UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Statistical modeling as an aid to academic research and control of citrus greening and citrus canker diseases in orange cultivation

Marcos Jardel Henriques

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Marcos Jardel Henriques

Statistical modeling as an aid to academic research and control of citrus greening and citrus canker diseases in orange cultivation

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
November 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

H518s

Henriques, Marcos Jardel
Statistical modeling as an aid to academic research and control of citrus greening and citrus canker diseases in orange cultivation / Marcos Jardel Henriques; orientador Francisco Louzada Neto. -- São Carlos, 2024.
114 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. ESTATÍSTICA APLICADA. 2. MODELOS NÃO LINEARES. 3. AMOSTRAGEM. 4. GREENING (DOENÇA DE PLANTA). 5. CANCRO (DOENÇA DE PLANTA). I. Louzada Neto, Francisco, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Marcos Jardel Henriques

Modelagens estatísticas como auxílio à pesquisa acadêmica e controle das doenças greening e cancro cítrico na cultura da laranja

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
Novembro de 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Marcos Jardel Henriques, realizada em 27/09/2024.

Comissão Julgadora:

Prof. Dr. Francisco Louzada Neto (USP)

Prof. Dr. Diego Carvalho do Nascimento (UDA)

Prof. Dr. Pedro Luiz Ramos (PUC-Chile)

Profa. Dra. Vera Lucia Damasceno Tomazella (UFSCar)

Prof. Dr. Paulo Henrique Ferreira da Silva (UFBA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

ACKNOWLEDGEMENTS

I would like to express my gratitude to everyone who, in some way, supported me on this journey. Especially to my advisor, Professor Dr. Francisco Louzada Neto. I am also deeply grateful to Professor Dr. Oilson Alberto Gonzatto Junior for the entire journey we have shared through statistics, ever since it all began with our undergraduate course in the Bachelor's program in Statistics at the State University of Maringá (UEM) - PR. We never imagined that we were about to discover a "new ocean" hidden amidst the sciences. I would also like to extend my thanks to the funding agency CAPES. And last but not least, I am grateful to everyone who, in some way, rooted for me.

RESUMO

HENRIQUES, M. J. **Modelagens estatísticas como auxílio à pesquisa acadêmica e controle das doenças greening e cancro cítrico na cultura da laranja**. 2024. 114 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Esta tese apresenta soluções estatísticas para alguns problemas relacionados à agricultura. Um deles propõe uma plataforma que gera planos amostrais fundamentados na estatística teórica, utilizando computação e considerando conhecimentos das ciências agrárias. A plataforma foi desenvolvida para gerar planos amostrais automáticos, visando agilizar a detecção da proporção da doença *greening* em lavouras de laranja. Isso porque, no Brasil, é exigido que se faça senso para se detectar tais proporções. Para esse primeiro caso, a modelagem foi estruturada por meio de técnicas de amostragem, através de hierarquias envolvendo distribuições de contagem e proporção, especificamente a Beta-Binomial e a FlexShape-Binomial. O segundo problema abordado nesta tese consiste no seguinte: há alguns anos, muitas revistas científicas da área de ciências agrárias passaram a exigir a realização de dois ensaios idênticos, em diferentes épocas, para a possibilidade de submissão às revistas. Ou seja, somente com os resultados dos dois ensaios, seria possível submeter o artigo para tais periódicos. Assim, com dois bancos de dados que quase atendem a essa exigência (ou seja, experimentos realizados de forma quase idênticas), foi proposta uma abordagem estatística para demonstrar a equivalência entre os dois experimentos, utilizando modelagens bayesianas para se comparar prioris informativas e posteriores. As diferenças entre os dois bancos de dados ocorreram durante a coleta dos dados. Para este segundo momento da tese, os dados são provenientes de experimentos planejados para detectar variedades de laranja resistentes à doença do cancro cítrico. Para solucioná-lo, a proposta consiste em apresentar um modelo de regressão não-linear baseado na distribuição de probabilidade Gamma, associada às curvas de crescimento Logística, Gompertz, Weibull e Hill. No terceiro problema, a tese busca analisar um conjunto de dados experimentais, cujo objetivo foi identificar as melhores combinações de porta-enxertos de variedades de laranja que conferissem resistência à doença do cancro cítrico às novas plantas. Nesta etapa, a modelagem foi realizada através da distribuição de probabilidade Beta Inflacionada de Zeros Bayesiana Longitudinal. Os três problemas principais da tese foram solucionados, e, além disso, direta ou indiretamente, outros problemas e resultados agronômicos como a descoberta de novas variedades resistentes à doença do cancro cítrico foram detectadas.

Palavras-chave: FlexShape-Binomial, Beta-Binomial, Gamma, Beta Inflacionada, Modelo Logístico, Gompertz, Weibull, Hill, Inferência Bayesiana, Dados Longitudinais.

ABSTRACT

HENRIQUES, M. J. **Statistical modeling as an aid to academic research and control of citrus greening and citrus canker diseases in orange cultivation**. 2024. 114 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

This thesis presents statistical solutions to several problems related to agriculture. One of them proposes a platform that generates sampling plans based on theoretical statistics, utilizing computation and considering knowledge from agricultural sciences. The platform was developed to generate automatic sampling plans, aiming to expedite the detection of the proportion of the *greening* disease in orange groves. This is because, in Brazil, it is required to conduct a census to detect such proportions. For this first case, the modeling was structured through sampling techniques, using hierarchies involving count and proportion distributions, specifically the Beta-Binomial and the FlexShape-Binomial. The second problem addressed in this thesis consists of the following: for some years, many scientific journals in the field of agricultural sciences have required two identical trials, conducted at different times, to allow submission to these journals. In other words, only with the results of both trials would it be possible to submit an article to such journals. Thus, using two datasets that almost meet this requirement (i.e., experiments conducted in nearly identical ways), a statistical approach was proposed to demonstrate the equivalence between the two experiments, utilizing Bayesian modeling to compare informative priors and posteriors. The differences between the two datasets occurred during data collection. For this second part of the thesis, the data are derived from experiments designed to detect orange varieties resistant to citrus canker disease. To address this, the proposal consists of presenting a non-linear regression model based on the Gamma probability distribution, associated with growth curves such as Logistic, Gompertz, Weibull, and Hill. In the third problem, the thesis seeks to analyze a set of experimental data, aiming to identify the best combinations of rootstocks for orange varieties that confer resistance to citrus canker disease in new plants. At this stage, the modeling was conducted using the Bayesian Longitudinal Zero-Inflated Beta probability distribution. The three main problems of the thesis were solved, and, in addition, directly or indirectly, other problems and agronomic results, such as the discovery of new varieties resistant to citrus canker disease, were achieved.

Keywords: FlexShape-Binomial, Beta-Binomial, Gamma, Inflated Beta, Logistic Model, Gompertz, Weibull, Hill, Bayesian Inference, Longitudinal Data.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

The use of statistics has been addressing the most diverse problems in society, especially those that require theoretical knowledge to support data analyses. Statistical theories are numerous, and through their applications, questions that demand data analysis are, in the vast majority of cases, answered. These applications range from healthcare and agricultural sciences to maritime studies, artificial intelligence, and even studies of the universe, among others. In other words, statistics universally serves as an essential support for decision-making in various fields of knowledge (MARTIN, 2001; TUFFÉRY, 2011; SENRA, 2011; CRESPO, 2020; SØBJERG *et al.*, 2021; LECOUTERE; CHU, 2024).

Currently, *statistics* is deeply integrated into research, development, and innovation, permeating all fields of knowledge. It plays crucial roles across various sectors of society, ranging from the most advanced theoretical and/or applied scientific research to industries, solving the most complex and varied problems, and influencing financial sectors. In other words, statistics is involved in virtually all problems requiring data analysis for their solutions. This is because statistical analysis guides decisions in diverse contexts, from resource allocation to the most detailed decision-making processes (SNEE, 1983; CLEARY, 2008; TAUFINA *et al.*, 2019; DURÁN; WIVES, 2018; RAMOS, 2016; SARAIVA *et al.*, 2012; BIEDERMANN; KOTSOGLOU, 2024).

All scientific and technological advancements in statistics are the result of centuries of study and are often related to applications in various fields of knowledge that required data analysis for their solutions. One of the first fields to benefit massively from the use of statistics was agricultural sciences, specifically agronomy, through the discoveries and scientific developments primarily achieved by Ronald Fisher (FISHER, 1935; SALSBURG, 2009). Since then, the field of agricultural sciences has been increasingly shaped by statistics. These fields are so closely connected that many researchers in agronomy often refer to the area as agricultural experimentation (FISHER, 1935; BOX; HUNTER; G, 2005; MONTGOMERY, 2008; BRIEN; DEMÉTRIO, 2009; SERMARINI *et al.*, 2020; SANTOS *et al.*, 2024). Some researchers have

even created specific terminology for this field. Many, for instance, refer to "plots" instead of "replicates" when discussing experimental units, as a pure statistician would denominate.

Going further, one can find several synonymous terms between agricultural sciences and statistics, where some researchers standardize values and terminology based on data from their studies (VERONESI *et al.*, 1995; SCAPIM; CARVALHO; CRUZ, 1995). Thus, continuing in this direction and leveraging numerous advancements in statistics and the expertise of many researchers, this thesis will address three problems in agriculture. In the following paragraphs of this introduction, these problems will be briefly presented.

In one of the proposals (the first), this thesis will present a proprietary platform that generates sampling plans backed by statistics, utilizing computational tools. The platform will be extremely useful to accelerate the detection of the proportion of plants with the greening disease in orange orchards (a highly relevant proposal for this area, since, by law, a census of agricultural fields is required, among other aggravating factors that make the work very costly). To solve this first problem, a platform (that develops sampling plans) will be built, capable of identifying the proportion of greening infection in orange orchards. The modeling will consist of hierarchies involving count and proportion distributions, such as the beta-binomial distribution. The FlexShape-binomial case will also be addressed. In other words, for this stage, an application was developed so that researchers, agronomists, farmers, and other professionals in the field can practically apply the results of this study.

Dumelle *et al.* (2022), working on comparisons between design-based and model-based approaches for spatial sampling and inference in finite populations, arrived at several conclusions. Design-based approaches rely on random sampling for inference, while model-based approaches depend on distributional assumptions. These authors compared these methods in a spatial context for finite populations using simulated and real data. They also found that model-based inference tends to outperform design-based inference, even for asymmetric data where the distributional assumptions of the model-based approach are violated.

Braswell *et al.* (2020), working in laboratories with citrus root samples, observed early and improved detection of *Candidatus Liberibacter asiaticus* in citrus, the vector insect of the bacteria causing greening disease. These authors worked with simple random sampling and, at times, divided orange plantations into clusters. The detection was not automatic but performed through laboratory analyses, which increases both time and costs. These authors employed DNA-based detection strategies using central vein leaf samples. On other occasions, these same authors worked with similar or identical methodologies in an attempt to capture the vector insect and subsequently conducted laboratory investigations to determine whether these insects had been contaminated. Based on these results, they aimed to predict the proportion of greening disease infection.

Although Brazilian law mandates a census in orange plantations, basic sampling techniques are often used (for verification purposes due to a lack of inspection agents). It is worth

noting that the entire process is still conducted manually, without any specialized software for this type of task. Even contrary to Brazilian law, the best that has been done so far in terms of sampling is cluster sampling conducted by the Fundo de Defesa da Citricultura - FUNDECITRUS (Science and Sustainability for Citriculture) (BASSANEZI *et al.*, 2024).

The second problem addressed by this thesis is the fact that many scientific journals in agricultural sciences/agronomy have, for many years, required two identical trials conducted at different times before submission (ADVANCES. . . , 2024; SEMINA. . . , 2024; CROP. . . , 2024). This often leads to experimental differences. Thus, with two datasets from experiments conducted by the thesis team, a statistical approach was proposed to demonstrate equivalence between the two experiments (a highly relevant proposal for agronomy/agricultural sciences, given the requirements of these journals). The differences between these datasets occurred during data collection. Hence, a statistical approach was proposed to demonstrate equivalence between the two experiments, using Bayesian modeling to compare informative priors and posteriors. With this type of result, it is possible to challenge such requirements of certain scientific journals. The priors and posteriors were obtained through the development of a nonlinear regression model based on Gamma probability distributions, associated with the Hill, Logistic, Gompertz, and Weibull growth curves. In other words, two identical modelings were carried out; however, for the second analysis, the results of the first were used as informative priors. It is worth noting that, generally in the agronomic field, analyses of this type of data are conducted separately using Analysis of Variance (ANOVA), as initially analyzed by the authors who conducted the trials that generated these data (GONÇALVES-ZULIANI, 2014a).

In the third problem, the thesis seeks to analyze an experimental dataset whose objective was to identify the best rootstock combinations of orange varieties that confer resistance to citrus canker disease in new plants. At this stage, the modeling was conducted using the Bayesian Longitudinal Zero-Inflated Beta distribution.

In summary, in addition to the three agronomic solutions presented, the thesis also offers agronomic results that contribute to improving outcomes related to the two agricultural diseases discussed earlier, including the identification of orange genotypes resistant to citrus canker disease. Since most farmers worldwide no longer practice agriculture that "imitates nature" (which would require a radical lifestyle change for the nearly 8 billion inhabitants of Earth), for now, in addition to the various genetic improvement methods known to science, the best approach is to find plants inherently resistant to possible diseases or to prevent these diseases from reaching the crops.

The next section of this thesis will present the Literature Review.

1.1 Literature Review

Hierarchical models can involve multiple levels of hierarchy. Even when consolidated into a single model, they allow for the inclusion of random effects at each hierarchical level. These effects correspond to the random errors that arise from variations between the units at each level (GOLDSTEIN, 1999).

From a theoretical perspective, the levels of a hierarchical model are composed of a random sample of the units under consideration. Thus, hierarchical models can analyze these data structures, allowing for individual specifications at each level, which are subsequently integrated into a unified model (RAUDENBUSH; BRYK, 2002).

The hierarchical model outlined by Casella and Berger (2010) is an example that addresses conditional probability results to form such a hierarchy. In their work, these authors illustrated the concept of the Binomial-Poisson hierarchy, presenting some specific observations in the given context.

> An insect lays a large number of eggs, each with a probability of survival p. On average, how many eggs will survive? The "large number" of eggs laid is a random variable, usually assumed to be $PO(\lambda)$. Furthermore, if we assume that the survival of each egg is independent, we have Bernoulli trials. Thus, if we let X be the number of survivors and Y the number of eggs laid, we have,
>
> $$X|Y \sim BI(Y,p),$$
> $$Y \sim PO(\lambda),$$
>
> a hierarchical model. Note that, for simplification, the notation $X|Y \sim BI(Y,p)$ was used, meaning that the conditional distribution of X given $Y = y$ is $BI(y,p)$ (CASELLA; BERGER, 2010).

Some authors have already worked with this Binomial-Poisson hierarchy. Working with data related to the efficacy of antidepressant compounds, Shkedy *et al.* (2005) were able to develop a proposal identical to the example presented by Casella and Berger (2010). In the study in question, the examined data consisted of binary outcomes obtained from a crossover experimental design. These data had previously been analyzed by Shkedy *et al.* (2005), who proposed techniques to estimate treatment effects using both generalized linear mixed models (GLMM) and generalized estimating equations (GEE) for binary data with clustering. In this case, it was assumed that the number of responses within each binomial observation remained constant, an assumption that may not be suitable for behavioral experiments like this one (DRL-72). In this context, the attempts recorded in each binomial observation may be influenced by the level of the administered dose. Therefore, Shkedy *et al.* (2005) refined the approach initially proposed by Shkedy *et al.* (2005) and introduced a Bayesian hierarchical model that combines binomial and Poisson elements. This model considers the number of responses as a random variable that follows a Poisson distribution. The conclusions drawn from the GLMM and binomial-Poisson

models show similarities, but the latter offers an additional advantage by allowing the estimation of the relationship between the successes achieved and the attempts made.

Through hierarchies, investigating HIV and teratology data, Comulada and Weiss (2007), while working with models for binomial data with random numbers of attempts, proposed Bayesian multivariate Poisson models for bivariate responses, correlated through random effects. Additionally, they extended the models for the analysis of longitudinal binomial outcomes and longitudinal multivariate data.

Beta-binomial models are widely used for overdispersed binomial data, with the probability of success being modeled according to a beta distribution. The number of binary trials in each binomial is assumed to be non-random and unrelated to the probability of success. However, in many behavioral studies, the binomial observations demonstrate more complex structures. Thus, a general beta-binomial-Poisson mixture model was proposed to allow a relationship between the number of trials and the probability of success for overdispersed binomial data (ZHU; EICKHOFF; KAISER, 2003).

Griffiths (1973), when working with data on the common cold and influenza, encountered the need to model using the beta-binomial distribution. The data consisted of information on infections from both diseases in urban households. Since the database comprised households with no cases of the diseases, the author worked to adjust a truncated beta-binomial model.

Analyzing Parkinson's data at two points in time, Lora and Singer (2008) and Lora and Singer (2011) modeled the data using beta-binomial/Poisson and beta-binomial/gamma-Poisson compositions, respectively. In the first study, the beta-binomial/Poisson regression models were used to model repeated bivariate counts. During this initial phase, the authors employed these models to capture repeated bivariate measures with covariates, utilizing multivariate Poisson distributions. They used maximum likelihood methods to fit the data from a study involving patients with Parkinson's disease and concluded that training sessions are beneficial for improving both agility and the ability to perform specified finger movements. The proposed model assumes that the covariances among total attempts are equal. In the second study, these same authors worked with beta-binomial/gamma-Poisson regression models for repeated counts with random parameters. As mentioned, they initially used beta-binomial/Poisson models, which were widely utilized by several authors at the time for modeling multivariate count data. Lora and Singer (2008) extended these models to accommodate repeated multivariate count data with overdispersion in the binomial component. To overcome some limitations of this model, Lora and Singer (2011) considered a beta-binomial/gamma-Poisson alternative that allows for both overdispersion and different covariances among the Poisson counts. In the 2011 study, they considered maximum likelihood estimates for the parameters using a Newton–Raphson algorithm and compared both models in a practical example.

The authors Shkedy *et al.* (2005) worked with the hierarchical binomial-Poisson model. That is, these authors utilized a hierarchical binomial-Poisson model for analyzing a crossover

design for correlated binary data when the number of trials depends on the dose. The analyzed data pertain to an antidepressant compound, with collected information stemming from binary results of a crossover design derived from experiments. The experiment was conducted in this manner and for the following purpose: the differential reinforcement of a 72-second low-rate (DRL-72) schedule is a standard behavioral testing procedure for screening a potential antidepressant compound. The data analyzed in the article are binary results from a crossover design for this experiment. Shkedy *et al.* (2004) proposed estimating the treatment effects using generalized linear mixed models (GLMM) or generalized estimating equations (GEE) for clustered binary data. The models proposed by Shkedy *et al.* (2004) assumed that the number of responses in each binomial observation would be fixed. This may be an unrealistic assumption for a behavioral experiment, as the number of responses (the number of trials in each binomial observation) is expected to be influenced by the level of dose administered. Thus, in the second article Shkedy *et al.* (2005), these authors extended the model proposed by Shkedy *et al.* (2004) and suggested a Bayesian hierarchical binomial-Poisson model, which assumes that the number of responses is a Poisson random variable. The results obtained from the GLMM and binomial-Poisson models are comparable. However, the latter model allows for estimating the correlation between the number of successes and the number of trials.

Other authors have also worked with hierarchical models, but with a different focus. Fabio, Paula and Castro (2012), working with a Poisson model, managed to incorporate a variable intercept, where it was assumed that the random effect would have a generalized log-gamma (GLG) distribution. This random effect was designed to handle the overdispersion of count data, capturing correlations between groups. To achieve their initial objectives, they obtained the marginal models through numerical integration methods and derived the multivariate negative binomial model using a specific parameter configuration of the hierarchical model.

There are many works in the literature that present models of proportion rates using the Beta distribution (CHOI, 2023; NAWA; NADARAJAH, 2023; BAREIKIS; MANSTAVIČIUS, 2024; NAWA; NADARAJAH, 2024; SHARMIN; ZULKAFLI; ALI, 2024), and even analyses of datasets involving proportions with zero inflation, modeled with the Inflated Beta distribution (OSPINA; FERRARI, 2012b; PENG; LI; LIU, 2016; PEREIRA; BOTTER; SANDOVAL, 2012; MARTINEZ, 2008; ABDEL-KARIM, 2017).

Regarding Inflated Beta regression models, some authors have developed their research through this type of modeling. (MARTINEZ, 2008; OSPINA; FERRARI, 2012b), observing the potential of the beta regression model and the scientific research conducted up to that point, recognized the need to develop a model that could handle proportions while capturing excess zeros (or ones), which is very common in datasets of this nature. Thus, these authors proposed "a combination of a beta distribution and a Bernoulli distribution, which are used to represent data situated in the intervals $[0,1]$, $[0,1)$ and $(0,1]$, focusing on degenerate values at zero and one." Several validation studies of this proposal have been conducted, allowing the development of

diagnostic techniques for the inflated beta regression model, and ultimately, this new theory was illustrated through a dataset.

Vila *et al.* (2024), Working with a model for rates and bimodal proportions (beta model), they noticed that the beta distribution is not suitable for modeling bimodal data within the unit interval. In an effort to obtain a distribution capable of capturing this bimodality, they proposed a bimodal beta distribution constructed using an approach based on the alpha-skew-normal model. These authors discussed various properties of this distribution, such as bimodality, real moments, entropies, and identifiability. Furthermore, they proposed a new regression model based on the proposed distribution and discussed the residuals.

Amid studies involving proportion modeling, (PEREIRA; BOTTER; SANDOVAL, 2012) introduced a truncated inflated beta distribution (TBEINF). These authors state that both the beta distribution and the inflated beta distribution are excellent options to start modeling and analyzing datasets of proportions. However, they assert that there are cases where they do not fit well to variables that do not take values in the open interval $(0, c)$, $0 < c < 1$. Thus, what they present is a mixture of the trinomial distribution with the beta distribution, constrained to the open interval $(c, 1)$. The properties are studied through Monte Carlo (MC) simulation, and this methodology was illustrated with data from the National Social Security Institute (INSS).

Peng, Li and Liu (2016), analyzing metagenomic data, achieved good results through zero-inflated beta regression. Current technological advances have propelled scientific research across various fields. As a result, metagenomic data have expanded far beyond previous expectations, thanks to tools that enable large-scale DNA sequencing (NGS). This has promoted significant advancements in areas involving genetic data, including clinical sectors.

In addition to sample size limitations (small samples) and dimensional complexity, metagenomic data are characterized by proportions with a high number of zeros and sparse nature. These researchers were able to validate something in the literature with this data analysis. By analyzing human metagenomic data, they confirmed and identified biologically important taxa. This aligns with the ultimate goal of science: to provide results that can trigger advancements and improvements in society.

Abdel-Karim (2017), in their studies, introduced the "Extended Zero-One Inflated Beta" model and three-part regression models adjusted for proportion data analysis. These models are designed for proportional response data that exhibit distinct probabilities of zero, allowing the response variable to take values of zero and one. The proposed models are specifically developed to address the peculiarities of asymmetry and heteroscedasticity present in fractional datasets, aiming to estimate unknown parameters accurately. Through extensive Monte Carlo simulations, the performance of both approaches was evaluated in terms of bias and root mean squared error. Additionally, practical applications of these models were illustrated with a real dataset.

Throughout this thesis, we will also work with the Beta-Binomial hierarchy. Many

authors have worked with this probability model (CHEN; LI; ZHU, 2023; NAJERA-ZULOAGA *et al.*, 2023; ZHOU; LIN, 2023; ĆMIEL *et al.*, 2023; AGHAYERASHTI; SAMANI; GANJALI, 2023).

One of the most recent studies involving this hierarchy discusses modeling the weekly number of cities reporting new cases of cryptosporidiosis (CHEN; LI; ZHU, 2023). The authors propose a GARCH model involving the beta-binomial model. In practice, the model incorporates covariates and utilizes a logit transformation to capture time series characteristics. They demonstrated the existence of a stationary and ergodic solution by introducing a contraction condition to the conditional mean process and a Markov structure to the covariate process.

Najera-Zuloaga *et al.* (2023) proposed modeling involving clinical data through beta-binomial regression. The data consisted of Patient-Reported Outcomes (PROs), which are frequently used as primary outcomes in clinical research studies. This information is typically measured on ordinal scales and tends to exhibit overdispersion. Researchers in this field often analyze these dimensions separately. Thus, the authors propose a multidimensional model, including several PROs in a joint analysis. Their proposal was evaluated and compared to independent analysis through a simulation study and an application to real data from patients with respiratory diseases. The results showed advantages over conventional analyses in terms of significance and interpretation of parameters.

Using hierarchical models, (ĆMIEL *et al.*, 2023) proposed a statistical tool based on beta-binomial distributions that can construct a robust gene co-regulation network (CoRegNet) among tens of thousands of experiments. These authors demonstrated that their analysis and proposal brought new insights to the literature in this field. Previously, it was known that genes were generally co-regulated linearly; however, they discovered an interesting set of genes that are co-regulated non-linearly. That is, half the time they change in the same direction, and the other half they change in the opposite direction. Moreover, these researchers uncovered a set of gene pairs that follow Simpson's paradox. They worked with public data, as millions of RNA sequencing samples have been deposited in public databases, providing a rich resource for biological research. These datasets encompass tens of thousands of experiments and offer comprehensive information about human cellular regulation. However, a significant challenge is how to integrate experiments conducted under different conditions. These researchers further state that by utilizing such data, CoRegNet offers a powerful approach to identify functionally related gene pairs, potentially revealing new biological insights.

Other authors Aghayerashti, Samani and Ganjali (2023) also achieved good results through modeling illustrated by the Beta-Binomial model. The study involved a Bayesian modeling of latent variables for correlated mixed classification responses and beta-binomial with missing data for an international statistical literacy project poster competition. The data come from a real dataset from the International Statistical Literacy Project (ISLP), during a poster competition held in 2020-2021 for undergraduate students in Iran.

Tripathi, Gupta and Gurland (1994) worked on parameter estimates for the beta-binomial model. Their proposal consisted of suggesting alternative methods for estimating parameters in both beta-binomial and truncated beta-binomial models. Some of these methods are advantageous because they produce estimators based on linear equations. This, in a way, facilitates the construction of confidence intervals and hypothesis tests regarding the parameters. For the beta-binomial distribution, a simple estimator based on moments or factorial moment ratios has high asymptotic relative efficiency for most parameters and is an attractive and viable alternative for calculating the maximum likelihood estimator. It is also simpler to compute than an estimator based on the mean and zeros, proposed by (CHATFIELD; GOODHARDT, 1970).

