovam a queda posteriormente a quarta repicagem, devido a falta de formação dos brotos no período de 30 e 60 dias.

Apesar disso, alguns trabalhos têm disponibilizado métodos estatísticos com melhor adequação para explorar dados relacionados aos índices de propagação ou dados de contagem de brotos, a exemplo disso o material de Mendes et al. (1999) aborda a questão do uso do modelo de regressão de Poisson. Embora o modelo Poisson seja uma boa indicativa para este tipo de dados, pode ocorrer de não apresentar estimativas fidedignas em alguns casos em que os dados apresentem superdisperssão. Por esse motivo, encontrar outro tipo de modelagem foi o fator primordial na concepção de melhores estimativas. Optar por meios mais flexíveis como o uso do GAMLSS na modelagem desses dados, proporcionou resultados opostos com outros tipos de modelagem testadas para a mesma massa de dados. Sabendo desses preceitos, a modelagem Binomial Negativa do tipo I se sobressaiu tanto no diagnóstico gráfico quanto nas estimativas encontradas para explicar qual cultivar seja ele Imperial ou Smooth Cayenne tenha maior índice de propagação de brotos. A partir destas estimativas constatou-se que a variedade Imperial tem maior índice de propagação de brotos.

Mendes et al. (1999) analisando o processo de taxas de multiplicação em seis espécies de bananas maçã, relatou a presença de variação na produtividade de brotos no decorrer das seis repicagens feitas, propondo uma perda das taxas após a quarta repicagem, exceto para a espécie B que permaneceu crescendo posteriormente o sexto subcultivo. A literatura cita que essa instabilidade no interior do genótipo pode originar-se de mudanças fisiológicas dos brotos, que aparecem fixados em diversos locais de transferência da planta matriz.

Em contrapartida, Hamad e Taha (2008) alegam que o extenso período de multiplicação ocorra de forma padrão, demostrando que 30% da constituição dos brotos surgem na fase inicial dos 30 dias e 40% surgem nos últimos 15 dias, no qual constitui o intervalo de 75 dias de incubação. Essas extremidades em desenvolvimento podem estar associadas ao nível de amadurecimento dos brotos, certo momento em que, segundo a literatura, os brotos desenvolvidos nos essenciais 30 dias, descenderam de gemas auxiliares localizadas em diferentes lugares onde o nível de maturação é mais desenvolvido, encerrando seu crescimento no decorrer da incubação, constituindo que esses brotos sejam avaliados na composição dos escritores semanalmente.

4 Considerações Finais

Inicialmente, como tratava-se de dados de contagem, para modelar o número de brotos de abacaxi utilizamos as técnicas clássicas de modelagem (Modelo Linear e Linear Generalizado). No decorrer da análise, percebeu-se que estes modelos não foram adequados devido as técnicas diagnósticas utilizadas (*hnp plot*). Neste sentido, partiu-se para tentar modelar o efeito de superdispersão pelos modelos QuasiPoisson e Binomial Negativo. No entanto estes não foram adequados também. A partir daí, ajustou-se modelos mais flexíveis que não precisassem necessariamente pertencer a família exponencial. Partimos então para classe de modelos gamlss.

Onde obteve-se as melhores estimativas com o modelo Binomial do tipo I (GAMLSS), sabendo que este é adequado de acordo com as técnicas diagnósticas, ao se utilizar o subcultivo como efeito ajustado no parâmetro de dispersão. Após a escolha do modelo, procedeu-se o teste de Tukey via pacote `Bsagri` e `Multicomp`, constatou-se que houve efeito de interação entre subcultivo e cultivar, sendo que no subcultivo 1, as médias do número de Brotos não diferem estatisticamente entre si nos cultivares 1 e 2. Já a partir do subcultivo 2 até o subcultivo 6, a média do número de Brotos do cultivar 1, foi maior do que no cultivar 2 com a variedade Smooth Cayenne. Observando dentro do cultivar 1 com a variedade Imperial , o subcultivo 6 teve maior média do número de brotos em relação aos demais. Já dentro do cultivar 2 com a variedade Smooth Cayenne, os subcultivos 6 e 5, tiveram maior média na produção do número de Brotos relacionados aos demais.

Concluindo assim, que a partir destes resultados obtidos com um ferramental estatístico utilizando a modelagem GAMLSS adequada, é possível tomar a decisão correta de plantio do abacaxizeiro que irá proporcionar ao produtor um ganho satisfatório no manejo da micropropagação para produção de mudas com qualidade, gerando maior porcentagem de lucro por hectare plantada, com padrão de qualidade exigido pelo consumidor.

Referências

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: *Selected papers of hirotugu akaike*. [S.l.]: Springer, 1998. p. 199–213. Citado na página 18.

ANSCOMBE, F. J. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, JSTOR, v. 37, n. 3/4, p. 358–382, 1950. Citado na página 17.

BE, L.; DEBERGH, P. Potential low-cost micropropagation of pineapple (ananas comosus). *South African Journal of Botany*, Elsevier, v. 72, n. 2, p. 191–194, 2006. Citado na página 11.

BUUREN, S. V. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 16, n. 3, p. 219–242, 2007. Citado na página 19.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001. Citado na página 19.

CABRAL, J. R. S.; SOUZA, J.; FERREIRA, F. R. Variabilidade genética e melhoramento do abacaxi. *Recursos genéticos e melhoramento de plantas para o nordeste brasileiro*, Embrapa Semi-Árido, Embrapa Recursos Genéticos e Biotecnologia/Brasília-DF ..., v. 1, 1999. Citado na página 10.