Yamamoto and Yanagimoto (1992) worked on moment estimators for the beta-binomial distribution. That is, they proposed a new moment estimator for the dispersion parameter of the beta-binomial distribution. This estimator performs better than the usual moment estimators, as well as better than the moment estimator proposed by Tamura and Young (1987).

On the other hand, Negrao, Aquino and Bearzoti (2001) focused on a different aspect concerning hierarchical models. These authors concentrated on evaluating estimation methods for the parameters of the beta-binomial distribution via Monte Carlo simulation. Specifically, they aimed to assess the methods of moments, maximum likelihood, means and zeros, and analysis of variance in estimating the parameters of the beta-binomial distribution via Monte Carlo simulation. For this, they calculated evaluating statistics for estimators: bias, mean squared error, simple consistency, and standard deviation. The PROC UNIVARIATE procedure was also applied to the sequences of estimates to verify the asymptotic normality of the distribution of the estimators. It was concluded that the estimators from analysis of variance and maximum likelihood exhibited the best properties together. The moment estimator was consistent starting from a sample size of twenty, while the others were consistent for any sample value greater than five.

An adjusted sample size algorithm for clusters was developed using a beta-binomial model (FOSGATE, 2007). The authors aimed to design a computer algorithm to calculate sample sizes for estimating proportions while incorporating clustered sampling units using a beta-binomial model when information about intraclass correlation was unavailable. To achieve this, a computer algorithm was written in FORTRAN and evaluated for a hypothetical sample size situation. It is essential to incorporate clustering adjustment in sample size calculations when designing epidemiological studies (for example) to estimate disease burdens and other population proportions in the presence of correlated data. Beta-binomial models can be employed to account for clustering, and design effects can be estimated by generating beta distributions that encompass the correlation within the cluster.

The beta-binomial distribution used to estimate the number of false rejections in gene expression studies from microarray data was examined by Hunt, Cheng and Pounds (2009). In the analysis of differential gene expression in microarray data, it is common to assume independence

among null hypotheses, which implies independence in gene expression levels. This assumption of independence leads to the binomial distribution for the number of false rejections and results in an empirically derived false positive discovery rate estimator. The number of false rejections is then modeled using the beta-binomial distribution, and an estimator of the false positive discovery rate from the beta-binomial distribution is derived. This approach considers how the correlation among non-differentially expressed genes affects the distribution of false rejections. As an illustration, the authors use this method to compare the gene expression of soft tissue sarcoma samples with normal tissue samples.

However, several authors have already studied hierarchical models, and many others continue to explore this area, such as (SMITH, 1983), for instance, who worked on the Maximum Likelihood estimation of parameters for the Beta-Binomial distribution. Meanwhile, Navarro and Perfors (2012) wrote a manuscript that provides a detailed discussion of the Beta-Binomial model. One of the main objectives of this manuscript is to provide a more comprehensive resource to facilitate understanding and address some of the interesting technical issues that may arise when attempting to construct our own hierarchical models based on practical contexts.

With different focuses and distinct formats, many researchers have worked with Gamma probability distributions (RAMOS *et al.*, 2024; RAMOS *et al.*, 2021; RAMOS *et al.*, 2019; MOALA; RAMOS; ACHCAR, 2013; TOMAZELLA; LOUZADA-NETO; SILVA, 2006; MOLENBERGHS *et al.*, 2015; CORDEIRO *et al.*, 2014; ORTEGA; RIZZATO; DEMÉTRIO, 2009).

Além disso, há muitos trabalhos na literatura que trazem modelagens envolvendo curvas de crescimento (SUNWASIYA; CHANDOLIA; UTTAM, 2024; KIMANI, 2024; JABBAR; AL-SAEDI, 2024; SANGIN *et al.*, 2024; WAIZ; GAUTAM; WAIZ, 2019; BROWN; MAYER, 1988; CHEN; HOOVER, 2003; CHEN, 2007).

Dagogo *et al.* (2023) revisit the idea that traditional statistical methods for nonlinear models require a starting point (initial parameters or estimated values) to initiate the optimization process. The expression of the nonlinear model must be defined, parameters declared, and initial values specified, after which the parameters are estimated through some iterative method. In their work, these authors utilized a computer program to estimate three growth models: Weibull, Richards, and Gompertz, employing a modified version of the Levenberg-Marquardt method to solve a nonlinear regression model. The growth models were decomposed in terms of additive and multiplicative errors, assisting in identifying the most suitable model for growth studies. Consequently, the issue of initial parameters was addressed through second-order regression techniques before an iterative approach was conducted. The result includes the final parameter estimates, standard errors, p values, and model fit criteria, used to determine the most appropriate values for the growth model. This study successfully identified the Weibull growth model with additive error terms as the optimal growth model. These findings recommend the Weibull growth model for future studies involving growth curves.

McKellar and Lu (2003), in the second chapter of their book, through practical and

theoretical explanations, as well as applied methods, provide examples and discussions on growth curves: Hill, Logistic, Weibull, Gompertz, among others. Additionally, examples using the Gamma distribution are also presented. For these authors, the concept of the primary model is fundamental for areas such as predictive microbiology, biological sciences, and related fields. Models like these, for microbial growth, aim to describe the growth process with the fewest possible parameters while accurately defining the different phases of the curve. When the increase in population density is plotted against time, the resulting curve typically has four phases, referred to as the lag phase, exponential phase, stationary phase, and death or decline phase. As they point out, these are empirical applications of logistic and Gompertz functions, for instance. In recent decades, new generations of bacterial growth curve models have been developed, such as the Baranyi, Hills, Buchanan models, and heterogeneous population models, among others.

Buchwald (2007), researching better (or even different) ways to work with growth curves, discovered something new to describe bilinear data. In their research, they propose a completely generalized version of a linearized biexponential model ($LinBiExp$), to allow smooth and fully parameterizable transitions between two linear segments while maintaining a clear connection with linear models. This proposal is appropriate since linear models are widely used due to their unparalleled simplicity but cannot be applied to data that exhibit a turning point or rate change, even if the data shows good linearity sufficiently far from that point. In their work, Buchwald presents applications and brief conclusions. Various profiles of biological and medical interest are shown, including growth profiles, such as human height, agricultural crops, fruits, multicellular tumor spheroids, single-fission yeast cells, or even labor productivity, and decline profiles, such as age effects on cognition in patients developing dementia and lactation yields in dairy cattle. In all these cases, quantitative model selection criteria, such as Akaike information criteria and Bayesian Schwartz criteria, indicated the superiority of the bilinear model compared to less parameterized suitable alternatives, such as linear, parabolic, exponential, or classical growth models (e.g., logistic, Gompertz, Weibull, and Richards). The *LinBiExp* provides a versatile and useful bilinear functional form with five parameters that is convenient to implement, suitable for complete optimization, and utilizes intuitive and easily interpretable parameters. Thus, the author concludes that the *LinBiExp* allows the fitting of general bilinear-type data in a single nonlinear regression step, and the MSC can be used to judge its adequacy compared to other possible models. Being considered a natural extension of linear models, its parameters are intuitive and easily interpretable, enabling convenient applications in various fields.

Working with comparisons of nonlinear models to describe the growth curves of broiler chickens fed different levels of corn bran, Masoudi and Azarfar (2017) were able to monitor the growth of birds in the poultry industry. The study utilized the logistic, Gompertz, Lopez, and Richards models. The researchers concluded that the Gompertz growth model was the best for describing the body weight growth curves of broiler chickens. In general, the results showed that the Gompertz model better described the biological curves of chickens fed corn bran than other models. Furthermore, the growth parameters were affected by the corn bran. The study

involved planned experiments on the theme previously presented. Supporting research utilizing growth curves, Yang, Kozak and Smith (1978) had already noted the advantages of using the Weibull model for modeling them many years ago. These authors observed that the potential of Weibull-type functions as growth curves is quite flexible. They developed a new growth function that possesses sufficient flexibility in its form to encompass most patterns of biological growth. This function is created by including an expansion factor in the Weibull distribution function. Many biologically increasing phenomena can be effectively modeled by this function, allowing for variation in the numerical values of the scale, shape, and upper asymptote parameters. These researchers demonstrated the applicability of this function, showcasing height-age and volume-age curves for individual trees, as well as two volume-age curves for polymorphic sets of trees.

Zeide (1993), working with analyses of growth equations and analyzing forestry data, provides important observations about the data field that can assist us in interpreting biological data (i.e., how to explain it practically from the perspective of data science). In their work, this author notes that plant growth results from two opposing factors: the intrinsic tendency for unlimited growth (biotic potential) and the constraints imposed by environmental resistance and aging. The tendency for expansion prevails in the early life of a tree, while growth decline becomes prominent toward the end. Existing growth equations can be transformed (by differentiation, decomposition into division components, and application of logarithms) such that the components corresponding to these two factors are exposed. This transformation reveals two basic intrinsic forms in most analyzed equations. Their common characteristic is that growth expansion is proportional to the current size of the tree. The decline in growth of individual trees appears to be more variable and can be represented with equal precision by a variety of expressions. This may reflect that a greater number of factors hinder growth: resource scarcity, competition, reproduction, diseases, herbivory, disturbances, etc. Consequently, the growth path is inherently imprecise and can be viewed as a broad valley rather than a single line.

However, consulting various journal portals and researching the literature for the Gamma probability distribution (and some others) associated with the logistic, Gompertz, Weibull, and Hill growth curves reveals very few works, which have purposes quite distinct from the proposals of this thesis (NADARAJAH; AFUECHETA, 2017; MAZUCHELI; EMANUELLI, 2019; TJøRVE, 2003; TSOULARIS, 2001; SANGIN *et al.*, 2024; UDOUMOH; EBONG; IWOK, 2017; CHAN; AFUECHETA, 2016).

1.2 Motivations

There are three major motivations that triggered the development of this thesis.

1.2.1 Motivation 1

This motivation is related to the *greening* disease and aims to solve a sampling problem. Essentially, the challenge is to propose an online platform that allows end-users working in this agricultural sector to handle sampling automatically.

Like every agricultural crop, citriculture requires various management practices to achieve good production results in terms of both quantity and quality. However, one of the biggest challenges is diseases (AMORIM; REZENDE; CAMARGO, 1997). Among these, *huanglongbing* (currently better known as *greening*) stands out. For any infected plants in orange groves, it is necessary to remove them, eradicate them entirely (100%), and plant new, healthy saplings in their place. In other words, since *greening* disease has no phytosanitary control, eradicating the orange tree is the only way to eliminate the infection. Thus, any citrus tree infected with this disease must be immediately removed from the grove and destroyed (JUNIOR *et al.*, 2017). This is essential to prevent epidemics of this disease, which could render large areas unsuitable for citrus production within a few years (ALVES[1]; BELOTI[1] *et al.*, 2014).

The first report of *huanglongbing* (*greening*) infection occurred in China in 1919, where it received its name Reinking *et al.* (1919). Later, in 1937, the disease was detected in South Africa and named *greening* (MERWE, 1937). The first report in Brazil was in 2006 (TEIXEIRA *et al.*, 2005; TEXEIRA *et al.*, 2005), and currently, only Europe remains free of the disease (JúNIOR,). The disease is transmitted by vector insects. For example, orange groves infected by *greening* in Florida (USA), which had virtually no infected plants around 2010, reached over 70% infected plants within five years, with 30% of fruits lost in the 2014 harvest (VIAGEM, 2014; GERALDELLO *et al.*, 2015). The disease is present in all Brazilian states. In the citrus belt of São Paulo and the Triângulo/Southwest of Minas Gerais, for instance, the infection rate by 2014 was 14%, while in Paraná it was between 2% and 4% (VIAGEM, 2014). By 2021, the infection rate in orange groves of these same regions (São Paulo and Minas Gerais) had risen to around 22.37% (CITRICULTURA, 2021). Orange groves in the agricultural region of Paranavaí, Paraná, currently have an infection rate of 10%. An alert from monitoring coordinators indicates strong evidence of underreporting in data related to the current *greening* infection rate in these groves, which increases concerns (MARQUES, 2023).

In the mid-2010s, entire farms in California, USA, became unable to produce any citrus species due to underestimating the risk posed by the disease. When infected, young plants fail to produce, mature plants experience significant fruit drop, and eventually die. This disease is transmitted by the psyllid insect (*Diaphorina citri*). Therefore, the measures to minimize the appearance of the disease in groves include the use of healthy saplings, eradication of diseased plants, and control of the vector insect (AMORIM; REZENDE; CAMARGO, 1997). Due to the severity of the disease, Brazilian law imposes limits and fines on violators. Instruction Normative No. 53, issued by the Ministry of Agriculture in October 2008, states that fines can reach nearly 90,000 reais for farmers who fail to eradicate diseased plants or completely eradicate groves

when infection exceeds 28% (STEPHANES, 2008).

Furthermore, inspections of these groves are conducted by the agricultural defense agency of each state. However, as the law requires a **CENSUS** during these inspections, it has become impossible for technical agents to perform inspections due to the enormous workload. Evaluation forms were thus distributed to farmers, who committed to conducting evaluations, eradicating infected plants, and reporting their findings to sanitary defense agencies. Over time, it was observed that most farmers either withheld information or struggled to detect infected plants (which requires studies, training, and technical expertise). This could potentially render Brazil incapable of producing citrus in the near future.

Looking at other countries, despite spending over $2 billion combating the disease in recent years, the US has seen its orange production drop from approximately 7 million tons in the 2007/2008 harvest to around 1.65 million tons in the last harvest. Florida, once the world's largest citrus-growing region, produced approximately $650,000$ tons. In this state alone, annual losses exceed $1 billion (BASSANEZI, 2023). The incidence of *greening* in Brazilian citriculture reached 24.42% in 2022. In 2018, this rate was around 18%. This record has concerned the sector since 2023, according to a study by Cepea/Esalq-USP (BOTEON, 2023).

Given the economic significance of this citrus disease and the lack of phytosanitary control, this thesis proposes a platform that generates sampling plans with georeferenced data (from orange groves). Thus, agronomists, farmers, technicians, and other professionals involved can estimate the percentage/proportion of infection by this disease in groves using sampling techniques as end-users of this platform.

1.2.2 Motivation 2

Motivation 2 is related to citrus canker disease. It involves modeling two datasets to propose a viable solution that meets a strong requirement of most agricultural science journals, as shown in Table 1: conducting two identical trials at different times for submission to the journal. For decades, many scientific journals in agronomy/agricultural sciences have required researchers to duplicate their experiments across different seasons (BARRERO; (VICE), 2023).

- When the experiment is conducted in controlled environments (laboratories, greenhouses, etc.), a repetition over time is required (conducting two trials at different moments. Example: one in January and another in March of the same year). In this case, the justification for the repetition is to increase the reliability of the results.

- For field experiments, it is requested that at least the repetitions be conducted in different growing seasons. Example: one experiment during the 2021/2022 season (from October to March) and the other during the 2022/2023 season (from October to March). For field experiments, this request is justified to allow the results to be compared, enabling

Table 1 – Journals in the fields of agronomy and agricultural sciences that require the duplication of trials conducted at different times

Journal	**Electronic Address**
Advances in Weed Science	https://awsjournal.org/
Semina Ciências Agrárias	https://ojs.uel.br/revistas/uel/index.php/semagrarias
Acta Scientiarum Agronomy	https://periodicos.uem.br/ojs/index.php/ActaSciAgron
Bragantia	https://www.scielo.br/j/brag/
Scientia Agrícola	https://www.scielo.br/j/sa/
Pesquisa Agropecuária Brasileira	https://seer.sct.embrapa.br/index.php/pab
Ciência Rural	https://www.scielo.br/j/cr/
Ciência e Agrotecnologia	https://www.scielo.br/j/cagro/
Weed Science	https://www.scielo.br/j/aws/
Weed Technology	https://www.cambridge.org/core/journals/weed-technology
Bioscience Journal	https://seer.ufu.br/index.php/biosciencejournal
Crop Protection	https://www.sciencedirect.com/journal/crop-protection
Crop Science	https://onlinelibrary.wiley.com/journal/1439037X
Agriculture	https://www.mdpi.com/journal/agriculture
⋮	⋮
⋮	⋮

Fonte: The authors

verification of whether the results remain consistent across different agricultural years. Moreover, in a way, it would also be possible to increase/verify the reliability of the results.

In this sense, the trials may not always be conducted identically, either due to random errors, variations in management, or even the execution of the experiments. Discrepancies may arise in experimental designs, data collection, and the occurrence of measurement errors. Such inconsistencies may also occur due to trials being conducted by different individuals or other factors leading to variations in execution. Additionally, the classic "measurement error" may occur during data collection, not to mention the various confounding factors that may affect data analysis.

As a result, when two experiments are conducted with some discrepancies (not caused by chance), many researchers are required to conduct at least a third experiment to ensure, at the very least, the submission of the work to certain scientific journals. This can lead to several issues: additional costs, extended time to complete the research, frequent changes in research teams impacting activities such as undergraduate research, master's, doctoral, and post-doctoral studies. These challenges are further exacerbated by the expiration of research grants (which have fixed durations) and budgetary limitations for a new trial.

In this context, equipped with two datasets fitting this exact problem, the team responsible for this thesis proposed an approach enabling researchers to evaluate whether there are significant differences between the results of the two conducted experiments. For this motivation 2, the difference between the two conducted trials lies in the method of data collection.

In theory, this approach could save time, financial resources, and effort, potentially facilitating the publication of the study in a scientific journal in the field of agricultural sciences.

Thus, motivation 2 involved modeling two datasets. Specifically, the results achieved by modeling the first dataset (where data were collected differently than agreed upon) will be used as informative priors in the modeling of the second dataset (second experiment, where the data were collected as agreed upon in our meetings), which will form the solution for motivation 2. Besides providing something essential for a Bayesian approach: informative priors, the modeling of the first dataset will also deliver novel results for the agronomic field: identifying orange crop genotypes resistant to *citrus canker*. Although in the first trial the data were collected incorrectly, over time, the trends in the data within the generated graphs yield positive and valuable results for agronomy.

The idea is to compare the posterior distributions of the two modelings. If they do not show significant differences/distances, we can conclude that the experimental results are statistically equivalent at a given significance level.

1.2.3 Motivation 3

Motivation 3 is related to the *citrus canker* disease and consists of finding rootstock combinations of orange varieties that are resistant to this disease.

The first recorded reports of this disease date back to 1827 to 1831 in India (FAWCETT; JENKINS, 1933). This disease is caused by the bacterium *Xanthomonas citri subsp. citri* (SCHAAD *et al.*, 2006) and causes significant damage to citrus production worldwide, resulting in losses for rural producers as well as industries (GOTTWALD *et al.*, 2002; VILORIA *et al.*, 2004). Surrounded by a yellow halo, in addition to the fruit itself, citrus canker affects branches, stems, and leaves of citrus trees. The disease begins with necrosis, which can cause cracks in the fruit, providing an entry point for other diseases and pests, potentially leading to fruit drop (PADMANABHAM; VIDHYASEKARAN; RAJAGOPALAN, 1973).

There are several ways for the bacterium that causes *citrus canker* to develop in orange plants. These include attacks by pests that spread the disease (such as the citrus leaf miner (*Phyllocnistis citrella*) GOTTWALD, GRAHAM and SCHUBERT (1997)), as well as wind BEHLAU *et al.* (2008), contaminated agricultural machinery, tools, and materials used in crop management, among others. So far, the most effective means of controlling the disease is the use of agricultural pesticides (GOTTWALD *et al.*, 2002; BOCK; PARKER; GOTTWALD, 2005). Among these, the most effective pesticides are copper-based formulations that protect the tissue of young leaves and reduce the bacterial load on plants (BEHLAU *et al.*, 2010). However, copper is a toxic element that accumulates in the environment, contaminating soils, rivers, and potentially reaching groundwater (DEWDNEY; GRAHAM, 2018).

The most sustainable and efficient method that agriculture has developed for controlling

plant diseases across various crops and their associated diseases is the development of genetically resistant plants. Plant breeding programs, combined with plant pathology programs, have been conducting tireless research in this direction (POMPEU; BLUMER, 2006). This would be the best solution for controlling citrus canker, given the type of agriculture encouraged today (VILORIA *et al.*, 2004).

In this context, and knowing that orange varieties have different resistances and behaviors against citrus canker, and that their combinations can even generate stronger descendants, this third challenge consists of analyzing experimental data from agronomic trials aimed at identifying rootstock combinations resistant to *citrus canker*. There is something particular about this dataset that makes modeling more challenging: it contains a large number of zeros.

1.2.4 The Orange Crop

Given that the three main motivations of this thesis involve a single agricultural crop, it is important to emphasize it: the orange crop. Brazil is the largest producer of oranges in the world, as well as the world's largest exporter of orange juice Netto and Regina (2023), annually exporting approximately 80% of the orange juice traded globally. In the 2020/2021 harvest, Brazil accounted for approximately 33% of the global production of this *commodity* and nearly two-thirds of the total juice volume produced worldwide (VIDAL, 2021). Just from the export of orange juice produced in Brazil, nearly US$ 1.5 billion is exported annually. It is forecasted that orange production will increase from 16.4 million tons in the 2022/2023 harvest to 17.2 million tons in 2031/32. Production is expected to grow annually by approximately 0.5% over the next decade. However, compared to past decades, there has been a decline in both production and cultivated area. The state of São Paulo is Brazil's largest producer, leading the ranking with 75% of the national production. Furthermore, in the last harvest, the state produced an average of 31 tons per hectare cultivated, while the national average was 26.7 tons per hectare (GASQUES *et al.*, 2022).

The orange belongs to the genus *Citrus*, the family *Rutaceae*, and the species *Citrus sinensis* (RIBEIRO, 2010). There is strong evidence that the orange fruit originated in antiquity through a cross between pomelo and mandarin. Historically, there are records and indications that citrus fruits originated in India, the Himalayas, and East Asia, and were later taken from there to Africa, Europe, and consequently the rest of the world. In the Americas, they arrived in 1500 with the Portuguese (CRISÓSTOMO; NAUMOV, 2009).

This plant is one of the variations within the citrus crop. Within citrus cultivation, there are lemons, limes, mandarins, oranges, and other variations. In Brazil, among citrus fruits, the orange crop holds the greatest economic importance. What most concerns the various sectors working with this fruit are citrus diseases, as they are the main factors affecting its production. Citrus plants, in general, can be affected by bacterial, fungal, and viral diseases. Thus, investments and scientific developments that support its advancement are indispensable.

The citrus diseases causing the most significant damage are: Greening (caused by the bacterium *Candidatus Liberibacter spp*), Citrus Canker (caused by the bacterium *Xanthomonas citri*), Blossom End Rot (caused by the fungus *Colletotrichum spp*), and Citrus Leprosis (caused by the *Citrus leprosis virus*) (AMORIM; REZENDE; CAMARGO, 1997). Among these four, the two most dangerous are those caused by bacteria (Greening and Citrus Canker) (ALVES[1]; BELOTI[1] *et al.*, 2014).

Thus, proposing solutions for the problems and losses caused by these two diseases is crucial for the production chain of this agricultural crop.

CHAPTER

2

DEVELOPMENT OF AN AUTOMATIC PLATFORM FOR SAMPLING PLANS IN THE DETECTION OF GREENING IN ORANGE ORCHARDS: A HIERARCHICAL APPROACH WITH BETA-BINOMIAL AND FLEXSHAPE-BINOMIAL MODELS.

2.1 Methodology

From meetings among the research team members, an understanding of the areas of knowledge involved was developed to initiate the construction of sampling plans. Based on the disease proportion, how it arises and spreads in the orchards, among other questions raised, studies on statistical theories were undertaken, along with the development of sampling plans. In other words, given the problem, theoretical studies and the implementation of the practical part of this work were carried out. For this purpose, some probability distributions were used: Bernoulli, Binomial, Beta, FlexShape, Beta-Binomial, and FlexShape-Binomial. After all studies and analyses, the implementation phase resulted in an online platform capable of generating sampling plans that detect the proportion of greening disease infection in orange orchards.

With this knowledge in hand, the construction of maps for the orange farms began. Initially, the Agência de Defesa Agropecuária do Paraná (ADAPAR) provided the central coordinates of one of the farms. Using Google Earth, its location was identified to begin the map construction. Geographic coordinates were collected, and a grid was overlaid using R software, based on desirable assumptions that met criteria established in meetings (with theoretical statistical and agronomic foundations). Among the assumptions was the construction

of the first sampling plan for a population of $1,000$ orange trees. The farm was divided into plots of $1,000$ plants each. Consequently, it was decided to divide each area, containing $1,000$ plants, into four clusters of 250 plants each. The objective, then, was to determine the number of individuals to be observed in each cluster containing 250 plants. Assuming that the disease dispersion follows a Beta-Binomial distribution, the sample sizes to be randomly drawn and collected within each cluster were calculated. The same procedure was then performed using the FlexShape-Binomial probability distribution.

2.1.1 Beta-Binomial Distribution

As we are dealing with modeling bacterial diseases in orange trees (large, medium, or small), when we look at just one individual, we may want to model the occurrence or absence of the respective disease in that plant. When we are trying to model a set of them, the aggregation of the situation mentioned earlier culminates in models composed of the methodology used in the situation referred to at the beginning of this paragraph.

In this sense, since the visual inspection of the evaluated plant is categorical—confirming the presence or absence of the disease—each individual inspection can be treated as a random variable X that represents the outcome of a Bernoulli trial. That is, we say that $X = 1$ if the presence of the disease is confirmed, and $X = 0$ otherwise.

Defining that the probability of the disease being confirmed is equal to a parameter $p \in (0,1)$, and that the probability of non-confirmation is equal to its complement, i.e., $1-p \in (0,1)$, the model is established.

$$X \mid p \sim \text{Bernoulli}(p) \qquad \text{with} \qquad \begin{array}{c} X \mid p \in \{0,1\}, \\ 0 < p < 1. \end{array} \tag{2.1}$$

The Bernoulli Model has a probability function defined by

$$\Pr(X = x \mid p) = p^x (1-p)^{1-x} \qquad \text{with} \qquad \begin{array}{c} x = 0 \quad \text{ou} \quad x = 1, \\ 0 < p < 1. \end{array}$$

The expected value and variance of a Bernoulli random variable are expressed as follows:

$$\mathbb{E}(X \mid p) = p \qquad \text{and} \qquad \text{Var}(X \mid p) = p(1-p) \qquad \text{with} \qquad 0 < p < 1.$$

If a quantity equal to $n \in \mathbb{N}$ of plants is collected in a given planting area, the set of conclusions obtained from each individually collected plant constitutes a model that can represent the total number of positive or negative confirmations of the random variable $X \mid p$. This new quantity (of positive confirmations)

$$Y \mid p = X_1 \mid p + X_2 \mid p + \cdots + X_n \mid p,$$

can be statistically defined as the realization of a binomial random variable. That is, the number of positive confirmations $Y \mid p$ in a known group of n plants is defined by the model,

$$Y \mid p \sim \text{Binomial}(p,n) \qquad \text{with} \qquad \begin{array}{c} Y \in \{0,1,2,\ldots,n\}, \\ n \in \mathbb{N} \quad \text{and} \quad 0 < p < 1. \end{array} \tag{2.2}$$

The Binomial Model has a probability function defined by

$$\Pr(Y = y \mid p) = \binom{n}{y} p^y (1-p)^{n-y} \qquad \text{with} \qquad \begin{array}{c} y = 0,1,2,\ldots,n, \\ n \in \mathbb{N} \quad \text{and} \quad 0 < p < 1. \end{array}$$

The expected value and variance of a random variable Y with a Binomial distribution are expressed as follows,

$$\mathbb{E}(Y \mid p) = np \quad \text{and} \quad \text{Var}(Y \mid p) = np(1-p) \qquad \text{with} \qquad \begin{array}{c} n \in \mathbb{N}, \\ 0 < p < 1. \end{array}$$

In a practical experiment, it is common for infections not to be distributed uniformly across a planting area; that is, in many cases, diseases exhibit focal aggregation because the infection starts from a central point. Consequently, thinking of individuals throughout a planting area as an idealized statistical population under the assumption of the existence of *clusters* (for reasons of sampling adequacy, practicality, or even convenience) must take into account that the proportions of infected plants in a group of regions do not remain constant across them. That is, statistically, it is equivalent to considering a proportion that is randomly distributed over the planted area.

A suitable model for rates and proportions, such as the proportion P, is the beta model, denoted by,

$$P \sim \text{Beta}(\alpha, \beta) \qquad \text{with} \qquad \begin{array}{c} 0 < P < 1, \\ \alpha > 0 \quad \text{and} \quad \beta > 0. \end{array} \tag{2.3}$$

The Beta Model has a probability density function expressed by

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \qquad \text{with} \qquad \begin{array}{c} 0 < p < 1, \\ \alpha > 0 \quad \text{and} \quad \beta > 0. \end{array}$$

The expected value and variance of a random variable P with a Beta distribution are given by,

$$\mathbb{E}(P) = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \text{Var}(P) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{with} \qquad \begin{array}{c} \alpha > 0, \\ \beta > 0. \end{array}$$

If a researcher is interested in understanding a set of the previous ideas, a suitable model for a random variable that combines the binomial and beta models is the Beta-Binomial model. The Beta-Binomial distribution, being a composition resulting from the binomial and beta distributions, is a natural extension of a binomial model in situations where it is assumed that the frequency of the parameter has a beta distribution. Therefore, a random variable $Y|P$, that is, Y conditionally independent of the random variable P, describes the total number of plants for which the disease confirmation is positive, considering that the proportion P of diseased plants is not constant.

To obtain the probability function for $Y|P$, it is necessary to determine the marginal probability of Y based on the joint probability density function of the random vector (Y,P). This function is expressed by,

$$f_{Y,P}(y,p) = \Pr(Y = y \mid p) f(p) \qquad \text{with} \qquad \begin{array}{c} y = 0,1,\ldots,n,\ 0 < p < 1, \\ n \in \mathbb{N},\ \alpha > 0,\ \beta > 0. \end{array}$$

where $\Pr(Y = y \mid p)$ and $f(p)$ are the probability and probability density functions of the Binomial and Beta models, respectively. The result of this product is given by,

$$f_{Y,P}(y,p) = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} \qquad \text{with} \qquad \begin{array}{c} y = 0,1,\ldots,n, 0 < p < 1, \\ n \in \mathbb{N}, \alpha > 0, \beta > 0. \end{array}$$

The marginal probability function for Y is determined by integrating $f_{Y,P}(y,p)$ over the entire parameter space of p, that is, the interval $(0,1)$, thus,

$$\begin{aligned} \Pr(Y = y) &= \int_0^1 f_{Y,P}(y.p)\mathrm{p} \\ &= \int_0^1 \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}\mathrm{p} \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{y+\alpha-1}(1-p)^{n-y+\beta-1}\mathrm{p} \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)}, \end{aligned}$$

and defines the model,

$$Y \sim \text{Beta-Binomial}(\alpha,\beta,n), \qquad \text{with} \qquad \begin{array}{c} y = 0,1,\ldots,n, \\ n \in \mathbb{N}, \alpha > 0, \beta > 0. \end{array}$$

Using the properties of conditional expectation and variance, the expectation and variance of a random variable with a Beta-Binomial distribution are obtained. Note that,

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|P)] = \mathbb{E}[nP] = \frac{n\alpha}{\alpha+\beta} \qquad \text{with} \qquad n \in \mathbb{N}, \alpha > 0, \beta > 0.$$

and that,

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[\,\mathrm{Var}(Y|P)\,] + \mathrm{Var}[\,\mathbb{E}(Y|P)\,] \\
&= \mathbb{E}[nP(1-P)] + \mathrm{Var}[nP] \\
&= n\{\ \mathbb{E}(P) - \mathbb{E}(P^2) + n\mathrm{Var}(P)\ \} \\
&= n\{\ \mathbb{E}(P) - \mathrm{Var}(P) - [\mathbb{E}(P)]^2 + n\mathrm{Var}(P)\ \} \\
&= n\left\{ \frac{\alpha}{\alpha+\beta} - \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} - \left[\frac{\alpha}{\alpha+\beta}\right]^2 + \frac{n\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \right\} \\
&= \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{with} \qquad n \in \mathbb{N}, \alpha > 0, \beta > 0.
\end{aligned}$$

Thus, the expected value and variance of a random variable Y with a Beta-Binomial distribution with parameters α and β, given a known n, are given by,

$$\mathbb{E}(Y) = \frac{n\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{with} \qquad n \in \mathbb{N}, \alpha > 0, \beta > 0.$$

Observe in Figure 4 the general shape of the probability function for a random variable with a Beta-Binomial distribution with parameters $\alpha = 0.5; 1; 2; 5$ and $\beta = 0.5; 1; 2; 5$, and $n = 30$. In general, these shapes are also assumed by the beta distribution, such as the *bathtub* shapes when $0 < \alpha < 1$ and $0 < \beta < 1$; J or inverted J shapes with varying intensity when $\alpha > \beta$ and $0 < \beta < 1$ or $\beta > \alpha$ and $0 < \beta < 1$, respectively; *constant* when $\alpha = \beta = 1$; *linearly increasing* or *linearly decreasing* when α is a multiple of β or β is a multiple of α, respectively; *symmetric bell* when $\alpha = \beta > 1$ and *asymmetric bell to the left or right* when $\alpha > \beta = 1$ or $\beta > \alpha > 1$, respectively.

In an experimental context, the parameters α and β of the Beta-Binomial distribution do not have a direct practical meaning. Some authors, such as Hughes, Madden & Munkvold Hughes, Madden and Munkvold (1996), suggest a parameterization that considers the parameters p and ρ, expressed by,

$$p = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \rho = \frac{1}{\alpha+\beta+1}.$$

As a consequence of this choice, the expected value and variance of the random variable can be rewritten as,

$$\mathbb{E}(Y) = np \qquad \text{and} \qquad \mathrm{Var}(Y) = np(1-p)[1+(n-1)\rho],$$

In other words, now the parameter p, when multiplied by n, directly represents the expected value of Y. On the other hand, note that the variance of Y is equal to the product of the variance of the binomial by a number that is always positive and dependent on the parameter ρ. That is,

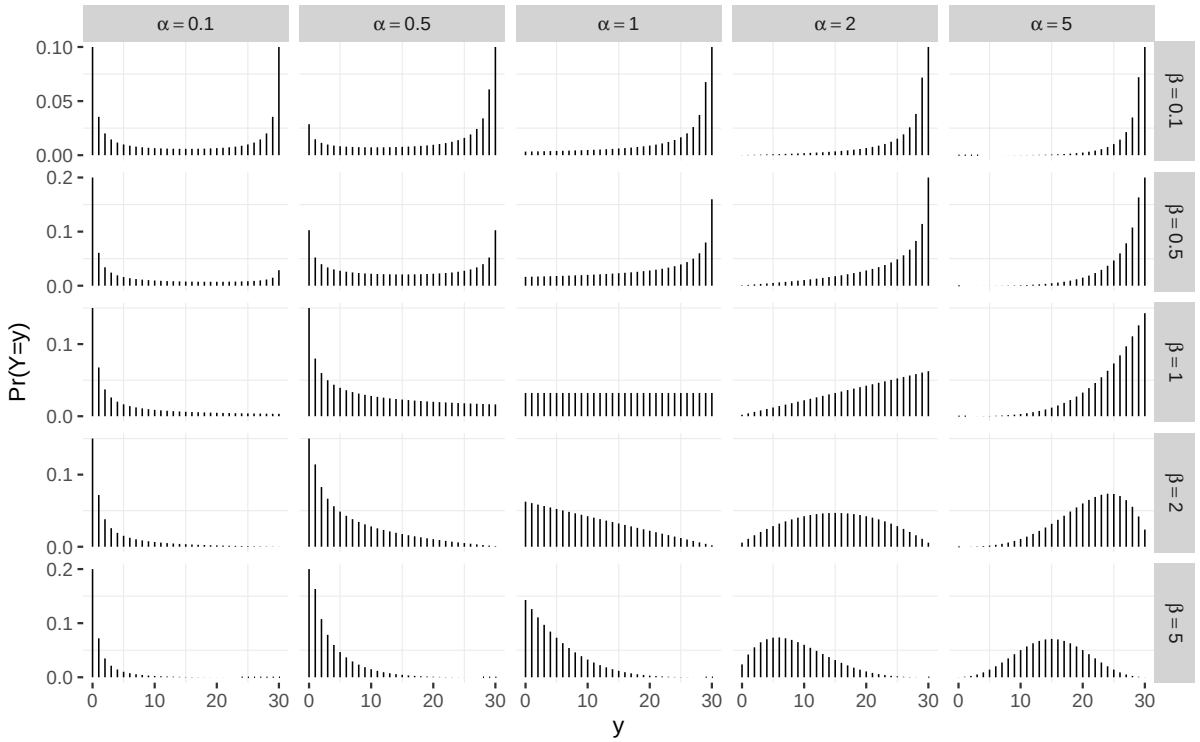


Figure 1 – Illustration of the behavior of the probability function of a Beta-Binomial distribution for different values of the parameters $\alpha = 0.5; 1; 2; 5$ and $\beta = 0.5; 1; 2; 5$ and $n = 30$.

the parameter ρ now indicates a factor of inflation for the binomial variance, or a measure for the *degree of aggregation* of the proportion of infected plants.

An interesting characteristic of the parameter ρ is that when $\rho \to 0$, the variance of Y tends to the binomial variance. In the limit, this means that there is no aggregation in the infections (HUGHES; MADDEN; MUNKVOLD, 1996; ENNIS; BI, 1998). Figure 2 illustrates this approximation well.

2.1.2 FlexShape-Binomial Distribution

Instead of considering the Beta distribution for the random variable P in the hierarchy $Y|P$, it is possible to consider any other unit distribution of interest. Nascimento *et al.* (2023) proposed the FlexShape probability distribution and illustrated it through analyses of proportion data. Such data were treated from a quality control perspective. The data modeled with this distribution belong to the area of water particle monitoring (for rainfall monitoring) from a station in the Atacama Desert - Chile, known as the driest desert in the world.

It is said that P has a FlexShape distribution with parameters μ and σ, denoted by $P \sim \text{FS}(\mu, \sigma)$, if the probability density function of P has the form,

$$f(p) = \frac{30}{5+\sigma}\left[\sigma(4p^2-4p+1)+1\right]\left[\mu(2p-1)+1\right]p(1-p) \qquad \text{with} \qquad 0 < p < 1.$$

com $-1 \leq \mu \leq 1$ and $\sigma \geq 0$.

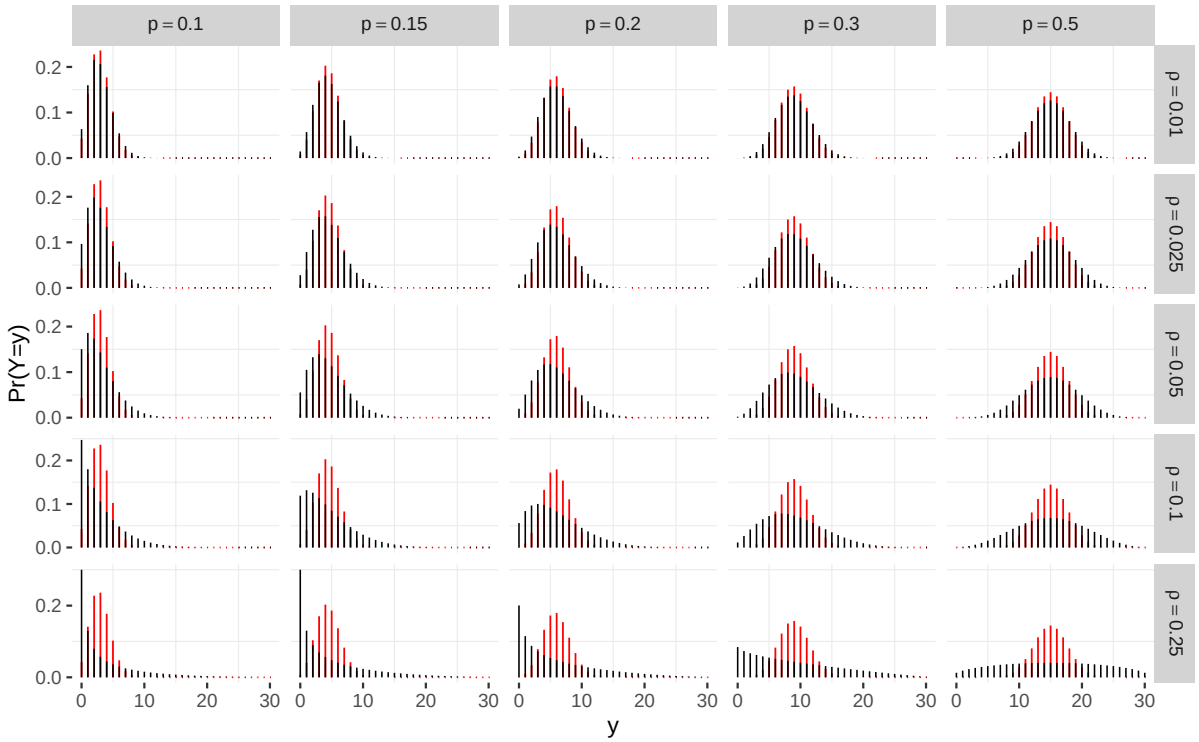


Figure 2 – Comparison between the behaviors of the binomial probability functions (in red) and Beta-Binomial (in black) for different values of the parameters $p = 0.1; 0.15; 0.2; 0.3; 0.5$ and $\rho = 0.01; 0.025; 0.05; 0.1; 0.25$.

These authors Nascimento *et al.* (2023) claim that they have developed a new probability distribution, which, so far, according to the literature in the field, is a new unit distribution $(0,1)$ that can accommodate asymmetry and, to the best of their knowledge, also asserts that it is the first distribution of this nature capable of accommodating bimodality. Figure 3 presents some illustrations of the behavior of this distribution.

The r-th moment $\mathbb{E}[P^r]$ is given by,

$$\mathbb{E}[P^r] = \frac{1}{2^r}\sum_{j=0}^{r}\binom{r}{j}E[\gamma^j], \qquad com \qquad r = 1,2,3,\ldots,$$

where

$$\mathbb{E}[\gamma^{2h-1}] = \frac{15\mu}{2(5+\sigma)}\left[\frac{1}{2h+1} + \frac{\sigma-1}{2h+3} - \frac{\sigma}{2h+5}\right],$$

and

$$\mathbb{E}[\gamma^{2h}] = \frac{15}{2(5+\sigma)}\left[\frac{1}{2h+1} + \frac{\sigma-1}{2h+3} - \frac{\sigma}{2h+5}\right],$$

with $h = 1,2,3,\ldots$

It has mean and variance

$$\mathbb{E}(P) = \frac{(3\mu+7)\sigma+7\mu+35}{70+14\sigma} \quad \text{and} \quad \text{Var}(P) = \frac{(7+3\sigma)(35+7\sigma-3\mu^2\sigma-7\mu^2)}{196(5+\sigma)^2},$$

with $-1 < \mu < 1$ and $\sigma > 0$.

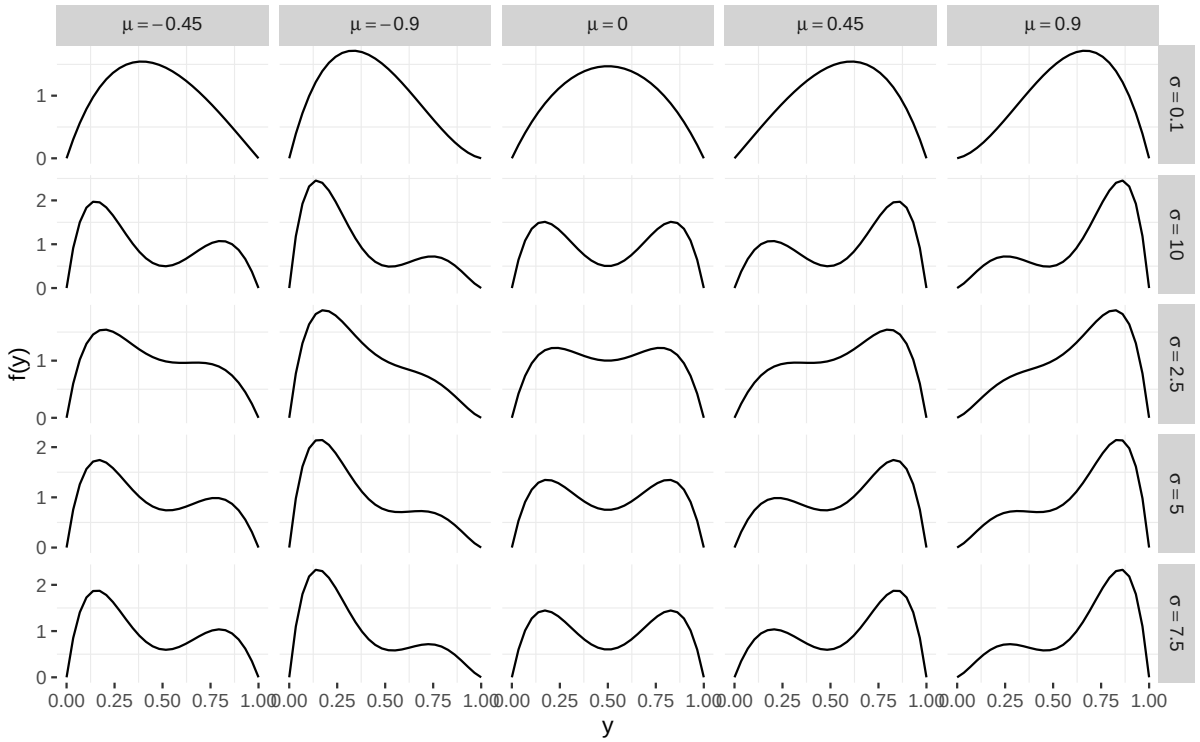


Figure 3 – Behavior of the FlexShape distribution for different combinations of the skewness parameter $(-0.9 \leq \mu \leq 0.9)$ and the bimodality parameter $(0.1 \leq \sigma \leq 10)$.

To obtain the probability function for $Y|P$, it is necessary to determine the marginal probability of Y based on the joint probability density function of the random vector (Y,P). This function is expressed as,

$$f_{Y,P}(y,p) = \Pr(Y = y \mid p) f(p) \qquad \text{with} \qquad \begin{array}{c} y = 0,1,\ldots,n, 0 < p < 1, \\ n \in \mathbb{N}, \alpha > 0, \beta > 0. \end{array}$$

where $\Pr(Y = y \mid p)$ and $f_{Y,P}(y,p)$ are the probability and probability density functions of the Binomial and FlexShape models, respectively. The result of this product is given by,

$$f_{Y,P}(y,p) = \binom{n}{y} \frac{30}{5+\sigma} \left[\sigma(4p^2 - 4p + 1) + 1 \right] \left[\mu(2p-1) + 1 \right] p^{y+1} (1-p)^{n-y+1},$$

with $y = 0,1,\ldots,n$, $0 < p < 1$, $n \in \mathbb{N}$, $-1 < \mu < 1$ and $\sigma > 0$.

The marginal probability function for Y is determined by integrating $f_{Y,P}(y,p)$ over the

entire parameter space of p, that is, the interval $(0,1)$, thus,

$$\begin{aligned}
\Pr(Y=y) &= \int_0^1 f_{Y,P}(y,p)\dot{\mathrm{p}} \\
&= \int_0^1 \binom{n}{y}\frac{30}{5+\sigma}\big[\,\sigma(4p^2-4p+1)+1\,\big]\big[\,\mu(2p-1)+1\,\big]p^{y+1}(1-p)^{n-y+1}\dot{\mathrm{p}} \\
&= \binom{n}{y}\frac{30}{5+\sigma}\int_0^1 \big[\,\sigma(4p^2-4p+1)+1\,\big]\big[\,\mu(2p-1)+1\,\big]p^{y+1}(1-p)^{n-y+1}\dot{\mathrm{p}} \\
&= \binom{n}{y}\frac{30\sigma}{5+\sigma}\frac{\Gamma(n-y+2)\Gamma(y+2)}{-\Gamma(7+n)}\Bigg\{(\mu-1)n^3+\Big[(4-6\mu)y+3\mu-7\Big]n^2+ \\
&\qquad \Big[(12\mu-4)y^2+(24-6\mu)y+14\mu-10\Big]n-8\mu y^3-24y^2-28\mu y-24\Bigg\},
\end{aligned}$$

and defines the model,

$$Y\sim \text{FlexShape-Binomial}(\mu,\sigma,n), \qquad \text{with} \qquad \begin{array}{c} y=0,1,\ldots,n, \\ n\in\mathbb{N}, -1<\mu<1, \sigma>0. \end{array}$$

Using the properties of conditional expectation and variance, one can obtain the expectation and variance of a random variable with a Beta-Binomial distribution. Note that,

$$\mathbb{E}(Y)=\mathbb{E}[\mathbb{E}(Y|P)]=\mathbb{E}[nP]=\frac{n\big[\,(3\mu+7)\sigma+7\mu+35\,\big]}{70+14\sigma} \qquad \text{with} \qquad \begin{array}{c} n\in\mathbb{N} \\ -1<\mu<1, \sigma>0. \end{array}$$

and that,

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[\,\mathrm{Var}(Y|P)\,]+\mathrm{Var}[\,\mathbb{E}(Y|P)\,] \\
&= \mathbb{E}[\,nP(1-P)\,]+\mathrm{Var}[\,nP\,] \\
&= n\big\{\ \mathbb{E}(P)-\mathbb{E}(P^2)+n\mathrm{Var}(P)\ \big\} \\
&= n\big\{\ \mathbb{E}(P)-\mathrm{Var}(P)-[\,\mathbb{E}(P)\,]^2+n\mathrm{Var}(P)\ \big\} \\
&= n\Bigg\{\frac{(3\mu+7)\sigma+7\mu+35}{70+14\sigma}-\frac{(7+3\sigma)(35+7\sigma-3\mu^2\sigma-7\mu^2)}{196(5+\sigma)^2}- \\
&\qquad \left[\frac{(3\mu+7)\sigma+7\mu+35}{70+14\sigma}\right]^2+\frac{n(7+3\sigma)(35+7\sigma-3\mu^2\sigma-7\mu^2)}{196(5+\sigma)^2}\Bigg\} \\
&= \frac{9n\big\{28(\sigma+7)(5+\sigma)/9-(\sigma+7/3)\left[\left(\mu^2-7/3\right)\sigma+7(\mu^2-35)/3\right]n\big\}}{196\,(5+\sigma)^2},
\end{aligned}$$

with $n \in \mathbb{N}$, $-1 < \mu < 1$ and $\sigma > 0$.

Observe in Figure 4 the general shape of the probability function for a random variable with FlexShape-Binomial distribution with parameters $\mu = -0.9; -0.45; 0; 0.45; 0.9$ and $\sigma = 0.1; 2.5; 5.0; 7.5; 10.0$. In a manner analogous to the Beta-Binomial case, these are also the shapes assumed by the FlexShape distribution.