CHAMBERS, J. M.; HASTIE, T. J. et al. *Statistical models in S*. [S.l.]: Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, CA, 1992. Citado na página 13.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado na página 19.

FAOSTAT, F. Available online: http://www. fao. org/faostat/en/# data. *QC (accessed on January 2018)*, 2017. Citado na página 10.

FERRARA, G.; VIDOLI, F. Semiparametric stochastic frontier models: A generalized additive model approach. *European Journal of Operational Research*, Elsevier, v. 258, n. 2, p. 761–777, 2017. Citado na página 16.

GUERRA, M. P. et al. Estabelecimento de um protocolo regenerativo para a micropropagação do abacaxizeiro. *Pesquisa Agropecuária Brasileira*, SciELO Brasil, v. 34, n. 9, p. 1557–1563, 1999. Citado na página 11.

HAMAD, A. M.; TAHA, R. Effect of benzylaminopurine (bap) on in vitro proliferation and growth of pineapple (ananas comosus l. merr.) cv. smooth cayenne. *Journal of Applied Sciences*, v. 8, n. 22, p. 4180–4185, 2008. Citado 3 vezes nas páginas 11, 29 e 30.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models: some applications. *Journal of the American Statistical Association*, Taylor & Francis, v. 82, n. 398, p. 371–386, 1987. Citado na página 11.

KOFI, O.; ADACHI, T. Effect of cytokinins on the proliferation of multiple shoots of pineapple in vitro. *SABRAO Journal*, v. 25, n. 1, p. 59–69, 1993. Citado na página 29.

LIU, D. et al. Climate-informed low-flow frequency analysis using nonstationary modelling. *Hydrological Processes*, Wiley Online Library, v. 29, n. 9, p. 2112–2124, 2015. Citado na página 16.

MCCULLAGH, P.; NELDER, J. Chapman and hall. *Generalized linear models*, 1989. Citado na página 13.

MENDES, I. d. C. et al. Biomassa c e atividade microbiana em solos de cerrado sob plantio direto e plantio convencional. *Embrapa Cerrados-Outras publicações técnicas (INFOTECA-E)*, Planaltina: Embrapa Cerrados, 1999., 1999. Citado 2 vezes nas páginas 28 e 30.

MONTGOMERY, D. C. *Design and analysis of experiments*. [S.l.]: John wiley & sons, 2017. Citado na página 20.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado 2 vezes nas páginas 11 e 13.

PAIVA, C. S. M.; FREIRE, D. M. C.; CECATTI, J. G. Modelos aditivos generalizados para posição, escala e forma (gamlss) na modelagem de curvas de referência. *Rev. bras. ciênc. saúde*, v. 12, n. 3, p. 289–310, 2008. Citado na página 18.

PINHEIRO, J. et al. *R Core Team (2018). nlme: Linear and nonlinear mixed effects models. R package version 3.1-137*. 2018. Citado na página 15.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. Citado 3 vezes nas páginas 11, 13 e 15.

RIPLEY, B. et al. Package 'mass'. *Cran R*, 2013. Citado na página 20.

SCHAARSCHMIDT, F.; SCHAARSCHMIDT, M. F. *Package 'BSagri'*. [S.l.], 2018. Citado na página 21.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado na página 18.

SILVA, B. et al. Micropropagação de mutante floral de abacaxizeiro ornamental. In: IN: CONGRESSO BRASILEIRO DE FRUTICULTURA, 24., 2016, SÃO LUIS. FRUTEIRAS .... *Embrapa Mandioca e Fruticultura-Resumo em anais de congresso (ALICE)*. [S.l.], 2016. Citado na página 24.

SOUZA, E. M. de. *Micropropagação do abacaxizeiro e estudos correlatos de modelagem estatística*. Disserta (Mestrado) — Universidade Federal do Recomcavo Baiano – UFRB, 2018. Citado 3 vezes nas páginas 11, 28 e 29.

STASINOPOULOS, D. et al. Flexible regression and smoothing. the gamlss packages in r. *GAMLSS for Statistical Modelling. GAMLSS for Statistical Modeling*, 2015. Citado na página 21.

STASINOPOULOS, M. D. et al. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: Chapman and Hall/CRC, 2017. Citado 2 vezes nas páginas 16 e 18.

TEAM, R. C. *R: A language and environment for statistical computing; 2015*. 2018. Citado 2 vezes nas páginas 12 e 20.

VENABLES, W.; RIPLEY, B. Statistics and computing. *Modern applied statistics with S. Springers, New York, USA,*, 2002. Citado 2 vezes nas páginas 13 e 14.

VIHINEN, M. et al. Multicomp: a program package for multiple sequence comparison. *Bioinformatics*, Oxford University Press, v. 8, n. 1, p. 35–38, 1992. Citado na página 21.

VOUDOURIS, V. et al. Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, Taylor & Francis, v. 39, n. 6, p. 1279–1293, 2012. Citado na página 18.

WOOD, S. *mgcv 1.3. R package*. 2006. Citado na página 11.

ZEILEIS, A.; KLEIBER, C.; JACKMAN, S. Regression models for count data in r. *Journal of statistical software*, University_of_Basel, v. 27, n. 8, p. 1–25, 2008. Citado 2 vezes nas páginas 14 e 15.