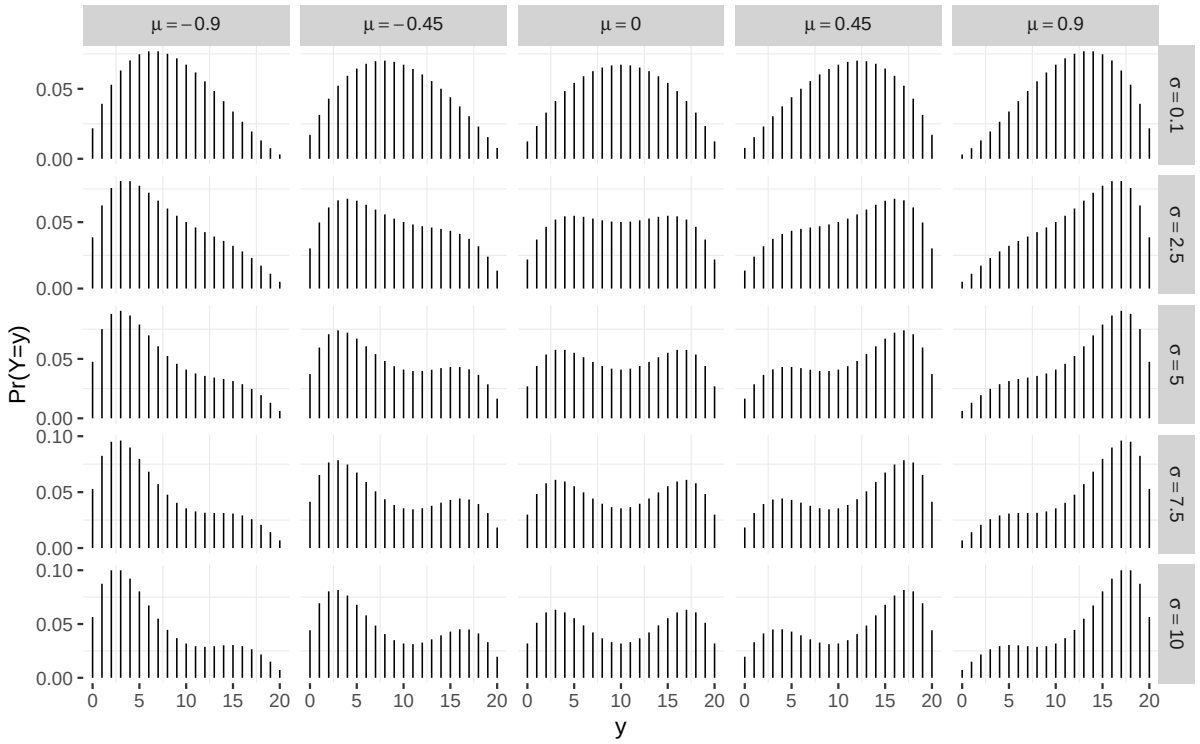


Figure 4 – Illustration of the behavior of the probability function of a FlexShape-Binomial distribution for different values of the parameters $\mu = -0.9; -0.45; 0; 0.45; 0.9$ and $\sigma = 0.1; 2.5; 5.0; 7.5; 10.0$ and $n = 20$.

Again, in an experimental context, the parameters μ and σ of the FlexShape-Binomial distribution do not have a direct practical meaning. In this sense, a parametrization is also suggested that considers the parameters p and ρ, expressed as,

$$p = \frac{(3\mu+7)\sigma+7\mu+35}{70+14\sigma} \qquad \text{and} \qquad \rho = \frac{(7+3\sigma)\left(3\mu^2\sigma+7\mu^2-7\sigma-35\right)}{(3\mu\sigma+7\mu+7\sigma+35)(3\mu\sigma+7\mu-7\sigma-35)},$$

As a consequence of this definition, the expected value and variance of the random variable can also be rewritten as,

$$\mathbb{E}(Y) = np \qquad \text{and} \qquad \text{Var}(Y) = np(1-p)[1+(n-1)\rho],$$

that is, now the parameter p when multiplied by n directly represents the expected value of Y. On the other hand, note that the variance of Y is equal to the product of the binomial variance by a number that is always positive and dependent on the parameter ρ. In other words, the parameter ρ now indicates a factor of inflation for the binomial variance, or even a measure of the *degree of aggregation* of the proportion of infected plants.

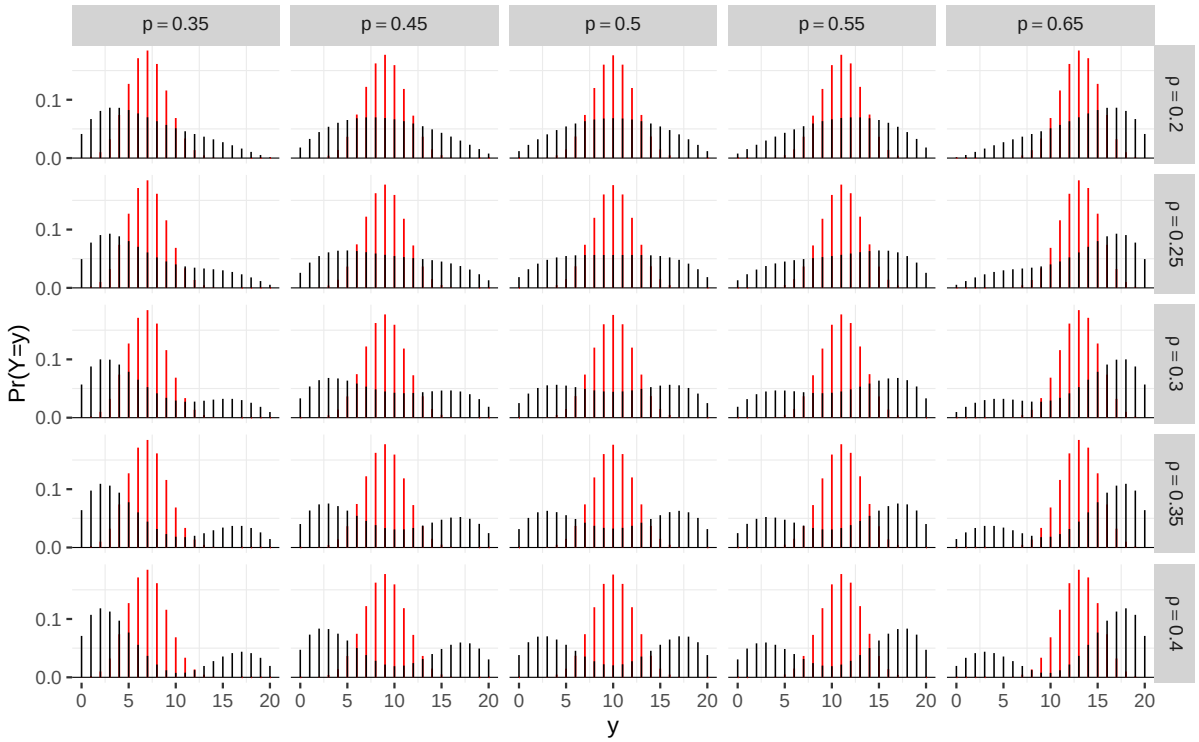


Figure 5 – Comparison of the behaviors of the Binomial probability functions (in red) and FlexShape-Binomial (in black) for different values of the parameters $p = 0.1; 0.15; 0.2; 0.3; 0.5$ and $\rho = 0.01; 0.025; 0.05; 0.1; 0.25$.

2.1.3 Determination of sample size

Considering the central context of this study, which deals with obtaining sampling plans, it is necessary to consider analytical expressions for the estimators of p and ρ. In this sense, if a random sample of size nN is considered, consisting of N independent *clusters* of size n, the number of occurrences Y_j in the j-th *cluster*, where $j = 1, \ldots, N$, is computed. Thus, a set of independent observations $(y_j, n), j = 1, \ldots, N$ is organized, which constitutes a random sample of size N of realizations of a random variable with Beta-Binomial (or FlexShape-Binomial) distribution with parameters (p, ρ), assuming n is known.

The first sample moment, denoted by $\overline{Y}$, and the second sample central moment, denoted by S^2, are computed as follows:

$$\overline{Y} = \sum_{j=1}^{N} \frac{Y_i}{N}, \qquad \text{and} \qquad S^2 = \sum_{j=1}^{N} \frac{(Y_i - \overline{Y})^2}{N-1},$$

and then are equated to the theoretical first moment and the theoretical second central moment, respectively. Therefore, the estimators $\hat{p}$ and $\hat{\rho}$ for p and ρ, determined based on the method of moments, are the solutions to the following system of equations,

$$\begin{cases} \mathbb{E}(Y) &= \overline{Y}, \\ \text{Var}(Y) &= S^2 \end{cases} \qquad \text{or even} \qquad \begin{cases} n\hat{p} &= \overline{Y}, \\ n\hat{p}(1-\hat{p})\left[1+(n-1)\hat{\rho}\right] &= S^2. \end{cases}$$

From the first equation, it follows that

$$n\hat{p} = \overline{Y} = \frac{1}{N}\sum_{j=1}^{N} Y_i \qquad \Longleftrightarrow \qquad \hat{p} = \frac{1}{nN}\sum_{j=1}^{N} Y_i.$$

Substituting this result into the second equation, it follows that,

$$\begin{aligned} n\hat{p}(1-\hat{p})\left[1+(n-1)\hat{\rho}\right] &= S^2 \\ 1+(n-1)\hat{\rho} &= \frac{S^2}{n\hat{p}(1-\hat{p})} \\ (n-1)\hat{\rho} &= \frac{S^2}{n\hat{p}(1-\hat{p})} - 1 \\ (n-1)\hat{\rho} &= \frac{S^2 - n\hat{p}(1-\hat{p})}{n\hat{p}(1-\hat{p})} \\ \hat{\rho} &= \frac{S^2 - n\hat{p}(1-\hat{p})}{n\hat{p}(1-\hat{p})(n-1)}. \end{aligned}$$

Therefore, the solution of the system that determines the moment-based estimators for p and ρ has the following form

$$\hat{p} = \frac{1}{nN}\sum_{j=1}^{N} Y_i \qquad \text{and} \qquad \hat{\rho} = \frac{S^2 - n\hat{p}(1-\hat{p})}{n\hat{p}(1-\hat{p})(n-1)}.$$

In the work of Hughes, Madden, and Munkvold Hughes, Madden and Munkvold (1996), an approach is described to obtain the sample size based on certain reliability criteria, specifically, criteria are defined that take into account the coefficient of variation and the length of the confidence interval. In any case, it is necessary to have in mind an expression for the standard error (se) of the mean proportion estimator $\hat{p}$.

Keeping in mind the expression for $\hat{p}$, the variance of this estimator is determined as follows,

$$\begin{aligned}
\mathrm{Var}(\hat{p}) &= \mathrm{Var}\left(\frac{1}{nN}\sum_{j=1}^{N} Y_j\right) \\
&= \frac{1}{(nN)^2}\sum_{j=1}^{N}\mathrm{Var}(Y_j) \\
&= \frac{1}{(nN)^2}\sum_{j=1}^{N} np(1-p)\left[1+(n-1)\rho\right] \\
&= \frac{1}{(nN)^2} nNp(1-p)\left[1+(n-1)\rho\right] \\
&= \frac{p(1-p)\left[1+(n-1)\rho\right]}{nN}.
\end{aligned}$$

Now, the standard error of the estimator $\hat{p}$ is defined as the square root of its variance, thus,

$$\mathrm{ep}(\hat{p}) = \sqrt{\frac{p(1-p)\left[1+(n-1)\rho\right]}{nN}}.$$

With a well-defined expression for the standard error of $\hat{p}$, one can establish the most interesting criterion to determine an adequate sample size, whether based on the Coefficient of Variation, the length of the Confidence Interval, or the Effect of Experimental Design.

2.1.3.1 Calculation Based on the Coefficient of Variation

If the criterion to be used is defined by the coefficient of variation, denoted by C, the relationship that provides the basis for determining the sample size is expressed by,

$$C = \frac{\mathrm{ep}(\hat{p})}{p}.$$

From this relationship, given an arbitrary coefficient of variation C_0, the following is obtained,

$$\begin{aligned}
C_0 &= \frac{1}{p}\sqrt{\frac{p(1-p)\left[1+(n-1)\rho\right]}{nN}} \\
{C_0}^2 &= \frac{1}{p^2}\frac{p(1-p)\left[1+(n-1)\rho\right]}{nN} \\
np{C_0}^2 &= \frac{(1-p)\left[1+(n-1)\rho\right]}{N} \\
N &= \frac{(1-p)\left[1+(n-1)\rho\right]}{np{C_0}^2}. \qquad (2.4)
\end{aligned}$$

2.1.3.2 Calculation Based on the Length of the Confidence Interval

If the criterion to be used is defined via confidence interval, then based on the central limit theorem, a confidence interval for $\hat{p}$ is written as,

$$\hat{p} \pm z_{\alpha/2}\text{ep}(\hat{p}),$$

where $z_{\alpha/2}$ is the upper quantile of the standardized normal distribution that accumulates a probability equal to $\alpha/2$. Essentially, there are two ways to use the length of the confidence interval as a basis for determining the sample size.

1) If it is desired to define half of the length of the confidence interval as a fixed proportion H of the mean proportion p, one can simply consider the equation,

$$z_{\alpha/2}\text{ep}(\hat{p}) = Hp.$$

From this expression, one obtains,

$$\begin{aligned} Hp &= z_{\alpha/2}\text{ep}(\hat{p}) \\ Hp &= z_{\alpha/2}\sqrt{\frac{p(1-p)\left[1+(n-1)\rho\right]}{nN}} \\ p^2\left(\frac{H}{z_{\alpha/2}}\right)^2 &= \frac{p(1-p)\left[1+(n-1)\rho\right]}{nN} \\ N &= \frac{(1-p)\left[1+(n-1)\rho\right]}{np}\left(\frac{z_{\alpha/2}}{H}\right)^2. \end{aligned} \tag{2.5}$$

2) On the other hand, if half the length of the confidence interval is defined as a fixed measure h, one simply needs to consider the equation,

$$z_{\alpha/2}\text{ep}(\hat{p}) = h.$$

from which one obtains,

$$\begin{aligned} h &= z_{\alpha/2}\text{ep}(\hat{p}) \\ h &= z_{\alpha/2}\sqrt{\frac{p(1-p)\left[1+(n-1)\rho\right]}{nN}} \\ \left(\frac{h}{z_{\alpha/2}}\right)^2 &= \frac{p(1-p)\left[1+(n-1)\rho\right]}{nN} \\ N &= \frac{p(1-p)\left[1+(n-1)\rho\right]}{n}\left(\frac{z_{\alpha/2}}{h}\right)^2. \end{aligned} \tag{2.6}$$

If it is possible to determine initial estimates for the parameters p and ρ, one can estimate the appropriate number of *clusters* to be sampled as suggested by the resulting expressions in (2.4), (2.5), or (2.6). Otherwise, it is also possible to carry out an initial estimation procedure and determine the sample size based on the ideas of Fosgate (FOSGATE, 2007).

2.1.3.3 Calculation Based on the Design Effect

This procedure for calculating a sample size takes into account a rudimentary estimate for the expected average and maximum proportion. These estimates are used to deduce the degree of correlation between the *clusters* and the effect caused by the design (DE). This method considers an effective sample size and updates it based on the assumptions of the beta-binomial model. The method consists of following these steps:

The researcher in the area of interest is asked about the average proportion, $\hat{p}$, that they believe to be the true value in the study population;

- If the reported proportion is less than or equal to 0.5, the researcher must indicate the percentile they believe accumulates 95% of the probability distribution, that is, the researcher should state the maximum proportion, $\hat{p}_{max}$, that they expect to find in reality;
- If the reported proportion is greater than 0.5, the researcher must indicate the percentile they believe accumulates 5% of the probability distribution, that is, the researcher should state the minimum proportion, $\hat{p}_{min}$, that they expect to find in reality.

These percentiles reflect the researcher's opinion about the expected variation in the proportion among the studied *clusters*, and the difference between these percentiles indicates the precision expected by the researcher, which corresponds to the measure h, and is defined by,

$$h = \hat{p}_{max} - \hat{p} \qquad \text{or} \qquad h = \hat{p} - \hat{p}_{min}.$$

A rudimentary estimate for the correlation among the *clusters* can be requested from the researcher (in this case, the number of *clusters* can be obtained directly from expressions like (2.6)). On the other hand, if this is not feasible, it is possible to determine an initial estimate as follows:

- The precision provided by the researcher corresponds to half of the confidence interval, which is equal to the product of the standard error and the confidence coefficient. The standard deviation of the chosen distribution for the proportion (Beta or FlexShape) will replace the standard error in the formula for the confidence interval. The standard deviation of a random variable X with a beta and flex-shape distribution with parameters p and ρ is expressed by,

$$\sqrt{\text{Var}(X)} = \sqrt{p(1-p)\rho},$$

where

$$p = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \rho = \frac{1}{\alpha+\beta+1},$$

for the Beta distribution and

$$p = \frac{3\mu\sigma+7\mu+7\sigma+35}{14(5+\sigma)} \quad \text{and} \quad \rho = \frac{(7+3\sigma)\left(3\mu^2\sigma+7\mu^2-7\sigma-35\right)}{(3\mu\sigma+7\mu+7\sigma+35)(3\mu\sigma+7\mu-7\sigma-35)},$$

for the Flex-Shape distribution.

Now, the standard deviation can be replaced by the provided precision and rewritten so that the equation can be solved for ρ, thus,

$$\tilde{\rho} = \frac{(h/z_{\alpha/2})^2}{\tilde{p}(1-\tilde{p})}.$$

The determination of an estimate for ρ indicates an approximation for the correlation between the *clusters*, and with it, it is possible to determine the inflation factor of the variance,

$$\mathrm{DE} = 1 + (n-1)\rho,$$

Considering an effective sample size n_E that represents the number of sample units assuming a non-*clustered* population, which contains statistical information equivalent to that which will be provided by the data from a *clustered* population. This quantity can be calculated as,

$$n_E = \frac{nN}{\mathrm{DE}}.$$

Note that with this expression, it is possible to determine what is most interesting for the researcher, either the appropriate number of *clusters* (if the size of each *cluster* is fixed and known) or the number of individuals to be sampled in each *cluster* (in the situation where all *clusters* are sampled).

$$\frac{n_E}{n}\mathrm{DE} = N, \qquad \text{or} \qquad \frac{n_E}{N}\mathrm{DE} = n.$$

Having an adequate number of *clusters* to be sampled, it is necessary to analyze the study area and establish the criteria that will highlight the existing *clusters*. This choice was made taking into account the considerations of the researchers who requested the outline of an appropriate sampling plan. After several meetings, a consensus was reached on how the sampling plans would be developed for the possible detection of the *Greening* disease in orange cultivation.

2.1.4 Web platform built and usage illustration

Based on the theoretical deductions made, a web plataform in R/Shiny was developed (available at this link <https://cemeai.shinyapps.io/Amostrador-Citrus/>) to facilitate sample planning for the areas of interest.

The agronomy engineers from the Agricultural Defense Agency of Paraná - ADAPAR, provided the central coordinates of one of the farms where there was interest in using the sampling methodology to be developed. This made it possible to begin the planning process. First, the farm was identified using *Google Earth*, where the area was outlined and exported to initiate the computational work. Based on this area, a sampling plan was developed.

Figure 6 – Study Farm. A lot containing orange trees with the plantation of interest. with geographic coordinates — Lat: 23° 07’ 43.2” S and Lon: 52° 12’ 37.91” – elevation of 501m and altitude equals 818m.

As confirmed by the agronomists at ADAPAR, there is a multitude of orange cultivars, each with its own requirements, ranging from different planting spacings to varying canopy sizes. Thus, the orange orchards are very diverse from one property to another; in most cases, areas of the same size do not have the same number of trees among different producers. Therefore, it was decided to work with plots containing one thousand plants. Following this requirement, each area of one thousand plants was divided into four clusters of 250 orange trees each, as shown in

Figure 6. Such suggestions were accepted, given that if an infection occurred in an area of 250 plants, it would likely be spread over a considerable number of surrounding plants. Regarding the division of areas into lots of 1000 plants, meeting ADAPAR's requirements, authors such as (TROYO *et al.*, 2008) recommend subdividing farms into smaller areas due to their large territorial extents.

Thus, with the work directed and restricted to the assumptions and requirements to be followed, and with the units defined, efforts continued through means that would make it possible to develop the sampling plans. Firstly, using *Google Earth*, the georeferenced coordinates that outline the area of the farm to be studied were obtained. These details should be entered into the platform created, in Field 5 (Figure 7).

- **Latitude points:**

  -23.1282411424187; -23.1286826117191; -23.1290885884515; -23.1293382361441;

  -23.1295554590774; -23.1294605731739; -23.129353897703; -23.129304004956;

  -23.1292565440262; -23.1292542500099; -23.129279776398; -23.1293313657747;

  -23.1294088566612; -23.12953420456; -23.1291474282998; -23.1288466781369;

  -23.1285846783308; -23.1283045702802; -23.1282345790479; -23.1281789506247;

  -23.1281460047817; -23.1281534093235; -23.1281841708567; -23.1282411424187.

- **Longitude points:**

  -52.2111611712132; -52.2111538910133; -52.2111505093702; -52.2111414521805;

  -52.2111300757091; -52.2110297752498; -52.2109048339988; -52.2107973058136;

  -52.2106650173041; -52.210528996305; -52.2103434510098; -52.2101947220154;

  -52.2100672267355; -52.2099267885012; -52.2099407320112; -52.2099504864915;

  -52.2099603247846; -52.2099791414985; -52.2101537283278; -52.21034428548;

  -52.2105852374098; -52.2107668399842; -52.2109641239278; -52.2111611712132.

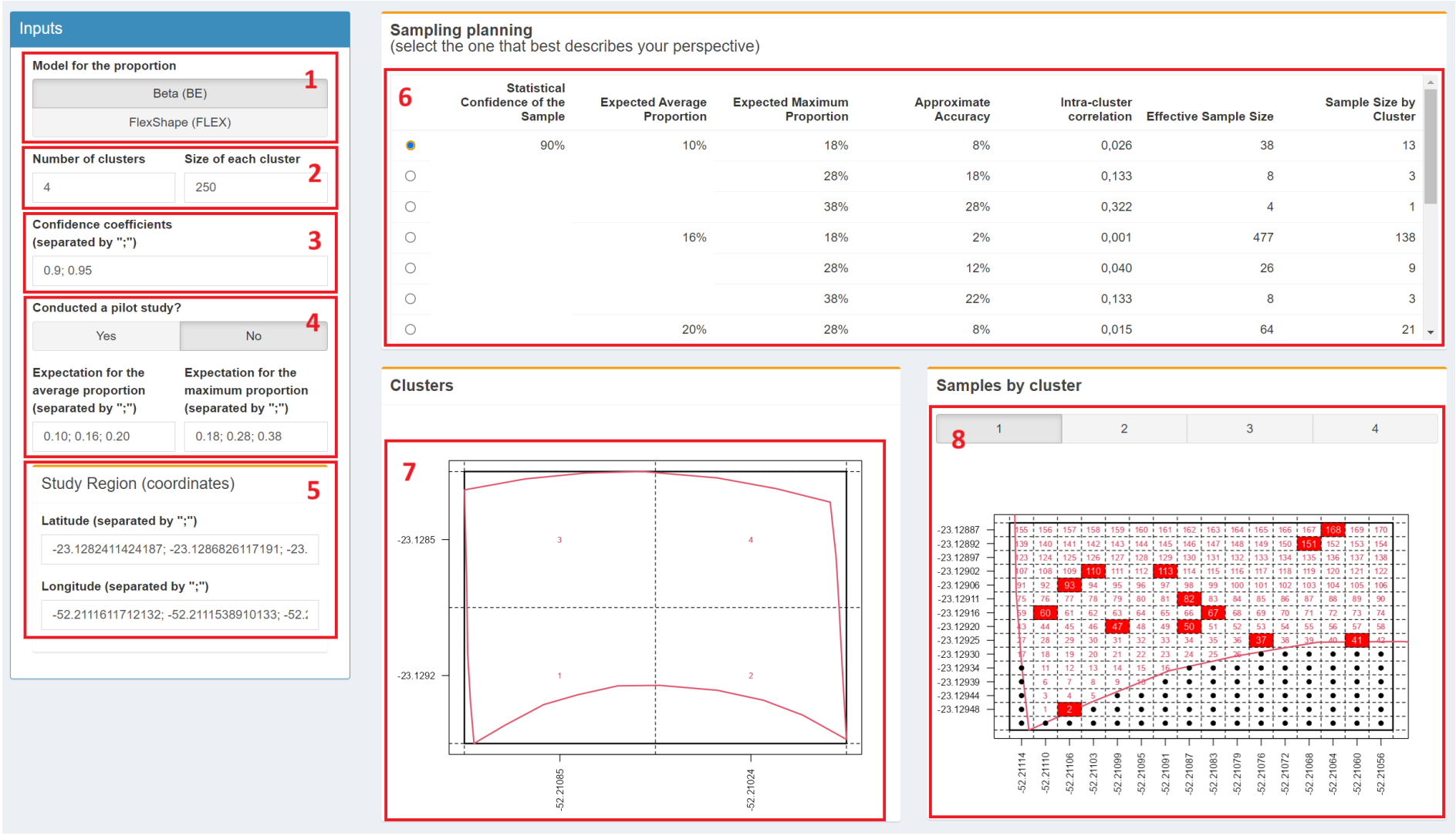


Figure 7 – Screenshot of the web platform.

According to the theoretical foundations of the fields of knowledge involved in this research, the information, the suggestions, and the expertise of the research group that planned and conducted the scientific research described in this work, the following criteria were established during the meetings that took place (during the decision-making and analysis period).

The sample sizes are calculated under the consideration that the analysis will be conducted on a population of 1000 plants partitioned into 4 *clusters*, therefore: $N = 4$. Each of the *clusters* consists of approximately 250 plants, therefore: $n = 250$. The total number of individuals that can be sampled within the population of 1000 plants is expressed by $nN = 1000$. These details should be entered into the platform, in Field 2 (Figure 7).

The aim is to determine the ideal quantity, n_0, of individuals that should be sampled in each of the *clusters* to establish a statistically significant sample. Then, a routine (on the Web Plataform) to identify the rectangle encompassing the previously obtained region. With the coordinates of the rectangle's vertices, this area was divided *clusters* (in this example, $N = 4$. The resulting division can be seen in the field Field 7 (Figure 7)), each containing approximately $n = 250$ plants (in this example). A code was implemented to identify each smaller rectangular area through its four vertices and its central point. With the coordinates of each *cluster* identified, the codes were applied again to each of the *clusters*, arbitrarily dividing them into a set of approximately $n = 250$ smaller regions, each potentially representing a tree. With all the coordinates in hand and a graph of the area and its divisions created, it was possible to construct the graphical visualization (the resulting can be seen in the field Field 8 (Figure 7)). As in previous stages, a code was implemented for this purpose. As an exclusion criterion, the central points of the smaller areas that did not belong to the study area were discarded. With each region eligible for sampling identified, it was possible to follow the suggestions of (LIN; POUSHINSKY; MAUER, 1979) and randomly choose which units would be collected.

To provide an overall view of the possible variations of the initial criteria that must be established by researchers in the agronomic field, some combinations of the requested estimates were considered. The result can be seen in the platform, in Field 6 (Figure 7), whose columns:

- **Statistical Confidence of the Sample:** In this plan, three distinct confidence levels were considered for the farm. $1 - \alpha = 0.90$; and 0.95. These details should be entered into the platform, in Field 3 (Figure 7);
- **Expected Average Proportion** and **Expected Maximum Proportion**: In this plan, distinct situations were considered. $\hat{p} = 0.10$; 0.16; and 0.20 and $\hat{p} = 0.18$; 0.28; and 0.38. These details should be entered into the platform, in Field 4 (Figure 7).
- **Approximate Accuracy**: The estimative for the precision, based on $\hat{p}$ and $\hat{p}_{max}$;
- **Intra-cluster correlation**: The estimative for ρ;
- **Effective Sample Size**: The sample size, independent of clusters;

- **Sample Size by Cluster**: n_0.

The selected line highlights the choice of the agronomic researchers and indicates that, under the assumption that the Beta-Binomial model is capable of representing how the infection is distributed, with a confidence of 90%, a sample of 13 individuals from each *cluster* should be sufficient to enable the identification of an average proportion as expected by the researchers. The sampled areas are highlights in Field 8 (Figure 7).

2.2 Conclusion

After all the necessary meetings and studies, supported by theoretical foundations in statistics and agronomy, it was possible to develop a platform capable of generating sampling plans using georeferenced data from orange groves. Thus, all collaborators involved in inspecting the citrus greening disease in orange orchards will be able to perform their work much faster and more effectively, without the need to conduct a full census of citrus plantations.

CHAPTER

3

BAYESIAN MODELING FOR EQUIVALENCE ASSESSMENT IN AGRONOMIC TRIALS: COMPARISON OF INFORMATIVE PRIORS IN DETECTING CITRUS VARIETIES RESISTANT TO CITRUS CANKER.

3.1 Analysis of the first dataset (for this first dataset, some researchers on the team made mistakes during the data collection process).

3.1.1 Methodology

3.1.1.1 The experiment

The experiment was installed in a greenhouse. Twelve 12 genotypes of orange were planted in pots. There were 5 replicates of each genotype, that is, in the experiment, there was a total of 60 pots with orange plants. Six randomly chosen leaves from each plant received 6 perforations each. In this way, each plant represented a replicate in the experiment.

The evaluations were performed through the measurements of the diameter of the lesions as follows: between September and November 2016, 8 evaluations were carried out. The diameter of the lesion was measured from one of the perforations of each sheet (observational units), randomly chosen. That is, at each observation, a lesion was randomly selected rom each leaf and its diameter was measured, totaling 360 observations (60 pots x 6 observation each). In the test sweet oranges were used (*Citrus sinensis*) of the pear variety was used in the experiment.

The sheets were perforated with needles measuring (0.55 x 0.20 mm). Inoculated *Xanthomonas citri subsp. citri*, to a concentration of 10^8 UFC/mL through spectrophotometer 600 nm. The orange varieties used in the research were: Valência (VALEN), Valência 2 (Valen), Valência Puka (Puka), Valência Paloma (Paloma), Perâ Oriçanga (Pera Ori), Perâ Irradiada (Irradiada), Perâ Itapetininga (Itapetininga), Perâ Maringá (Pera mga), Perâ IAC (Pera IAC), Sanguinea Mombuca (Sanguine), Hamlin (Hamlin), Precoce Oriçanga (Prec. Ori.). Statistical analyzes were performed using a generalized nonlinear models adopting the Bayesian paradigm.

The following subsections will present a brief explanation about how we are going to incorporate the preliminary information, about the past studies regarding the genotypes resistant to citrus canker disease in a probabilistic point of view, added to the obtained experimental results. Then, presents the growth curves models which presents the evolution across time of the experimentation combined with the statistical computing.

3.1.1.2 Bayesian Inference

The Bayesian methodology is based on the premise that all uncertainties must be modeled through direct probability calculations, and thus, statistical inference draws conclusions based on probability laws (CHRISTENSEN *et al.*, 2010). In this context, Bayesian inference treats the parameter of interest as a random quantity, associating it with a probability distribution, which is the key aspect that distinguishes Bayesian Inference from Frequentist Inference. By combining prior information with sample data, the posterior distribution of the parameter is obtained. This process is conducted using Bayes' Theorem, which allows updating initial beliefs about a parameter based on new evidence or sample data (GELMAN *et al.*, 2014).

Bayes' Theorem establishes a relationship between the prior distribution, which represents the prior knowledge or belief about the parameter, and the posterior distribution, which reflects the updated knowledge after incorporating the observed data (ROBERT, 2007). The basic formula of the theorem is given by:

$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\Pr(D)},$$

where $\Pr(\theta|D)$ is the posterior distribution of the parameter θ given the data D, $\Pr(D|\theta)$ is the likelihood, $\Pr(\theta)$ is the prior distribution, and $\Pr(D)$ is the normalizing constant (or evidence).

The main advantage of Bayesian inference is its flexibility in combining prior information with new data, which is particularly useful in situations with small samples or where previous information is relevant (CHRISTENSEN *et al.*, 2010). The choice of the prior can significantly influence the results; however, as more data is incorporated, the impact of the prior tends to diminish, and the posterior distribution more robustly reflects the available evidence (BERNARDO; SMITH, 1994).

Another important feature of Bayesian inference is its applicability to complex models, such as hierarchical models, where multiple levels of uncertainty can be modeled (GELMAN *et al.*, 2014). Furthermore, modern computational methods, such as Markov Chain Monte Carlo (MCMC), allow the practical application of Bayesian inference to high-dimensional problems, where calculating the posterior distribution analytically would be impractical (BROOKS *et al.*, 2011).

Finally, Bayesian inference has been widely adopted in fields such as biostatistics, data science, and machine learning due to its ability to incorporate uncertainties and provide probabilistic estimates of the parameters of interest. Its flexibility and probabilistic focus make it a powerful tool for decision-making in environments with limited or uncertain information (BISHOP, 2006).

3.1.1.3 Gamma Distribution

A flexible distribution, part of the exponential family is the Gamma distribution, its range is continuous containing two parameters, which one gives its form α and the other its scale β. It can only be used considering scenarios which the model assumes positive asymmetric data to the right, incorporating fat tails. Its most usual density is given in the form

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \qquad \begin{matrix} y > 0, \\ \alpha, \beta > 0 \end{matrix},$$

with expected mean and variance

$$\mathbb{E}(Y) = \frac{\alpha}{\beta} = \mu \qquad \text{and} \qquad \text{Var}(Y) = \frac{\alpha}{\beta^2} = \sigma^2.$$

However, it is encourage to reparameterize its function in order to work as a function of its average. Renaming $\alpha = \mu^2/\sigma^2$ and $\beta = \mu/\sigma^2$, then the respective reparameterized density is

$$f(y) = \frac{\mu}{\sigma^2 \Gamma(\mu^2/\sigma^2)} y^{\mu^2/\sigma^2 - 1} e^{-(\mu/\sigma^2)y}, \qquad \begin{matrix} y > 0, \\ \mu, \sigma > 0, \end{matrix}$$

now their expected mean and variance turn out to be more familiar as

$$\mathbb{E}(Y) = \mu \qquad \text{and} \qquad \text{Var}(Y) = \sigma^2.$$

3.1.1.4 Non-linear Gamma Regression Model

The Gamma model is a generalized linear model which consists of a random component, a link function and a systematic component, where the random component is linked to the systematic component through the link function. To construct a regression, in this work a generalized non-linear model will be constructed, where a non-linear function (on parameters) will be assigned to the systematic component since a growth curve will be assumed for it. That is, a growth model will be considered because the averages of the leaf disease diameters are increasing over time.

Consider a set of independent distributed random observations $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_N^\top)^\top$, where each component $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$, with $i = 1, \ldots, N$, denotes the vector with n_i repeated observations for i-th individual. Conditionally to random effects vector, denoted by $\boldsymbol{\gamma}_{i\mu}$ and $\boldsymbol{\gamma}_{i\sigma}$, the set $Y_{i1}, \ldots, Y_{in_i}$, with $i = 1, \ldots, N$, are independent (μ_{ij}, σ_{ij}) distributed random variables. The random effects are independent each other with Multivariate Normal distribution with mean vector equals $\mathbf{0}$ and covariance matrix $\boldsymbol{G}_\mu^{-1}$ and $\boldsymbol{G}_\sigma^{-1}$, respectively. Keep in mind the formulation of a Generalized Additive Model for Location, Scale and Shape (GAMLSS) Rigby and Stasinopoulos (2005), each parameter μ_{ij} and σ_{ij} of (μ_{ij}, σ_{ij}) distribution can be associated with the random variable Y_{ij} by a link function. The Gamma non-linear Regression Model can be defined by

$$\begin{aligned} Y_{ij} \mid \boldsymbol{\gamma}_{i\mu}, \boldsymbol{\gamma}_{i\sigma} &\sim (\mu_{ij}, \sigma_{ij}), \\ \boldsymbol{\gamma}_{i\mu} &\sim (\mathbf{0}, \boldsymbol{G}_\mu), \\ \boldsymbol{\gamma}_{i\sigma} &\sim (\mathbf{0}, \boldsymbol{G}_\sigma), \end{aligned} \qquad \begin{aligned} &i = 1, \ldots, N, \\ &j = 1, \ldots, n_i, \end{aligned}$$

furthermore, the parameters μ_{ij} and σ_{ij} satisfy the functional relationships

$$\begin{aligned} \mu_{ij} &= g_\mu(\boldsymbol{x}_{ij\mu}, \boldsymbol{\beta}_\mu, \boldsymbol{z}_{ij\mu}, \boldsymbol{\gamma}_{i\mu}), \\ \sigma_{ij} &= g_\sigma(\boldsymbol{x}_{ij\sigma}, \boldsymbol{\beta}_\sigma, \boldsymbol{z}_{ij\sigma}, \boldsymbol{\gamma}_{i\sigma}), \end{aligned}$$

where the functions g_μ and g_σ are non-linear functions of parameters vectors $\boldsymbol{\beta}_\mu$ and $\boldsymbol{\beta}_\sigma$, respectively, such as those described in Subsection 3.1.1.5, for example.

In this context, it is understood that

- $\boldsymbol{x}_{ij\mu} = (x_{1ij\mu}, \ldots, x_{J'_\mu ij\mu})^\top$ and $\boldsymbol{x}_{ij\sigma} = (x_{1ij\sigma}, \ldots, x_{J'_\sigma ij\sigma})^\top$ are vectors with explanatory variables observations associated with fixed effects of μ and σ to i-th individual in his j-th observation.
- $\boldsymbol{z}_{ij\mu} = (z_{1ij\mu}, \ldots, z_{q'_\mu ij\mu})^\top$ and $\boldsymbol{z}_{ij\sigma} = (z_{1ij\sigma}, \ldots, z_{q'_\sigma ij\sigma})^\top$ are vectors with explanatory variables observations associated with random effects of μ and σ to i-th individual in his j-th observation.
- $\boldsymbol{\beta}_\mu = (\beta_{1\mu}, \ldots, \beta_{J'_\mu \mu})^\top$ and $\boldsymbol{\beta}_\sigma = (\beta_{1\sigma}, \ldots, \beta_{J'_\sigma \sigma})^\top$ are parameters vectors associated with fixed effects in distribution parameters.
- $\boldsymbol{\gamma}_{i\mu} = (\gamma_{1\mu}, \ldots, \gamma_{q'_\mu \mu})^\top$ and $\boldsymbol{\gamma}_{i\sigma} = (\gamma_{1\sigma}, \ldots, \gamma_{q'_\sigma \sigma})^\top$ are random effects vectors of μ, and σ to i-th individual.
- $\boldsymbol{G}_\mu^{-1} = \boldsymbol{G}_\mu^{-1}(\boldsymbol{\lambda}_\mu)$ and $\boldsymbol{G}_\sigma^{-1} = \boldsymbol{G}_\sigma^{-1}(\boldsymbol{\lambda}_\sigma)$ are generalized inverse of symmetric matrices $\boldsymbol{G}_\mu = \boldsymbol{G}_\mu(\boldsymbol{\lambda}_\mu)$ and $\boldsymbol{G}_\sigma = \boldsymbol{G}_\sigma(\boldsymbol{\lambda}_\sigma)$, of order $q_\mu \times q_\mu$ and $q_\sigma \times q_\sigma$, respectively, which may depend of hyperparameters vectors $\boldsymbol{\lambda}_\mu$ and $\boldsymbol{\lambda}_\sigma$ associated to the parameters μ e σ.

It is important to note that a variety of curves may be conveniently chosen to represent the g_μ and g_σ functions, particularly in that study four specific growth models were selected. A brief exposition is presented in the next section.

3.1.1.5 Growth curves

The growth curves considered in this study are the Logistic, Gompertz, Weibull and Hill models. Their mathematical representations and descriptions are shown in Table 2. A graphical illustration of the models behavior with parameter variation can be seen in Figure 8.

Table 2 – Growth curves.

Model reference	Mathematical representation	Description
Logistic[1]	$g(x) = \frac{c}{1+\exp\{-a(x-b)\}}$	The logistic model was proposed when Verhulst, not agreeing with Thomas Malthus (population multiplies geometrically and food arithmetically), proposes a differential equation that ends up representing a much more precise population growth.
Gompertz[2]	$g(x) = c\exp\{-e^{-b-ax}\}$	The Gompertz model was proposed by the mathematician Benjamin Gompertz with the objective of modeling human mortality. However, it underwent modifications to model biological behaviors and their details. It is a sigmoid function that describes growth as being slower at the beginning and the end in a given period of time.
Weibull[3]	$g(x) = c\big(1-\exp\{-ax^b\}\big)$	The Weibull model is of great interest due to several characteristics, among them are its effectiveness and the ease of its interpretations due to graphic resources. This model provides at least reasonable accuracy even with small samples. It was proposed by Waloddi Weibull.
Hill[4]	$g(x) = \frac{c\,a^b}{a^b+x^b}$	The Hill model was proposed by Archibald Vivian Hill. His main goal was to describe the speed of muscle contraction.

[1] (VERHULST, 1838)
[2] (GOMPERTZ, 1825)
[3] (WEIBULL, 1951)
[4] (CHICHARRO; VAQUERO, 2006)

Source: The authors

Once we presented four appropriated models for nonlinear systematic components, illustrated in Figure 8, some questions start to be raised regarding the estimation point of view. Therefore, next subsections will present some alternatives for inference and model selection assumed in this work.

3.1.1.6 Models selection

The selection of models constitutes an important tool of statistical modeling, through it, the researcher has an additional basis to justify the choice of an appropriate model. Some of them are, the Deviance Information Criterion (DIC), the Extended Akaike Information Criterion (EAIC) and the Extended Bayesian Information Criterion (EBIC).

Based on these criteria, the model that presents the smallest measure is chosen. Defining $\mathrm{D}(\bar{\theta})$ such as *a posteriori mean of the deviance* and $\mathrm{D}(\widehat{\theta})$ such as *a posteriori measure of the quality of the model fit for the data*, the definitions of each criterion are as follows:

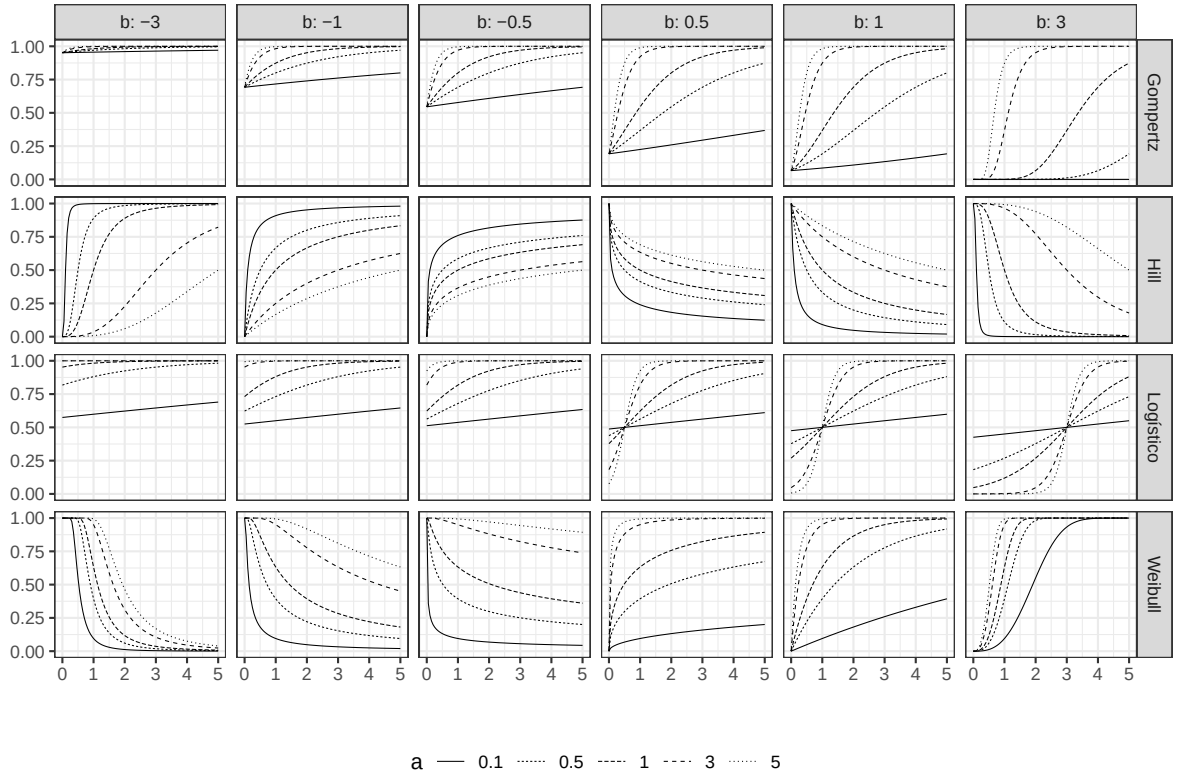


Figure 8 – Growth curves behavior illustration. In all cases, since parameter c represents the maximum value reached by any curve, it is fixed at $c = 1$.

Source: The authors

- DIC: One of the most used within Bayesian statistics, defined by $\text{DIC} = 2\text{D}(\bar{\theta}) - \text{D}(\widehat{\theta})$. (SPIEGELHALTER *et al.*, 2002)

- EAIC: Defined as $\text{EAIC} = \text{D}(\bar{\theta}) + 2k$, sendo k the number of parameters. (BROOKS *et al.*, 2002)

- EBIC: Defined by $\text{EBIC} = \text{D}(\bar{\theta}) + k\ln(n)$, being k the number EBIC: Defined effective parameters and n the sample size. (BRADLEY; THOMAS, 2008)

3.1.1.7 Convergence Criteria

To perform the convergence diagnostics, one can make use of graphical methods as well as numerical methods. Both are important tools and efficient in their purposes, but the numerical methods are more objective. Numerical methods include, for example, Gelman and Rubin (1992), Brooks and Gelman (1998), Geweke (1992), Heidelberger and Welch (1983) and autocorrelation. Among the most well-known graphic methods are trace, histogram, density and quantil-quantil estimators.

The Gelman and Rubin method compares variability between and within the chains, and a result very close to one obtained by the test statistic indicates chain convergence. The Geweke criterion is based on the equality test of the means of the first and last part of the Markov

chain. If the test statistic is between -2 and 2, it assumes chain convergence. Under the null convergence hypothesis, it is assumed that the samples are taken from a stationary distribution of the chain; the criterion of Heidelberger and Welch, tests the null hypothesis of stationarity. If the null hypothesis is rejected for a given value, the test is repeated, disregarding 10% of the first iterations. If the rejection occurs again, another initial 10% is discarded and the test is repeated. This is done until you eliminate 50% of the string. At the end, if the null hypothesis continues to be rejected, there is an indication of the lack of stationarity, and it is necessary to increase the number of iterations. Further details on this criterion can be found in Brooks e Roberts 1998 (BROOKS; ROBERTS, 1998). The self-correlation method identifies the size of the jump that the string must give to ensure independence between observations. This will allow the simulated sequence to mimic the behavior of independent random variables.

In relation to the graphical methods, when dealing with the trace graph, the equilibrium distribution is obtained by tracing two parallel lines along the chain. If the space between the lines is filled, it will be a good indicator of convergence. The method with the use of histograms aims to present the form of the distribution that generated the data. In this way, two superimposed histograms are assembled. One with the first third of the chain, after the discard of burnin, and the other with the final third of the chain. The closer the histograms are, the greater the similarity observed by the superposition, which will be a great indication of convergence. The method of density estimators, considered non-parametric, works as follows: after discarding burnin, the kernel of the first third is compared to the last third of the chain. The closer the convergence indicators are, the closer they are. And last but not least, we have the quantil-quantil, which compares the quatís of the first third of the chain with the latter, also after the discard of burnin. The closer a line is to the graph, the higher the convergence value of the analyzed chain.

3.1.1.8 Computational Procedure

The retrieval of the *a posteriori* distributions for the parameters was done with the aid of the package `rstan` Stan Development Team (2024) of *R* R Core Team (2018). Using MCMC process 10,000 values were generated, discarding the initial 5,000 values for adaptation of the simulation method. The convergence was attested based on the criteria of Gelman and Rubin, Geweke and also Heidelberger and Welch implemented in the package `coda` Plummer *et al.* (2006) of the *R*software. For any parameter θ, a non-informative *a priori* distribution was given, ie $\theta \sim \mathrm{N}(0, 100^2)$.

3.1.2 Proposed solution to the problem

3.1.2.1 Exploratory analysis

First, we conducted a descriptive analysis that tries to identify tendentious and discrepant behaviors, where it is applied several computational resources, based on statistical theories, defining what is or is not essential. On the table 3, there are some descriptive measures of all

genotypes used. The first column has exposed genotypes. The following two columns contain the maximum and minimum values of the lesion diameters, observed during all collections. And lastly, three other columns, the mean, standard deviation and coefficient of variation are respectively from the observed data. There is evidence of discrepancies between such measures. The lowest genotypes averages presented were among the varieties of Valencia, Valencia 2 and Irradiada. Moreover, Irradiada contains the lowest coefficient of variation.

Table 3 – Lesion diameter summary among genotype.

Orange Genotype	Minimum Value	Maximum Value	Mean Value	Standard Deviation	Coefficient of Variation
Hamlin	1.09	3.88	2.201	0.582	0.265
Irradiada	1.04	2.85	1.607	0.278	0.173
Itapetinig	1.10	4.09	2.276	0.591	0.260
Paloma	1.07	3.82	2.155	0.588	0.273
Pera IAC	0.89	4.04	2.127	0.543	0.255
Pera Mga	1.11	4.32	2.171	0.561	0.258
Pera Ori.	1.16	4.44	2.338	0.621	0.265
Prec. Ori	0.99	4.77	2.244	0.594	0.265
Puka	0.97	3.88	2.058	0.528	0.257
Sanguine	0.91	3.61	1.899	0.426	0.224
Valen.	0.54	3.00	1.545	0.397	0.257
VALEN.	0.54	2.50	1.515	0.331	0.219

Source: The authors

Figure 9 presents twelve graphs of profiles, where all values observed (in gray) are plotted, referring to the diameters of the 12 genotype lesions. In black lines present the means of these observations. Graphically, there are indications of different behaviors in both the values of the averages, as well as in the growth curves over time. Among the proposed models, the Range distribution associated with Logistic, Wibull, Gompertz and Hill growth models, the selected model was Hill. The estimates obtained for the parameters are in Table 4.

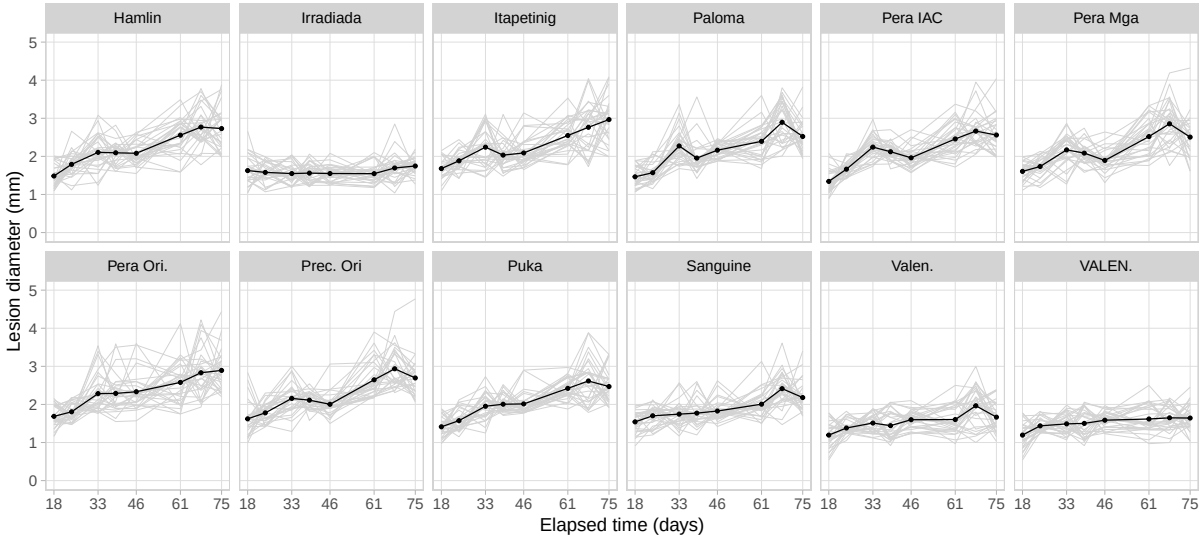


Figure 9 – Graphs of profiles of the observed values (in gray) and their averages (in black) referring to the diameters of the lesions of the 12 genotypes.

Source: The authors

3.1.2.2 Selected model fit analysis

The proposed model after selection among candidate models, has the form

$$\begin{aligned} Y_{ij} \mid \gamma_{ia}, \gamma_{ib}, \gamma_{ic} &\overset{\text{ind}}{\sim} (\mu_{ij}, \sigma), \\ \gamma_{ia} &\overset{\text{iid}}{\sim} \text{N}(0, \lambda_a), \qquad i = 1, \ldots, 60, \\ \gamma_{ib} &\overset{\text{iid}}{\sim} \text{N}(0, \lambda_b), \qquad j = 1, \ldots, 6, \\ \gamma_{ic} &\overset{\text{iid}}{\sim} \text{N}(0, \lambda_c), \end{aligned}$$

in this context, the parameter μ_{ij} satisfy the following functional relation

$$\mu_{ij} = g_\mu(t_{ij}, \boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{c_i}{1 + e^{a_i(t_{ij} - b_i)}},$$

in which

$$\boldsymbol{x}_i = (\boldsymbol{x}_{ia}, \boldsymbol{x}_{ib}, \boldsymbol{x}_{ic}) \qquad \text{and} \qquad \boldsymbol{\beta} = (\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \boldsymbol{\beta}_c),$$

with a_i, b_i and c_i, given by

$$\begin{aligned} a_i &= \boldsymbol{x}_{ia}^\top \boldsymbol{\beta}_a + \gamma_{ia}, \\ b_i &= \boldsymbol{x}_{ib}^\top \boldsymbol{\beta}_b + \gamma_{ib}, \\ c_i &= \boldsymbol{x}_{ic}^\top \boldsymbol{\beta}_c + \gamma_{ic}. \end{aligned}$$

and the vector parameter of interest, to be estimated, is $\boldsymbol{\theta} = (\boldsymbol{\beta}_a^\top, \boldsymbol{\beta}_b^\top, \boldsymbol{\beta}_c^\top, \sigma, \lambda_a, \lambda_b, \lambda_c)^\top$, where

- t_{ij} denote the time of evaluations, in days, elapsed up to the j-th evaluation of i-th subject;
- $\boldsymbol{x}_{ia} = \boldsymbol{x}_{ib} = \boldsymbol{x}_{ic} = (\text{GE}_{1i}, \ldots, \text{GE}_{12i})^\top$ denotes the indicator vector of observed genotype to i-th subject. Each of the 12 components are dichotomous variables denoting, respectively, the genotype origin: (1) *Hamlin*, (2) *Irradiada*, (3) *Itapetinig*, (4) *Paloma*, (5) *Pera IAC*, (6) *Pera Mga*, (7) *Pera Ori.*, (8) *Prec. Ori.*, (9) *Puka*, (10) *Sanguine*, (11) *Valen.*, (12) *VALEN.*;
- γ_{ia}, γ_{ib} e γ_{ic} denotes, respectively, the random effects from each subject observed on the parameters a_i, b_i and c_i.

Carry on with the analysis, after adjusting the four growth models, the parameters estimations of selected model are presented in Table 4 with the standard deviation (of posteriori distribution), the 95% credibility interval, the standard error associated with the estimates, as well as the convergence criteria used in this study are included. According to the proposed criteria, all chains converged.

Observing the general results obtained for the parameters a, b and c, there is a great variation among genotypes estimates, indicating that there is a peculiar behavior inherent to each genotype when exposed to this disease. Keeping in mind that, the b, and c parameters describe, on average, the time that the lesion diameter will reach its median value and the maximum diameter, respectively, one can interpret that:

- the lowest values of c are associated to the *Irradiada*, *VALEN.* and *Valen.* varieties, which expresses the higher resistance of these varieties in comparison to the others.
- the *Hamlin* variety shows the highest growth of lesion diameter over time, although it behaves very similarly to other varieties, such as *Prec. Ori.*, *Pera Ori.*, *Valen.* and *Paloma*.

In addition, it is also possible to see that there is heterogeneity within the observations with respect to the parameter c, indicated by the parameter estimate $\lambda_c = 0.0364$, compared to $\lambda_a \leq 1e-4$.

Table 4 – Summary measures for lesion diameter, taking into account the genotype.

				95% Credibility Interval		Standard Deviation	Convergence Diagnosis		
Parameter	Orange Genotype	Posteriori Mean	Standard Error	2.5%	97.5%	Time-Series	GR statistics	HW p-value	GW p-value
a	Hamlin	-0.0314	0.0001	-0.0399	-0.0248	0.0038	1.003	0.524	0.741
	Irradiada	0.0289	0.0001	0.0221	0.0376	0.0039	1.003	0.563	0.758
	Itapetinig	0.0197	0.0001	0.0131	0.0283	0.0039	1.003	0.559	0.737
	Paloma	-0.0168	0.0001	-0.0387	0.0017	0.0104	1.001	0.193	0.361
	Pera IAC	-0.0324	0.0001	-0.0612	-0.0073	0.0138	1.000	0.213	0.559
	Pera Mga	0.0098	0.0001	0.0012	0.0192	0.0046	1.002	0.700	0.775
	Pera Ori.	-0.0021	0.0001	-0.0217	0.0115	0.0083	1.001	0.178	0.583
	Prec. Ori	0.0094	0.0001	0.0021	0.0184	0.0042	1.002	0.485	0.758
	Puka	-0.0122	0.0001	-0.0329	0.0044	0.0097	1.000	0.352	0.335
	Sanguine	0.0225	0.0001	0.0158	0.0312	0.0039	1.003	0.512	0.734
	Valen.	-0.0332	0.0002	-0.0613	-0.0025	0.0145	1.001	0.373	0.618
	VALEN.	-0.0401	0.0001	-0.0670	-0.0169	0.0129	1.000	0.731	0.886
λ_a	—	0.0000	0.0000	0.0000	0.0002	0.0001	1.001	0.254	0.924
b	Hamlin	20.6249	0.0037	19.7307	21.5263	0.4589	1.000	0.647	0.802
	Irradiada	360.5689	0.0031	359.5833	361.5685	0.5058	1.000	0.471	0.862
	Itapetinig	150.6056	0.0029	149.6161	151.5862	0.5040	1.000	0.151	0.813
	Paloma	-5.0901	0.0029	-6.0425	-4.1396	0.4798	1.001	0.140	0.443
	Pera IAC	-5.9780	0.0030	-6.9233	-5.0337	0.4785	1.000	0.593	0.404
	Pera Mga	-0.9279	0.0031	-1.9135	0.0557	0.5010	1.000	0.137	0.541
	Pera Ori.	-7.3638	0.0031	-8.3330	-6.3755	0.4966	1.000	0.385	0.867
	Prec. Ori	10.3648	0.0031	9.3978	11.3422	0.4997	1.000	0.113	0.870
	Puka	-4.4416	0.0031	-5.4150	-3.4673	0.4960	1.000	0.312	0.782
	Sanguine	145.0472	0.0030	144.0799	146.0326	0.4975	1.000	0.906	0.754
	Valen.	-16.1639	0.0031	-17.1560	-15.1835	0.4988	1.001	0.695	0.351
	VALEN.	-19.9119	0.0031	-20.8906	-18.9362	0.4995	1.000	0.764	0.984
$\log(c)$	Hamlin	1.1731	0.0008	1.1059	1.2357	0.0328	1.001	0.366	0.945
	Irradiada	0.5285	0.0014	0.1454	0.8918	0.1901	1.000	0.117	0.814
	Itapetinig	1.3118	0.0009	1.1878	1.4380	0.0630	1.001	0.223	0.839
	Paloma	-0.1441	0.0009	-0.2500	-0.0324	0.0560	1.000	0.405	0.659
	Pera IAC	-0.2128	0.0009	-0.3100	-0.1020	0.0533	1.000	0.261	0.781
	Pera Mga	0.0586	0.0008	-0.0268	0.1444	0.0437	1.001	0.481	0.765
	Pera Ori.	-0.0112	0.0009	-0.1345	0.0998	0.0598	1.001	0.160	0.827
	Prec. Ori	0.1845	0.0008	0.1087	0.2625	0.0394	1.001	0.533	0.992
	Puka	-0.1547	0.0009	-0.2654	-0.0406	0.0576	1.000	0.373	0.644
	Sanguine	0.8317	0.0008	0.7004	0.9634	0.0675	1.001	0.454	0.840
	Valen.	-0.6077	0.0012	-0.7068	-0.4705	0.0607	1.001	0.694	0.533
	VALEN.	-0.6584	0.0008	-0.7449	-0.5642	0.0469	1.001	0.370	0.627
λ_c	—	0.0364	0.0001	0.0259	0.0492	0.0059	1.001	0.980	0.400

Source: The authors

In Figure 10 the estimated curves for the 12 genotypes of the study are plotted. It is visually observed that the Gamma model, associated with the Logistic growth model, well represent the average growth behavior of the diameter of the disease over time. Note that the *Irradiada* variety remained constant throughout the experiment, in addition to having maintained their diameters below 2mm, as can also be seen in Figure 11, which indicates that this variety has much more resistance to citrus canker in relation to the others. On the other hand, the variety *VALEN.* was susceptible in the first observations, but in most of the study it remained constant, that is, it also showed to be resistant to citrus canker disease. However, although the *Valen.* variety was less constant, over time, compared to *VALEN.*, both had similar results.

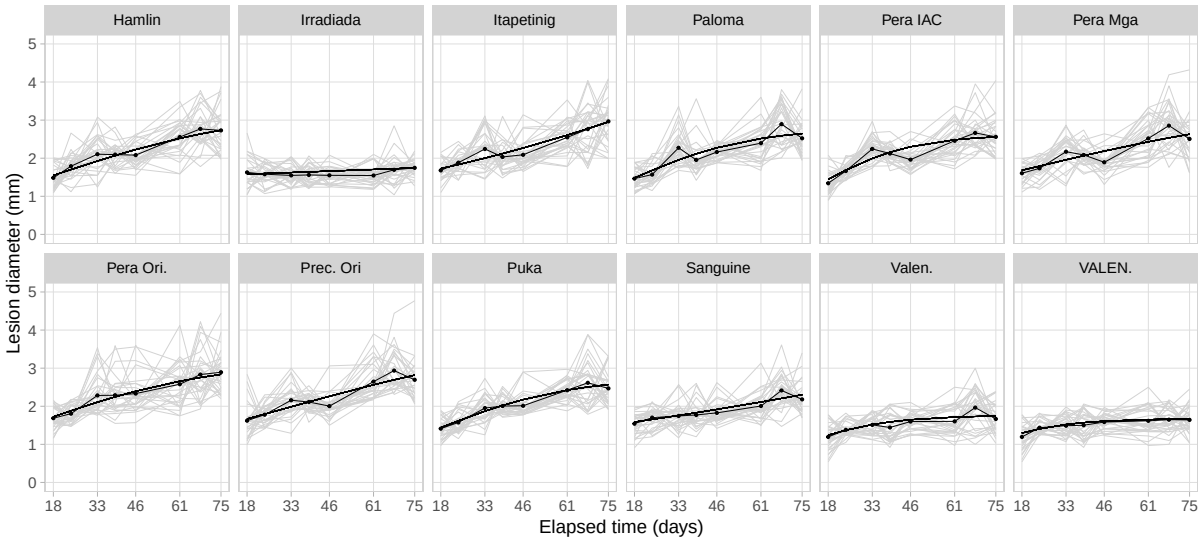


Figure 10 – Graphs of profiles of the observed values (in gray) and their estimates (in black) referring to the diameters of the 12 genotype lesions.

Source: The authors

In the first graph of Figure 11, associated with the parameter *a*, is exposed the HPD intervals for the representation of the time in which the diameter of the disease reaches its median value. In this context it is estimated 45 days for the *Hamlin* variety in contrast to approximately 10 days of the *Irradiada* variety. The maximum lesion size is described by the *c* parameter.

In the third graph of Figure 11, it is noted that the two varieties *Valen.* and *VALEN.* have different medians, but they have a very similar range, for the maximum lesion diameter. Already the *Irradiada* variety and *Valen.* have practically equal medians, however they are very discrepant in relation to the maximum size that their lesion diameters can reach.

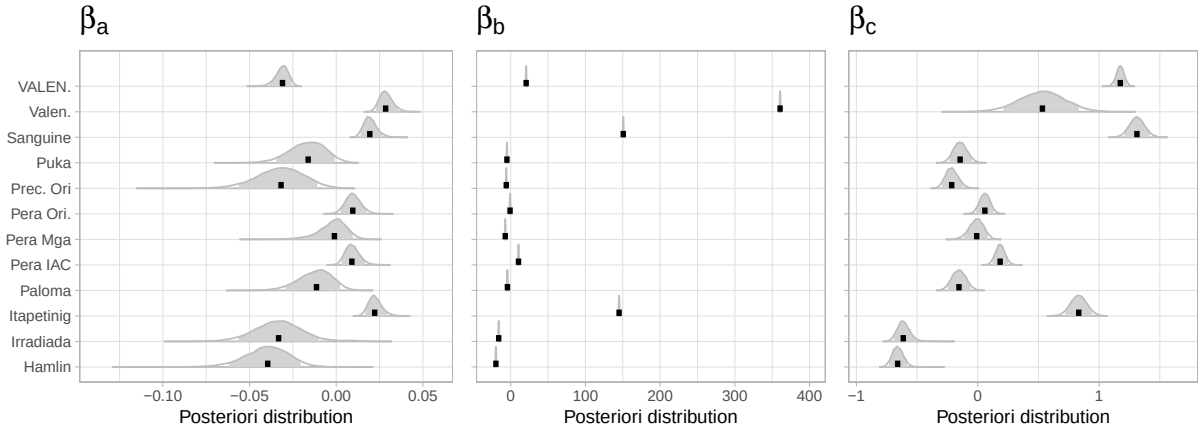


Figure 11 – HPD intervals for the estimates of the estimated parameters: *a* (left), *b* (center) and *c* (right).

Source: The authors

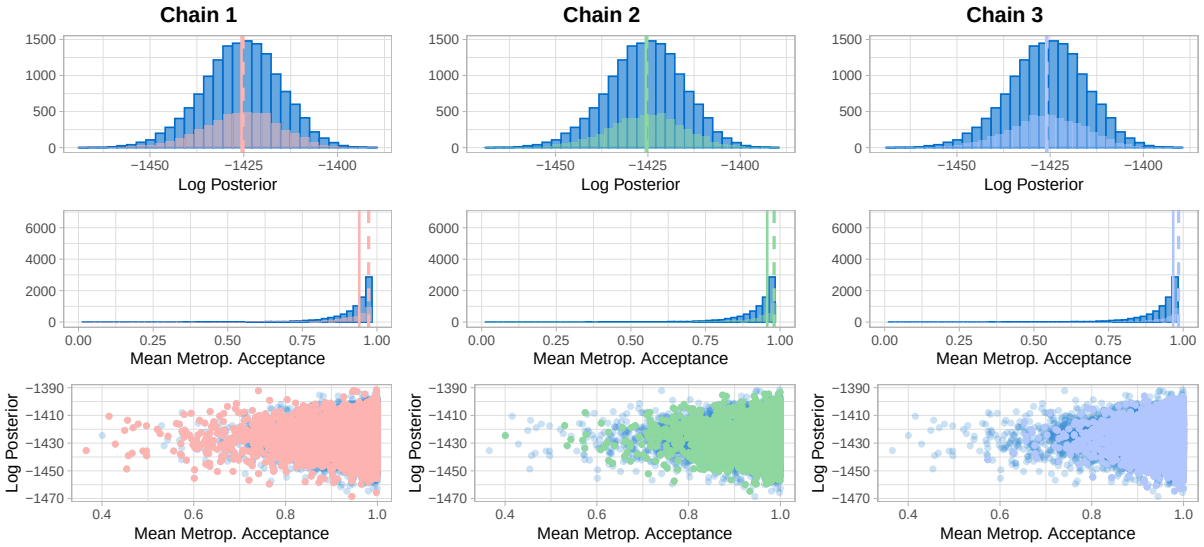


Figure 12 – .

Source: The authors

3.1.3 Conclusion

The generalized nonlinear model proved a reasonable alternative approach. Using the characteristic, from the adjusted growth curves, would endure the interpretability of the parameters allowed to characterize interesting aspects, like the exposure of several orange genotypes to citrus canker disease. Statistical evidence points out that the most resistance to citrus canker disease were Irradiada, Valencia and Valencia 2, respectively. In contrast, the Hamlin genotype was the most susceptible to the disease.

3.2 Analysis of the second dataset (at this stage, all data were collected correctly).

3.2.1 Methodology

3.2.1.1 Second experiment

The experiment was conducted in a greenhouse, using a completely randomized design (MONTGOMERY, 2008; BOX; HUNTER; G, 2005), where 12 orange genotypes were evaluated, planted in pots. There were 5 repetitions of each genotype, meaning there were 60 pots with orange plants in the experiment. Six leaves randomly selected from each plant received 6 perforations each, also randomly. Thus, each plant represented a replicate in the experiment. The leaves were perforated with needles measuring $(0.55 \times 0.20, \text{mm})$. The bacterium *Xanthomonas citri subsp. citri* was inoculated at a concentration of $10^8, \text{UFC/mL}$ using a spectrophotometer at $600, \text{nm}$. Evaluations were performed by measuring the diameter of the lesions as follows: the diameter of the six lesions on each leaf was measured, and the average was calculated to obtain a mean value for each leaf. Therefore, in each evaluation, six values were obtained for each

replicate (pot). Thus, in each evaluation, considering the entire experiment, 360 observations were recorded (keeping in mind that each observation noted in the database is an average for each leaf). A total of 8 evaluations were carried out, generating 2880 observations for the database. The sweet orange cultivars used in the trial were *Citrus sinensis*, of the pear variety.

Experiment 1, when conducted, did not follow the recommendations proposed by the research team. In the first trial, a mistake was made in measuring the diameter of the lesions. Instead of measuring all the perforations on each leaf, allowing for the calculation of the average value per leaf, only one perforation per leaf was measured in each evaluation, even randomly each time the experiment was assessed. This was noted when the statisticians began their work using descriptive statistics. The disease in focus (when it appears) maintains a certain growth over time (regarding the lesion area) or, at the very least, remains constant (when the lesion area does not increase in size). At certain moments, the descriptive statistics showed that these lesion areas even decreased and then started to grow again (which does not align with the agronomic literature).

Experiment 2 was planned exactly as intended.

3.2.2 Proposed solution to the problem

3.2.3 New exploration

In Figure 13, a direct comparison between the data collected in the first and second trials can be seen. The key point of attention is the reduction in variability; it is noticeable that the replicates from the second trial are less dispersed than those from the first. Moreover, patterns that were previously less evident for some genotypes became more pronounced in the second collection, as seen in the genotypes Irradiada, Valen., and VALEN.

After adjusting the four growth models, the parameters estimations of selected model are presented in Table 5 with the standard deviation (of posteriori distribution), the 95% credibility interval, the standard error associated with the estimates, as well as the convergence criteria used in this study are included. According to the proposed criteria, all chains converged.

Observing the general results obtained for the parameters a, b and c, the great variation among genotypes estimates remains, indicating that there is a peculiar behavior inherent to each genotype when exposed to this disease. In addition, it is also possible to see that there is heterogeneity within the observations with respect to the parameter c, indicated by the parameter estimate $\lambda_c = 0.337$, compared to $\lambda_a \leq 1e-3$.

In Figure 14, the estimated curves for the 12 genotypes in first and second the study are displayed. It can be visually observed that the Gamma model, combined with the Logistic growth model, effectively captures the average growth pattern of the disease diameter over time. Notably, the genotypes *Irradiada*, *Valen.* and *VALEN* throughout the experiment in the first data

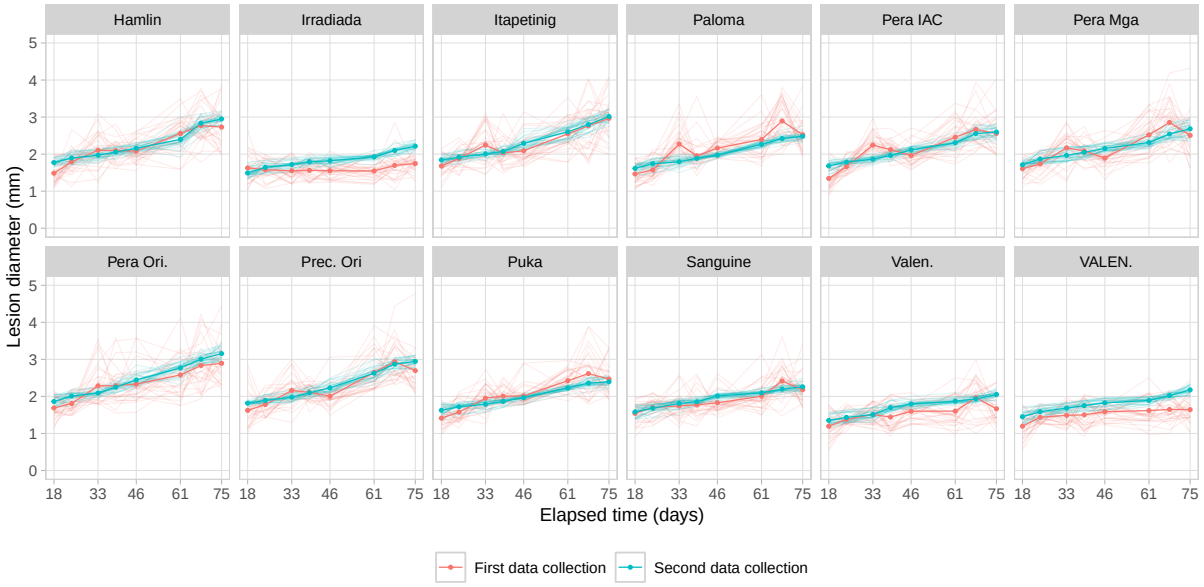


Figure 13 – Graphs of profiles of the observed values for the first and second data collection (in transparent red and blue, respectively) and their averages (in solid red and blue) referring to the diameters of the lesions of the 12 genotypes.

Source: The authors

Table 5 – Summary measures for lesion diameter, taking into account the genotype.

				95% Credibility Interval		Standard Error	Convergence Diagnosis		
Parameter	Orange Genotype	Posteriori Mean	Standard Deviation	2.5%	97.5%	Time-Series	GR statistics	HW p-value	GW p-value
	Hamlin	-0.0252	0.0000	-0.0268	-0.0236	0.0008	1.002	0.806	0.946
	Irradiada	0.0192	0.0000	0.0174	0.0211	0.0009	1.007	0.233	0.825
	Itapetinig	0.0140	0.0000	0.0123	0.0159	0.0009	1.004	0.870	0.955
a	Paloma	0.0024	0.0000	-0.0003	0.0050	0.0014	1.002	0.518	0.765
	Pera IAC	0.0014	0.0000	-0.0014	0.0041	0.0014	1.002	0.545	0.439
	Pera Mga	0.0061	0.0000	0.0039	0.0083	0.0011	1.003	0.543	0.710
	Pera Ori.	-0.0056	0.0000	-0.0084	-0.0028	0.0014	1.000	0.923	0.841
	Prec. Ori	0.0034	0.0000	0.0014	0.0055	0.0010	1.003	0.941	0.800
	Puka	0.0050	0.0000	0.0025	0.0076	0.0013	1.001	0.056	0.870
	Sanguine	0.0170	0.0000	0.0153	0.0190	0.0009	1.007	0.328	0.672
	Valen.	-0.0087	0.0001	-0.0158	-0.0020	0.0036	1.000	0.173	0.317
	VALEN.	-0.0022	0.0001	-0.0115	0.0045	0.0041	1.004	0.735	0.340
λ_a	—	0.0007	0.0001	0.0000	0.0019	0.0006	1.228	0.398	0.851
	Hamlin	20.6251	0.0000	20.6175	20.6326	0.0039	1.000	0.536	0.428
	Irradiada	360.5690	0.0000	360.5611	360.5767	0.0040	1.000	0.300	0.847
	Itapetinig	150.6056	0.0000	150.5980	150.6133	0.0039	1.000	0.501	0.584
a	Paloma	-5.0901	0.0001	-5.1107	-5.0702	0.0104	1.001	0.871	0.797
	Pera IAC	-5.9777	0.0002	-6.0049	-5.9506	0.0138	1.000	0.258	0.453
	Pera Mga	-0.9280	0.0000	-0.9369	-0.9190	0.0046	1.001	0.219	0.890
	Pera Ori.	-7.3634	0.0001	-7.3792	-7.3469	0.0083	1.000	0.343	0.916
	Prec. Ori	10.3648	0.0000	10.3565	10.3730	0.0042	1.001	0.472	0.916
	Puka	-4.4415	0.0001	-4.4606	-4.4223	0.0097	1.001	0.166	0.972
	Sanguine	145.0473	0.0000	145.0395	145.0549	0.0039	1.000	0.717	0.741
	Valen.	-16.1634	0.0002	-16.1924	-16.1346	0.0146	1.001	0.727	0.654
	VALEN.	-19.9117	0.0001	-19.9373	-19.8864	0.0129	1.000	0.250	0.388
	Hamlin	1.1746	0.0001	1.1671	1.1821	0.0039	1.001	0.791	0.859
	Irradiada	0.5292	0.0000	0.5215	0.5370	0.0039	1.000	0.469	0.986
	Itapetinig	1.3118	0.0000	1.3040	1.3194	0.0039	1.001	0.011	0.446
c	Paloma	-0.1434	0.0002	-0.1637	-0.1232	0.0104	1.000	0.234	0.716
	Pera IAC	-0.2111	0.0002	-0.2380	-0.1843	0.0136	1.001	0.504	0.508
	Pera Mga	0.0587	0.0001	0.0496	0.0679	0.0046	1.000	0.591	0.720
	Pera Ori.	-0.0108	0.0001	-0.0271	0.0057	0.0083	1.001	0.677	0.508
	Prec. Ori	0.1846	0.0001	0.1764	0.1926	0.0042	1.000	0.957	0.805
	Puka	-0.1540	0.0001	-0.1729	-0.1351	0.0096	1.001	0.970	0.554
	Sanguine	0.8317	0.0000	0.8241	0.8392	0.0039	1.000	0.421	0.407
	Valen.	-0.6056	0.0002	-0.6344	-0.5775	0.0145	1.000	0.179	0.269
	VALEN.	-0.6558	0.0001	-0.6806	-0.6309	0.0127	1.000	0.102	0.151
λ_c	—	0.3369	0.0013	0.2779	0.4076	0.0333	1.002	0.887	0.852

Source: The authors

collection appears to exhibit growth behavior in the second data collection. Furthermore, they all remain significantly more resistant to citrus canker compared to the others.

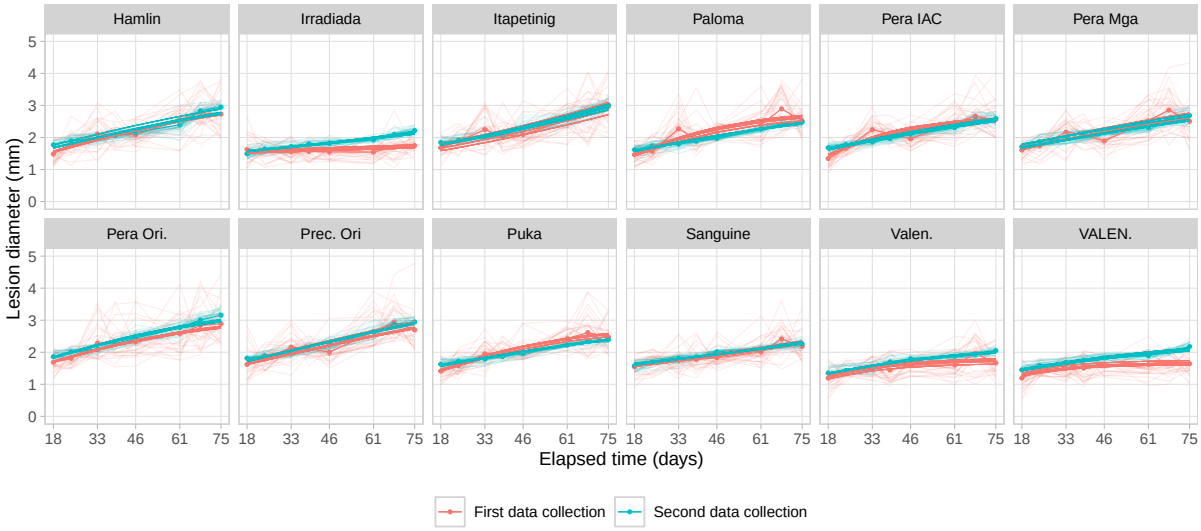


Figure 14 – Graphs of profiles of the observed values (in gray) and their averages (in black) referring to the diameters of the lesions of the 12 genotypes.

Source: The authors

Now, to carry out the comparison proposed in this study, observe in Figure 15 the posterior distributions of the first and second data collections. Keep in mind that the posterior distribution of the first collection was used as the prior for the model fitted to the data from the second collection. It is clear that the experimental errors from the first collection did not affect the results and interpretations regarding the parameters b and c of the proposed model. On the other hand, the parameter a appears to have been influenced in some genotypes, such as *VALEN*, *Valen.*, *Puka*, *Pera IAC*, and *Paloma*.

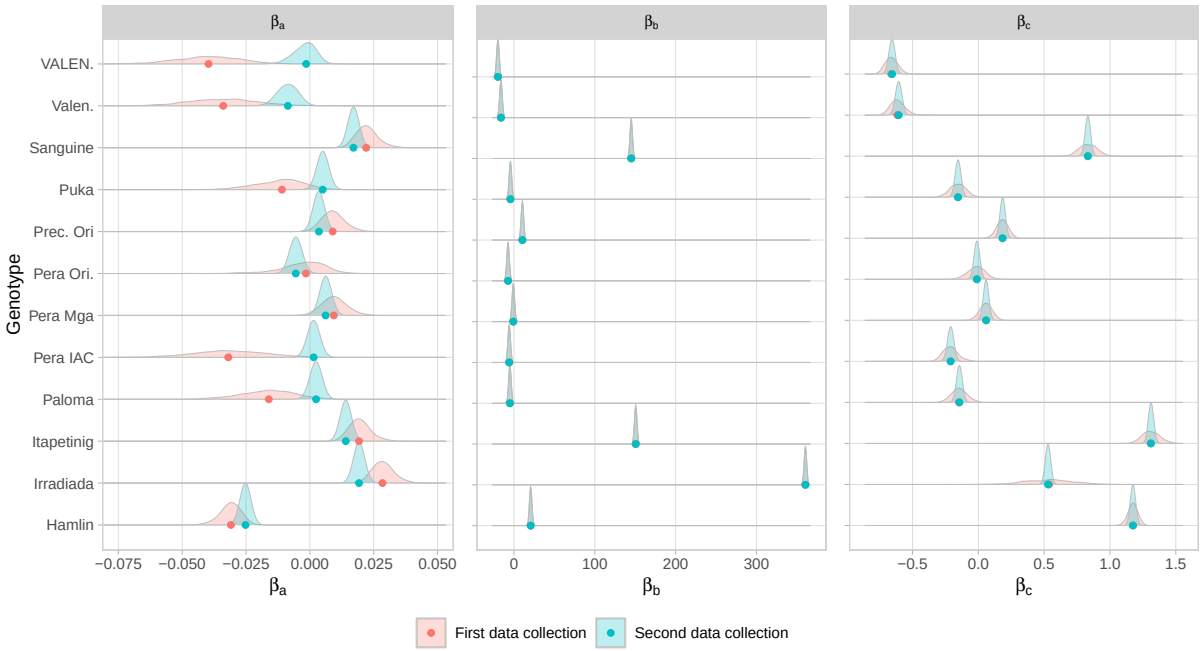


Figure 15 – HPD intervals for the estimates of the estimated parameters: a (left), b (center) and c (right) for the 12 genotypes and 2 data collections.

Source: The authors

Since the previous visualization may carry a certain degree of subjectivity, Figure 16 presents the posterior distribution for the differences between the distributions shown earlier. In this context, it is confirmed that the parameters *b* and *c* were indeed not influenced, and that the parameter *a* showed differences significantly distant from zero (from the perspective of the 95% credibility interval) only for the genotypes *VALEN* and *Pera IAC*.

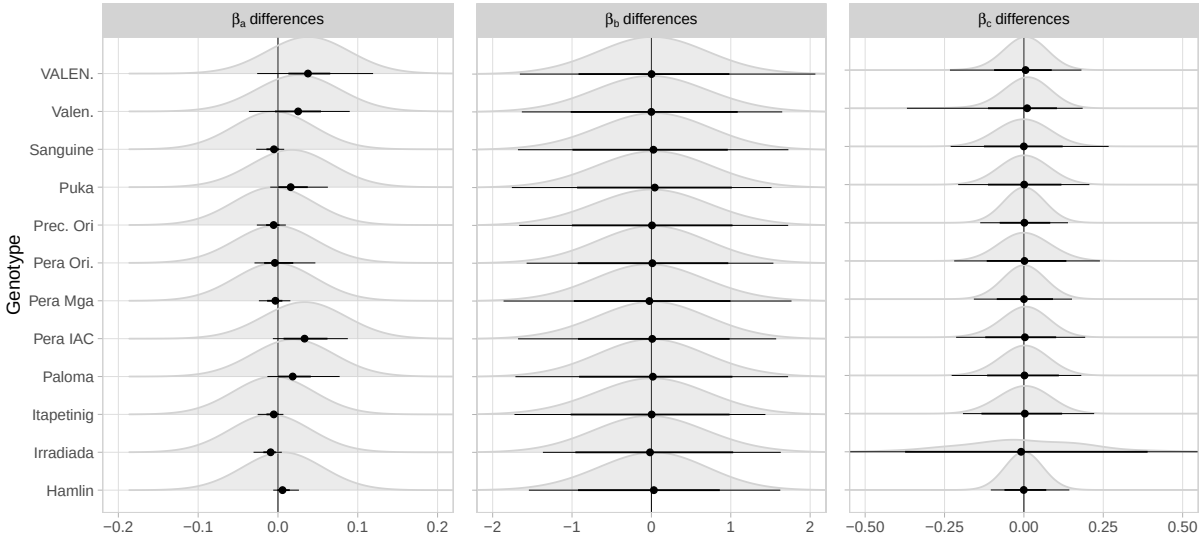


Figure 16 – Graphs of a posteriori differences for the 12 genotypes.

Source: The authors

3.2.4 Conclusion

The proposed method demonstrates a strong capability to evaluate the statistical repeatability of results in agronomic trials. By leveraging posterior distributions and comparing them across multiple data collections, the method provides a robust for assessing consistency in experimental outcomes. This allows for the detection of whether key parameters, such as growth rates or disease resistance measures, remain stable across repeated trials. As a result, the method reduces the subjectivity associated with visual comparisons and enhances the reliability of interpretations, particularly for genotypes under study.

Furthermore, the statistical repeatability of the method ensures that any observed variability can be attributed to true biological differences rather than experimental errors or random fluctuations. This is particularly valuable in agronomic studies, where environmental factors and natural variability can complicate the interpretation of results. By confirming that core parameters remain consistent across trials, the method provides stronger evidence for the validity and robustness of agronomic findings, allowing for more confident decision-making in breeding programs or agricultural management strategies.

CHAPTER

4

LONGITUDINAL BAYESIAN ZERO-INFLATED BETA REGRESSION FOR CITRUS CANKER RESISTANCE IN ORANGE ROOTSTOCKS

4.1 Methodology

Preliminary

Given a random variable Y with open support $U = (0,1) \subset \mathbb{R}$ (that is, Y represents proportions). These bounds correspond to the beta probability distribution, which is commonly used for modeling proportions.

A Beta Regression model is a statistical model that explores the existence and quantifies any potential relationships between Y and the presence of a set of factors that influence its behavior.

Beta Regression models have a range of applications and are commonly used to shed light on understanding various phenomena of interest across diverse knowledge domains (SMITHSON; VERKUILEN, 2006). These models can help answer questions related to average proportions of interest. These data are sourced from areas such as agriculture, including soil fertility, climate, agricultural pests, wood and food production, as well as fields such as healthcare, finance, engineering, education, among many others (KORHONEN *et al.*, 2007) (IBÁÑEZ; PRADES; SIMÓ, 2011) (MENDONÇA; SILVESTRE; PASSOS, 2011) (MORAES; ROCHA; MACHADO, 2012) (MULLEN; MARSHALL; MCGLYNN, 2013).

Despite the Beta Regression model being designed for proportion data, it has a limitation associated with its support as the model, in its original form is unable to model data that contains zeros. Thus, if the data are proportions, a zero-inflated version of the model must be used. This

is the approach adopted in this study.

Experimental design that generated the data

According to Gonçalves-Zuliani (2014b), the experiment was conducted in a rural area in the city of Paranavaí, Paraná, Brazil, at Latitude 23° 1' S, Longitude 50° 41' W, at an altitude of 467 meters above sea level. The genotypes were planted with a spacing of 2.5 meters between plants and 6.0 meters between rows. As four rootstocks were combined with nine genotypes, and the experiment was conducted with ten replicates for each rootstock-genotype combination, a total of 360 orange seedlings were planted.

All necessary cultivation practices for the successful development of this agricultural crop were carried out. Once the plants reached two years of age, five assessments were conducted in 2010, 2011, and 2012. To perform these assessments, four random branches were selected from each orange tree. The total number of leaves and the total number of diseased leaves were counted. This allowed for the calculation of the proportion of diseased leaves for each tree in each assessment. As five assessments were conducted, the database included 1800 observations.

Genotypes used in the experiment (top part of the grafting): 1 - Arapongas; 2 - Bianchi; 3 - EEL; 4 - IAC; 5 - IAC2000; 6 - IpiguaIAC; 7 - N58; 8 - N59; 9 - Olimpia.

Rootstocks (bottom part of the grafting): 1 - Laranja Caipira (Citrus sinensis (L.) Osbeck); 2 - Limão Cravo (Citrus limonia (L.) Osbeck); 3 - Tangerina Cleópatra (Citrus reshni hort. ex Tanaka); 4 - Tangerina Sunki (Citrus sunki (Hayata) hort. ex Tanaka).

Zero-Inflated Beta distribution

The Beta probability distribution, $\text{Be}(p,q)$ is a real-valued, continuous distribution with two shape parameters, denoted as $q > 0$ and $p > 0$, with density function

$$f_{\text{Be}}(y \mid p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \qquad 0 < y < 1,$$

Its parameters $q > 0$ and $p > 0$ shape the form, and the letter Γ represents the Gamma function itself.

The Beta probability distribution can take on various shapes due to its flexibility. These shapes can be constant, U-shaped, strictly increasing, or strictly decreasing. It will take on the form of a uniform distribution when $p = q \neq 1$. It will be strictly increasing (negatively skewed unimodal) when $p > q$; it will be strictly decreasing (positively skewed unimodal) when $p < q$, and it will be symmetrically unimodal (U-shaped) when $p = q \neq 1$.

The mean and variance of this probability distribution are as follows:

$$\mathbb{E}(Y \mid p,q) = \frac{p}{p+q} \qquad \text{and} \qquad \text{Var}(Y \mid p,q) = \frac{pq}{(p+q)^2(p+q+1)}.$$

To implement this model, the probability density function of the Beta distribution Beta(p,q) was reparameterized. For this we let $\sigma = p+q$ and $\mu = p/(p+q)$. Thus $\sigma > 0$ and $0 < \mu < 1$. Note also that $p = \sigma\mu$ and $q = \sigma(1-\mu)$. Thus, under this parameterization, the probability density function of Beta(p, q) takes the following form

$$g_{\text{Be}}(y \mid p,q) = \frac{\Gamma[\sigma]}{\Gamma[\sigma\mu]\Gamma[\sigma(1-\mu)]} y^{\sigma\mu-1}(1-y)^{\sigma(1-\mu)-1}, \qquad 0 < y < 1,$$

in this context, $\sigma > 0$ represents the scale parameter, $\mu \in (0,1)$ represents the location parameter.

The Beta distribution may not be used to model our data as it contains zeros. To do this, we use a special form of a mixture model, the Zero-Inflated Beta distribution. This distribution has a three-parameter probability density function

$$g_{\text{ZIBe}}(y \mid \nu,\mu,\sigma) = [\ \nu\ ]^{1_{\{0\}}(y)} [\ (1-\nu) f_{\text{Be}}(y;\mu,\sigma)\ ]^{1-1_{\{0\}}(y)}, \qquad 0 \le y < 1.$$

Continuing, the parameters represented: $\nu \in (0,1)$ for shape, $\mu \in (0,1)$ for location, and $\sigma > 0$ for scale. When $y = 0$, $1_{\{0\}}(y) = 0$, and when $y \neq 0$, $1_{\{0\}}(y) = 1$. Furthermore, the variance and mean of ZIBe(ν,μ,σ) will be

$$\mathbb{E}(Y \mid \nu,\mu,\sigma) = (1-\nu)\mu \quad \text{and} \quad \text{Var}(Y \mid \nu,\mu,\sigma) = (1-\nu)\left[\frac{\mu(1-\mu)}{\sigma+1} + \nu\mu^2\right].$$

Zero-Inflated Beta regression model

Some examples of modeling the expected value of the Beta distribution, Beta(p,q), were conducted years ago in previous works found in the literature of the field (JORGENSEN, 1997) (CEPEDA-CUERVO, 2001) (PAOLINO, 2001) (KIESCHNICK; MCCULLOUGH, 2003). Shortly thereafter, Ferrari and Cribari-Neto (2004) published a scientific article that brought theoretical and methodological advances to the field of statistics. Working directly with the Beta probability distribution, these authors proposed the Beta Regression Model. This became possible because, in their attempts to model the expected value in the Beta distribution Beta(p,q), they used the parameterization of the distribution in the form Beta(μ,σ) in conjunction with the theories of Generalized Linear Models. Information on this subject can also be found in the work of Nelder (NELDER; WEDDERBURN, 1972). These publications and studies led to various scientific articles and new modeling approaches (CEPEDA-CUERVO, 2014).

However, the previously mentioned works demonstrate promising results in cases where adequate sample sizes are available. For smaller sample sizes, other approaches use modified likelihood methods to provide more robust estimates and inferences (FERRARI; PINHEIRO, 2011).

Continuing with contributions related to modeling involving the Beta probability distribution, specifically focused on diagnostic analyses of this modelRocha and Simas (2011), Ferrari, Espinheira and Cribari-Neto (2011), Chien (2011), Anholeto, Sandoval and Botter

(2012) discussed and published more sophisticated works that introduced influence metrics, new graphical approaches, and new types of residuals for diagnostics. Also contributing to the growth of this field, Zhao *et al.* (2014) proposed a novel variable selection method (involving models with constant dispersion). In a similar vein, to further strengthen these statistical methodologies, expanding the range of possibilities and aiming to achieve even more robust and diverse approaches for comprehensive data analysis, Latif Latif and Yab (2015) presented a new way to optimize experimental design for experiments with the Beta Regression model: Beta(μ, σ). This approach can work with just one predictor variable.

The Zero-Inflated Beta Regression model, ZIBe(ν, μ, σ), was formalized and published by Ospina (OSPINA; FERRARI, 2012a). As it is relatively recent, its usage has not been as widespread as regular Beta Regression, according to the literature. Likelihood ratio testing for it is possible, and there are also comprehensive explanations available regarding the detection of residuals in modeling using the Zero-Inflated Beta Regression (PEREIRA; CRIBARI-NETO, 2012).

Tang *et al.* (2023) addressed challenges in modeling zero-inflated ecological data, emphasizing the difficulty of interpreting data with many zeros. They propose a zero-inflated Beta regression model handling two zero types: zeros from habitat inadequacy (modeled by Bernoulli) and zeros from randomness or incomplete sampling (Left Censoring). Their model, applied to plant cover data in South Africa's Cape Floristic Region, incorporates spatial effects for better prediction. Using Bayesian methods, they explore how environmental factors influence plant cover, showing that traditional methods inadequately address zero inflation.

Maluf, Ferrari and Queiroz (2024) contributed to beta regression, introducing robust estimators, the Logit Transformation Minimum Density Power Divergence Estimator (LMDPDE) and Logit Transformation Smooth Maximum Likelihood Estimator (LSMLE), which overcome sensitivity to outliers. These estimators avoid restrictive parameter conditions of prior methods, maintaining robustness and reliable asymptotic properties. Their new R package, `robustbetareg`, supports implementation and includes an algorithm to optimize tuning, demonstrating utility through health insurance coverage data.

Xia and Sun (2023) explored microbiome data, introducing models like ZIBe and Zero-Inflated Beta-Binomial (ZIBB) for handling compositional, zero-heavy data. ZIBe addresses high-dimensional metagenomic data challenges, while ZIBB manages zero inflation and overdispersion in microbiome counts. These models provide robust frameworks for hypothesis testing and model fitting, enhancing analysis of complex microbiome datasets.

Model definition

Given random observations independent distributed $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_N^\top)^\top$, here each element $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$, with $i = 1, \ldots, N$, denotes a vector of n_i repeated observations for i-th individual. Conditional on the random effects vectors, denoted by $\boldsymbol{\gamma}_{i\nu}$, $\boldsymbol{\gamma}_{i\mu}$ and $\boldsymbol{\gamma}_{i\sigma}$, the

set $Y_{i1},\ldots,Y_{in_i}$, with $i=1,\ldots,N$, are independent $\text{ZIBe}(\nu_{ij},\mu_{ij},\sigma_{ij})$ distributed random variables. The random effects are independent of each other with Multivariate Normal with mean vector equals $\mathbf{0}$ and with the covariance matrix $\boldsymbol{G}_{\nu}^{-1}$, $\boldsymbol{G}_{\mu}^{-1}$ and $\boldsymbol{G}_{\sigma}^{-1}$, respectively. Based on the construction of a Generalized Additive Model for Location, Scale and Shape (GAMLSS) (RIGBY; STASINOPOULOS, 2005), each parameter ν_i, μ_i and σ_i of $\text{ZIBe}(\nu_i,\mu_i,\sigma_i)$ through a link function, Y_i is linked to the model. At this moment the ZIBe regression will be defined as follows

$$\begin{aligned}
Y_{ij} \mid \boldsymbol{\gamma}_{i\nu},\boldsymbol{\gamma}_{i\mu},\boldsymbol{\gamma}_{i\sigma} &\overset{\text{ind}}{\sim} \text{ZIBe}(\nu_{ij},\mu_{ij},\sigma_{ij}),\\
\boldsymbol{\gamma}_{i\nu} &\overset{\text{iid}}{\sim} \text{N}(\mathbf{0},\boldsymbol{G}_{\nu}), && i=1,\ldots,N,\\
\boldsymbol{\gamma}_{i\mu} &\overset{\text{iid}}{\sim} \text{N}(\mathbf{0},\boldsymbol{G}_{\mu}), && j=1,\ldots,n_i.\\
\boldsymbol{\gamma}_{i\sigma} &\overset{\text{iid}}{\sim} \text{N}(\mathbf{0},\boldsymbol{G}_{\sigma}),
\end{aligned}$$

Furthermore, the parameters ν_{ij}, μ_{ij} and σ_{ij} satisfy the functional relationships

$$\begin{aligned}
g_{\nu}(\nu_{ij}) &= \eta_{ij\nu} = \boldsymbol{x}_{ij\nu}^{\top}\boldsymbol{\beta}_{\nu} + \boldsymbol{z}_{ij\nu}^{\top}\boldsymbol{\gamma}_{i\nu},\\
g_{\mu}(\mu_{ij}) &= \eta_{ij\mu} = \boldsymbol{x}_{ij\mu}^{\top}\boldsymbol{\beta}_{\mu} + \boldsymbol{z}_{ij\mu}^{\top}\boldsymbol{\gamma}_{i\mu},\\
g_{\sigma}(\sigma_{ij}) &= \eta_{ij\sigma} = \boldsymbol{x}_{ij\sigma}^{\top}\boldsymbol{\beta}_{\sigma} + \boldsymbol{z}_{ij\sigma}^{\top}\boldsymbol{\gamma}_{i\sigma}.
\end{aligned}$$

In this context, it is understood that for each distribution parameter ν,μ or σ represented generically by the symbol $\circ$ we have $\boldsymbol{x}_{ij\circ} = (x_{1ij\circ},\ldots,x_{J'_{\circ}ij\circ})^{\top}$ is a vector of observations drawn from an explanatory variables related to the fixed effects of the j-th evaluation of the i-th case; $\boldsymbol{z}_{ij\circ} = (z_{1ij\circ},\ldots,z_{q'_{\circ}ij\circ})^{\top}$ is a vector of observations drawn from an explanatory variables related to the random effects of the j-th evaluation of the i-th case; $\boldsymbol{\beta}_{\circ} = (\beta_{1\circ},\ldots,\beta_{J'_{\circ}\circ})^{\top}$, is a parameter vector related with fixed effects in distribution parameter; $\boldsymbol{\gamma}_{i\circ} = (\gamma_{1\circ},\ldots,\gamma_{q'_{\circ}\circ})^{\top}$, is a random effect vector related to i-th individual on the distribution parameter; $\boldsymbol{G}_{\circ}^{-1} = \boldsymbol{G}_{\circ}^{-1}(\boldsymbol{\lambda}_{\circ})$, are generalized inverse of symmetric matrices $\boldsymbol{G}_{\circ} = \boldsymbol{G}_{\circ}(\boldsymbol{\lambda}_{\circ})$, of order $q_{\circ}\times q_{\circ}$, which may depend on hyperparameter vectors $\boldsymbol{\lambda}_{\circ}$.

Finally, for the prior distribution of the regression coefficients $\boldsymbol{\beta}$, we used $\text{N}(0,0.001)$ to represent a relatively non-informative prior, minimizing its influence on the results and allowing the data to drive the inference. For the dispersion parameters $\boldsymbol{\lambda}$, we selected $\text{Gamma}(0.01,0.01)$ to reflect a broad, flat prior with high variance and low mean, accommodating a wide range of possible values.

Expectation and variance

Taking into account, the binding terms g_{σ} and g_{ν}, g_{μ} to ν, σ and μ, we have estimated values for the variance of Y_{ij}, in addition to covariance and correlation between Y_{ij} and $Y_{(ij')}$ that can be obtained conditional for random effects, or can be determined marginally by the expressions

$$\mathbb{E}(Y_{ij}) = \mathbb{E}[\mathbb{E}(Y_{ij} \mid \boldsymbol{\gamma}_{i\nu},\boldsymbol{\gamma}_{i\mu},\boldsymbol{\gamma}_{i\sigma})] = \mathbb{E}[(1-\nu_{ij})\mu_{ij}] = [1-\mathbb{E}(\nu_{ij})]\mathbb{E}(\mu_{ij}),$$

and

$$\begin{aligned}\mathrm{Var}(Y_{ij}) &= \mathrm{Var}\left[\,\mathbb{E}(Y_{ij} \mid \boldsymbol{\gamma}_{iv}, \boldsymbol{\gamma}_{i\mu}, \boldsymbol{\gamma}_{i\sigma})\,\right] + \mathbb{E}\left[\,\mathrm{Var}(Y_{ij} \mid \boldsymbol{\gamma}_{iv}, \boldsymbol{\gamma}_{i\mu}, \boldsymbol{\gamma}_{i\sigma})\,\right] \\ &= \mathrm{Var}\left[\,(1-\nu_{ij})\mu_{ij}\,\right] + \mathbb{E}\left\{(1-\nu_{ij})\left[\frac{\mu_{ij}(1-\mu_{ij})}{(\sigma_{ij}+1)+\nu_{ij}\mu_{ij}^2}\right]\right\},\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(Y_{ij};Y_{ij'}) &= \mathrm{Cov}\left[\,\mathbb{E}(Y_{ij} \mid \boldsymbol{\gamma}_{iv}, \boldsymbol{\gamma}_{i\mu}, \boldsymbol{\gamma}_{i\sigma}); \mathbb{E}(Y_{ij'} \mid \ bm\gamma_{iv}, \boldsymbol{\gamma}_{i\mu}, \boldsymbol{\gamma}_{i\sigma})\,\right] + \\ &\qquad \mathbb{E}\left[\,\mathrm{Cov}(Y_{ij} \mid \boldsymbol{\gamma}_{iv}, \boldsymbol{\gamma}_{i\mu}, \boldsymbol{\gamma}_{i\sigma}; Y_{ij'} \mid \boldsymbol{\gamma}_{iv}, \boldsymbol{\gamma}_{i\mu}, \boldsymbol{\gamma}_{i\sigma})\,\right] \\ &= \mathrm{Cov}\left[\,(1-\nu_{ij})\mu_{ij}; (1-\nu_{ij'})\mu_{ij'}\,\right] \\ &= \mathbb{E}\left[\,(1-\nu_{ij})\mu_{ij}(1-\nu_{ij'})\mu_{ij'}\,\right] - \mathbb{E}\left[\,(1-\nu_{ij})\mu_{ij}\,\right]\mathbb{E}\left[\,(1-\nu_{ij'})\mu_{ij'}\,\right],\end{aligned}$$

and

$$\mathrm{Corr}(Y_{ij};Y_{ij'}) = \frac{\mathrm{Cov}(Y_{ij};Y_{ij'})}{\sqrt{\mathrm{Var}(Y_{ij})\,\mathrm{Var}(Y_{ij'})}}.$$

Given the parameter estimates, the integrals in these expressions can be approximated with a numerical integration method, such as Gauss-Hermite Quadrature, for example, using the R package `cubature` (NARASIMHAN *et al.*, 2023). In this context, interval estimates can be obtained adopting the Delta Method, using the package `msm` (Christopher H. Jackson, 2011). On the other hand, they can even be obtained using MCMC samples, with their respective posterior distributions, obtained with packages `jags` (PLUMMER, 2018) and `coda` (PLUMMER *et al.*, 2006).

Model selection and convergence criteria

Model selection was conducted using three main criteria. The Deviance Information Criterion (DIC) (SPIEGELHALTER *et al.*, 2002) combines the deviance, which measures model fit, with a penalty for the effective number of parameters to prevent overfitting; a lower DIC value indicates a better model. The formula for DIC is $\mathrm{DIC} = 2\,\mathrm{D}(\bar{\theta}) - \mathrm{D}(\widehat{\theta})$. The Extended Akaike Information Criterion (EAIC) (BROOKS *et al.*, 2002) is an extension of the Akaike Information Criterion (AIC) that adds penalties for parameter count and model structure, making it suitable for complex models. A lower EAIC indicates a better balance between model accuracy and parsimony, and it is calculated as $\mathrm{EAIC} = \mathrm{D}(\bar{\theta}) + 2k$, with k being the number of parameters estimated. Lastly, the Extended Bayesian Information Criterion (EBIC) (BRADLEY; THOMAS, 2008), an enhancement of the Bayesian Information Criterion (BIC), is given by $\mathrm{EBIC} = \mathrm{D}(\bar{\theta}) + k\ln(n)$, where n represents the sample size. In all cases, $\mathrm{D}(\bar{\theta})$ denotes the a posteriori mean of deviations and $\mathrm{D}(\widehat{\theta})$ denotes a posteriori quality measure for data adjustment. In addition, k denotes the number of parameters estimated in the model.

To evaluate chain convergence, both numerical and graphical methods were employed. For numerical convergence diagnostics, three criteria were used. The Gelman and Rubin (GR)

statistic (GELMAN; RUBIN, 1992; BROOKS; GELMAN, 1998) runs multiple parallel MCMC chains and compares within—and between—chain variances. When the GR statistic is close to 1, typically with a threshold of GR ≤ 1.1, the chains are considered to have converged. Geweke's method (GW p-value) (GEWEKE, 1992) compares the means of the first and last parts of the MCMC chain to check for statistical similarity, where a p-value above 5% indicates convergence. Heidelberger and Welch (HW p-value) (HEIDELBERGER; WELCH, 1983) consists of a two-step process: testing for stationarity using the Cramer-von Mises test, and checking accuracy by ensuring the credible interval's half-width-to-mean ratio is below a threshold. If both tests are passed with a p-value above 5%, the chain is deemed converged.

Graphical methods also aided in convergence diagnostics. These included histograms and density estimates for a posteriori approximation, trace plots to visualize chain stability, limit mean plots, and autocorrelation functions.

Computational Aspects

The ZIBe(ν, μ, σ) distribution was adjusted to 1800 original observations for incidence per subject (36 clones $\times$ 5 evaluations $\times$ 10 repetitions), considering the influence of explanatory variables and effect of time.

- All of them used *a priori* were little or non-informative. In particular, normal flat, N$(0, 0.001)$ for the real parameters, and gamma flat Gamma$(0.01, 0.01)$ for positive parameters.
- In all simulated contexts, the `R` (R Core Team, 2018) software was used to compute the results using the `RStudio` (RStudio Team, 2015) interface.
- The simulations of the samples of the a posteriori distributions, as well as the models were achieved with packages `jags` (PLUMMER, 2018) and `coda` (PLUMMER *et al.*, 2006). In addition to coding the model to be sampled with `jags`, as carried out in this study, it is also possible to use a facilitating framework, such as the `zoib` package (KONG, 2023) or even adjust this model under the facilitated framework for STAN with the `brms` package (BüRKNER, 2021).
- The tests used for chain diagnosis are implemented in the `coda` package and the diagnostic graphs are implemented in the `ggmcmc` (MARÍN, 2016) package.
- 15,000 iterations were considered as adaptation time for the simulation. The burn-in was 35,000 and the sample used was 100,000 with jumps from 100 to 100 in order to generate independent samples.
- Descriptive and diagnostic charts were constructed using `ggplot2` (WICKHAM, 2016).

4.2 Proposed solution to the problem

To address and understand the occurrence of citrus canker infection in orange tree leaves, a statistical modeling approach was employed using the Zero-Inflated Beta Regression model ZIBe(ν, μ, σ). In this modeling, the disease incidence was represented by Y (in the experimental context, the number of diseased leaves divided by the total number of leaves collected), which, in turn, is influenced by two categorical covariates: rootstock and genotype. The rootstock was denoted as RS with the following categories: *Tangerina Cleópatra*, *Laranja Caipira*, *Tangerina Sunki* and *Limão Cravo*). The genotype, on the other hand, was represented as GE with the respective categories: *Olímpia*, *Arapongas*, *N59*, *Bianchi*, *Ipiguá IAC*, *ELL*, *IAC 2000*, *N58* and *IAC*.

Exploratory analysis

The exploratory analysis began by examining the behavior of observations grouped by evaluation. The histograms in Figure 17 depict the incidence measures observed across each of the five evaluations. It reveals that the proportion of uninfected plants varies; however, discerning specific patterns from this chart is challenging. Additionally, the dispersion of measures seems to fluctuate, but the average incidence among non-zero observations appears to exhibit a growing trend over time.

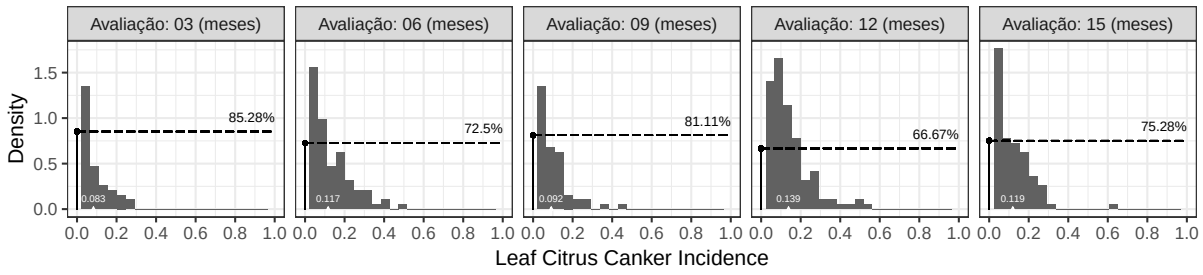


Figure 17 – Histograms for Leaf Citrus Canker Incidence to each tree analyzed considering the evaluation. The point on the x-axis indicates the sample mean of nonzero observations.

Figures 18 and 19 more clearly demonstrate trends over time. These figures present the observed incidence in plants already infected (illustrated in the boxplots below) alongside the proportion of uninfected plants (denoted by points above).

In Figure 18, a growth trend in incidence is observed once a plant is infected. This trend is fairly consistent across all rootstocks, with the possible exception of the Limao Cravo rootstock, which appears to maintain a more stable incidence level. Additionally, the ratio of null observations during the evaluations is noteworthy. Here, the Laranja Caipira rootstock is distinguished by its high and stable proportion of null observations. Conversely, the Limao Cravo rootstock shows a high vulnerability, as indicated by a significant decrease in the proportion of null observations over time.

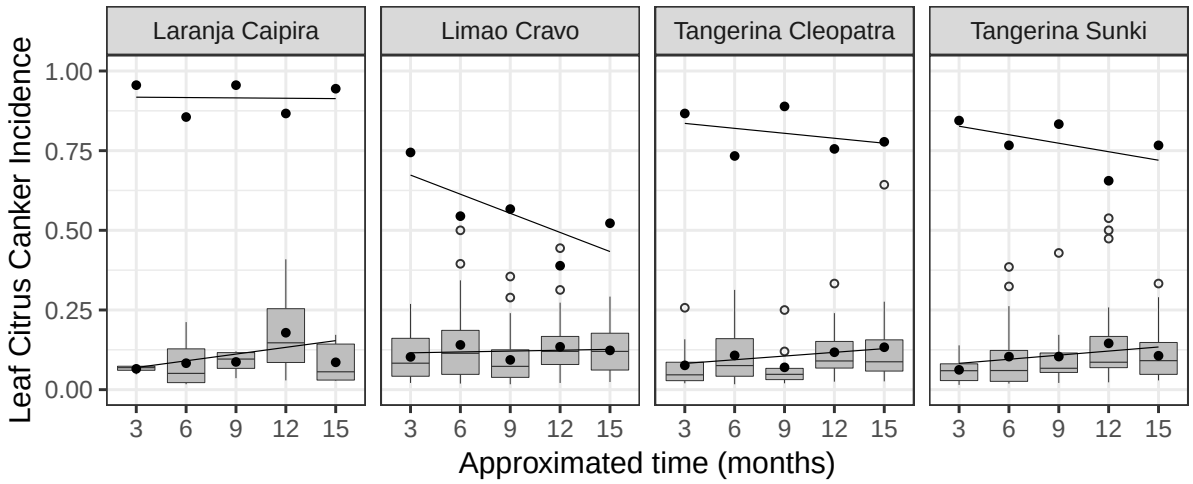


Figure 18 – Boxplots of the incidence of citrus canker disease (considering zeros and non-zeros) per replicate (plant) evaluated taking into account the evaluations or the rootstocks.

As can be seen in Figure 19, the non-zero incidence behavior, represented in boxplots, is more diverse among genotypes than among rootstocks. Some genotypes stand out by keeping the incidence low and stable, such as EEL and N59. The Arapongas genotype, strangely, shows a tendency to fall in incidence. The other genotypes show a tendency to increase incidence.

The proportion of null incidence is close in all genotypes, except for three, as the largest falls were observed in genotypes IAC and N58, and the IpiguaIAC genotype shows the highest overall proportion of null incidence. In addition, genotype N59, also strangely, exposes an increase in the proportion of zero incidences.

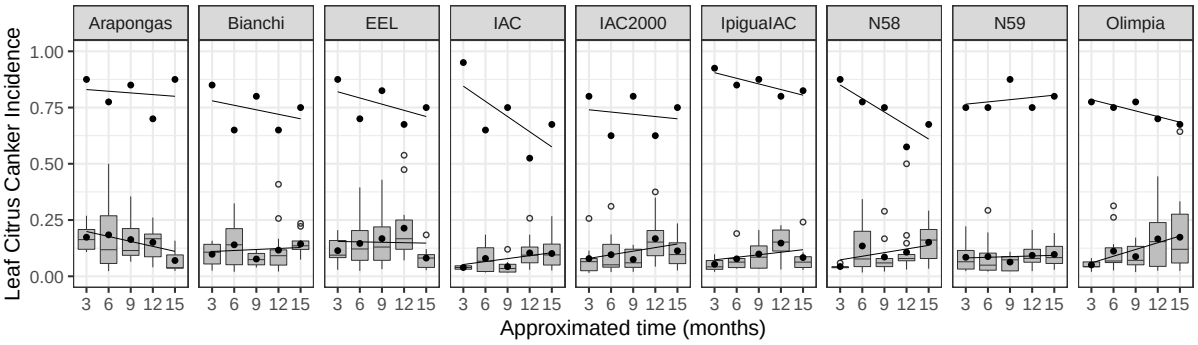


Figure 19 – Boxplots constructed with non-zero incidences and observed incidence ratios for incidence of leaf citrus canker per plant analyzed for genotype and assessment.

Selected Model Fit

To perform model selection, we proposed several possibilities considering, for example, the presence/absence of the effects of genotype, rootstock, evaluation, etc. Specifically, for the evaluation effect, we considered the presence of linear, quadratic, and cubic terms. These terms were included to understand a possible temporal behavior that is not strictly increasing/decreasing, as with the linear term, but that could represent a point of maximum/minimum incidence from

which the plant becomes progressively healthier/sicker, either due to some applied management or specific resistance of the studied genotype-rootstock combination.

After selecting among candidate models, the proposed model, shown in Tables 6 and 7, has the following form

$$\begin{aligned} Y_{ij} \mid \gamma_{0i\mu}, \gamma_{0i\nu} &\overset{\text{ind}}{\sim} \text{ZIBe}(\nu_{ij}, \mu_{ij}, \sigma), \\ \gamma_{0i\mu} &\overset{\text{iid}}{\sim} \text{N}(0, \lambda_\mu), \\ \gamma_{0i\nu} &\overset{\text{iid}}{\sim} \text{N}(0, \lambda_\nu), \end{aligned} \qquad \begin{aligned} &i = 1, \ldots, 360, \\ &j = 1, \ldots, 5, \end{aligned}$$

in this context, the parameters ν_{ij}, μ_{ij} and σ_i satisfy the following functional relations

$$\log\left(\frac{\nu_{ij}}{1-\nu_{ij}}\right) = \boldsymbol{x}_{ij\nu}^\top \boldsymbol{\beta}_\nu + \gamma_{0i\nu},$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \boldsymbol{x}_{ij\mu}^\top \boldsymbol{\beta}_\mu + \gamma_{0i\mu},$$

$$\log(\sigma_i) = \beta_{0\sigma},$$

where the vector of parameters to be estimated consists of $\boldsymbol{\theta} = (\boldsymbol{\beta}_\nu^\top, \boldsymbol{\beta}_\mu^\top, \beta_{0\sigma}, \lambda_\nu, \lambda_\mu)^\top$, with $\boldsymbol{\beta}_\nu = (\beta_{0\nu}, \beta_{1\nu}, \ldots, \beta_{13\nu})^\top$, $\boldsymbol{\beta}_\mu = (\beta_{0\mu}, \beta_{1\mu}, \ldots, \beta_{13\mu})^\top$, and belongs to the parametric space

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^{31} \mid \boldsymbol{\beta}_\nu \in \mathbb{R}^{14}, \boldsymbol{\beta}_\mu \in \mathbb{R}^{14}, \beta_{0\sigma} \in \mathbb{R}, \lambda_\nu \in \mathbb{R}_*^+, \lambda_\mu \in \mathbb{R}_*^+\}.$$

In addition, the observed covariable vectors for the i-th individual in j-th evaluation are denoted, respectively, by $\boldsymbol{x}_{ij\nu} = \boldsymbol{x}_{ij\mu} = (1, \text{AV}_{ij}, \text{AV}_{ij}^2, \text{PE}_i^\top, \text{GE}_i^\top)^\top$, where:

- AV_{ij} denotes the number of evaluations (each one equivalent to approximately 3 months) elapsed up to the j-th evaluation of i-th subject.
- $\text{PE}_i = (\text{PE}_{1i}, \ldots, \text{PE}_{3i})^\top$ denotes the indicator vector of the observed rootstock to the i-th subject. Each of the three components are dichotomous variable denoting, respectively, the rootstock origin: 1 – *Limão Cravo*, 2 – *Tangerina Cleópatra*, 3 – *Tangerina Sunki*. *Laranja Caipira* rootstock was used as a reference in the model.
- $\text{GE}_i = (\text{GE}_{1i}, \ldots, \text{GE}_{8i})^\top$ denotes the indicator vector of the observed genotype to i-th subject. Each of the eight components are dichotomous variables denoting, respectively, the genotype origin: 1 – *Arapongas*, 2 – *Bianchi*, 3 – *EEL*, 4 – *IAC*, 5 – *IAC2000*, 6 – *N58*, 7 – *N59*, 8 – *Olimpia*. *IpiguaIAC* genotype was used as a reference in the model.
- $\gamma_{0i\nu}$ and $\gamma_{i\mu}$ denote, respectively, the random effects from each subject observed on parameters ν and μ.

Table 6 – Proposed regression structures and criteria used to select the appropriate structure for the ν parameter of the distribution ZIBe(ν,μ,σ), that is, the proportion of uninfected plants.

Source: The authors

Step	Model	logit(ν)	logit(μ)	log(σ)	BIC	EAIC	EBIC
I.	1	$\nu_0+\gamma_{0\nu}$	μ_0	σ_0	33770.63	33778.63	33800.61
II.	1	$\nu_0+\gamma_{0\nu}$	μ_0	σ_0	33770.63	33778.63	33800.61
	2	$\nu_0+\gamma_{0\nu}+\text{AV}$	μ_0	σ_0	33742.37	33752.37	33779.85
	3	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2$	μ_0	σ_0	**33731.50**	**33743.50**	**33776.47**
	4	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{AV}^3$	μ_0	σ_0	33735.59	33749.59	33788.06
III.	1	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2$	μ_0	σ_0	33731.50	33743.50	33776.47
	2	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}$	μ_0	σ_0	33731.35	33745.35	33783.82
	3	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{GE}$	μ_0	σ_0	33730.44	33744.44	33782.91
	4*	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	μ_0	σ_0	**33723.05**	**33739.05**	**33783.01**

*Regression structure chosen for ν.

Table 7 – Proposed regression structures and criteria used to select the appropriate structure for the μ parameter of the distribution ZIBe(ν,μ,σ), that is, the mean incidence of infected plants.

Step	Model	logit(ν)	logit(μ)	log(σ)	DIC	EAIC	EBIC
I.	1	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}$	σ_0	33706.29	33724.29	33773.75
II.	1	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}$	σ_0	33706.29	33724.29	33773.75
	2	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}+\text{AV}$	σ_0	33710.30	33730.30	33785.26
	3	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}+\text{AV}+\text{AV}^2$	σ_0	**33680.34**	**33702.34**	**33762.79**
	4	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}+\text{AV}+\text{AV}^2+\text{AV}^3$	σ_0	—	—	—
III.	1	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}+\text{AV}$	σ_0	33680.34	33702.34	33762.79
	2	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}+\text{AV}+\text{PE}$	σ_0	33686.30	33710.30	33776.25
	3	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}+\text{AV}+\text{GE}$	σ_0	33689.72	33715.72	33787.16
	4*	$\nu_0+\gamma_{0\nu}+\text{AV}+\text{AV}^2+\text{PE}+\text{GE}$	$\mu_0+\gamma_{0\mu}+\text{AV}+\text{PE}+\text{GE}$	σ_0	**33673.68**	**33695.68**	**33756.13**

*Regression structure chosen for μ.

Based on estimates for parameters associated with ν, shown in the Table 8, we can see evidence that the adjusted coefficients to understand the time effect on uninfected plants proportion indicate a decreasing behavior in the estimates of ν with the passage of days, as $-1.0761x + 0.1104x^2 < 0$ whenever $2 < x \leq 6$. In addition, from this behavior, it can be understood that there is a lessening of this decrease from the fourth evaluation, approximately 12 months. Only the effects of *Arapongas*, *EEL* and *N59* genotypes are statistically equivalent to the effect of the genotype from the reference combination (*Laranja Caipira + IpiguaIAC*). The other genotypes and also the rootstocks present a statistically different influence.

Based on parameter estimates (Table 9) associated with the mean incidence among infected plants, i.e., associated with the μ parameter, of distribution $\text{ZIBe}(\nu, \mu, \sigma)$, it can be perceived that in the incidence among infected individuals, once *Arapongas* and *EEL* genotypes indicate a significant effect over the reference, this indicates the increase in incidence. All other effects, from genotypes, rootstocks, and even time, were not significant.

For the precision associated with the incidence measures between infected plants, the parameter estimate is given in Table 10.

Table 8 – Point and interval estimates and convergence criteria for regression coefficients and hyperparameters associated with the parameter ν of the ZIBe(ν, μ, σ) distribution.

				95% Credibility Interval		Standard Error		Convergence Diagnosis		
Associated variable	Parameter	Posteriori Mean	Standard Deviation	2.5%	97.5%	Naive	Time-Series	GR statistics	HW p-value	GW p-value
Intercepto	$\beta_{0\nu}$	5.8659	0.7265	4.4611	7.3294	1e-03	0.0233	1.002	0.1314	0.7770
Avaliação	$\beta_{1\nu}$	-1.0761	0.3275	-1.7292	-0.4338	5e-04	0.0165	1.009	0.2290	0.8387
Avaliação (quad.)	$\beta_{2\nu}$	0.1104	0.0398	0.0321	0.1897	1e-04	0.0019	1.020	0.1746	0.8862
Laranja Caipira – PE	$\beta_{3\nu}$	—	—	—	—	—	—	—	—	—
Limao Cravo – PE	$\beta_{4\nu}$	-2.5532	0.2613	-3.0837	-2.0574	4e-04	0.0023	0.992	0.6345	0.2272
Tangerina Cleopatra – PE	$\beta_{5\nu}$	-1.1315	0.2632	-1.6588	-0.6254	4e-04	0.0019	1.005	0.9188	0.2377
Tangerina Sunki – PE	$\beta_{6\nu}$	-1.3460	0.2616	-1.8702	-0.8447	4e-04	0.0019	0.998	0.6436	0.9698
Arapongas – GE	$\beta_{7\nu}$	-0.2669	0.3743	-1.0006	0.4671	5e-04	0.0030	0.993	0.2763	0.3032
Bianchi – GE	$\beta_{8\nu}$	-0.9591	0.3560	-1.6708	-0.2689	5e-04	0.0032	1.002	0.2157	0.9214
EEL – GE	$\beta_{9\nu}$	-0.6879	0.3601	-1.3974	0.0155	5e-04	0.0031	0.998	0.4006	0.8194
IAC – GE	$\beta_{10\nu}$	-1.1743	0.3513	-1.8736	-0.4959	5e-04	0.0032	1.013	0.0532	0.5047
IAC2000 – GE	$\beta_{11\nu}$	-1.0826	0.3546	-1.7899	-0.3983	5e-04	0.0032	0.991	0.0790	0.5245
IpiguaIAC – GE	$\beta_{12\nu}$	—	—	—	—	—	—	—	—	—
N58 – GE	$\beta_{13\nu}$	-0.9265	0.3578	-1.6360	-0.2311	5e-04	0.0031	0.989	0.0663	0.8833
N59 – GE	$\beta_{14\nu}$	-0.6024	0.3622	-1.3168	0.1039	5e-04	0.0031	1.017	0.5300	0.7420
Olimpia – GE	$\beta_{15\nu}$	-0.9040	0.3582	-1.6159	-0.2094	5e-04	0.0031	1.039	0.2636	0.4004
	λ_ν	0.7258	0.2038	0.3659	1.1635	3e-04	0.0032	0.997	0.7225	0.9986

○ Bold lines indicate effects that are not significant or that are statistically equivalent to those in the reference.

Table 9 – Point and interval estimates and convergence criteria for regression coefficients and hyperparameters associated with the parameter μ of the ZIBe(ν, μ, σ).

				95% Credibility Interval		Standard Error		Convergence Diagnosis		
Associated variable	Parameter	Posteriori Mean	Standard Deviation	2.5%	97.5%	Naive	Time-Series	GR statistics	HW p-value	GW p-value
Intercept	$\beta_{0\mu}$	-2.8019	0.4006	-3.5524	-2.0061	6e-04	0.0130	1.014	0.6814	0.9501
Avaliação	$\beta_{1\mu}$	0.2758	0.1855	-0.0903	0.6229	3e-04	0.0099	0.989	0.7322	0.8881
Avaliação (quad.)	$\beta_{2\mu}$	-0.0235	0.0222	-0.0652	0.0203	0e+00	0.0011	0.989	0.6823	0.9254
Laranja Caipira – PE	$\beta_{3\mu}$	—	—	—	—	—	—	—	—	—
Limao Cravo – PE	$\beta_{4\mu}$	-0.0525	0.1452	-0.3317	0.2370	2e-04	0.0012	1.039	0.5173	0.5888
Tangerina Cleopatra – PE	$\beta_{5\mu}$	-0.1446	0.1548	-0.4450	0.1623	2e-04	0.0012	1.102	0.8519	0.6464
Tangerina Sunki – PE	$\beta_{6\mu}$	-0.1450	0.1549	-0.4445	0.1620	2e-04	0.0012	1.042	0.6610	0.6442
Arapongas – GE	$\beta_{7\mu}$	0.4187	0.1923	0.0466	0.8006	3e-04	0.0015	1.030	0.9450	0.7647
Bianchi – GE	$\beta_{8\mu}$	0.2107	0.1787	-0.1354	0.5674	3e-04	0.0015	1.096	0.6228	0.9110
EEL – GE	$\beta_{9\mu}$	0.3767	0.1816	0.0249	0.7379	3e-04	0.0015	1.043	0.9639	0.9650
IAC – GE	$\beta_{10\mu}$	-0.1433	0.1812	-0.4941	0.2171	3e-04	0.0016	1.085	0.2849	0.5431
IAC2000 – GE	$\beta_{11\mu}$	0.1268	0.1775	-0.2161	0.4818	3e-04	0.0015	1.048	0.5693	0.1172
IpiguaIAC – GE	$\beta_{12\mu}$	—	—	—	—	—	—	—	—	—
N58 – GE	$\beta_{13\mu}$	0.1121	0.1817	-0.2383	0.4751	3e-04	0.0015	1.099	0.5130	0.3728
N59 – GE	$\beta_{14\mu}$	-0.0349	0.1899	-0.4050	0.3412	3e-04	0.0015	1.106	0.5478	0.8173
Olimpia – GE	$\beta_{15\mu}$	0.1910	0.1826	-0.1610	0.5558	3e-04	0.0015	1.064	0.2311	0.5019
	λ_{μ}	0.0285	0.0212	0.0044	0.0830	0e+00	0.0007	1.022	0.4998	0.4891

○ Bold lines indicate effects that are not significant or that are statistically equivalent to those in the reference.

Table 10 – Point and interval estimates and convergence criteria for regression coefficients and hyperparameters associated with the parameter σ of the ZIBe(ν, μ, σ) distribution.

				95% Credibility Interval		Standard Error		Convergence Diagnosis		
Associated variable	Parameter	Posteriori Mean	Standard Deviation	2.5%	97.5%	Naive	Time-Series	GR statistics	HW p-value	GW p-value
Intercepto	$\beta_{0\sigma}$	2.6780	0.0791	2.5242	2.8351	1e-04	0.0007	0.996	0.3028	0.8771

To understand possible conclusions more visually, consider the color matrix associated with the information extracted from Table 8. In this context, if a plant from the reference combination (or the rootstock or genotype classified as statistically equivalent) is considered to have a chance of being infected equal to one unit, then, in Figure 20 (3, 6, 9, 12 and 15 months), the equivalent chances (of the reference combination) for observing a zero incidence in any other combination of rootstock and genotype are shown. From this matrix it can be concluded that:

- The most positively expressive rootstock, whose estimated odds of observing a zero incidence are the largest, in general, is *Laranja Caipira*. On the other hand, the *Limao Cravo* rootstock is the most fragile in this sense, followed by the *Tangerina Sunki* and *Tangerina Cleopatra*.
- The genotypes that stand out most positively are *IpiguaIAC*, *Arapongas*, *EEL* and *N59*, all statistically equal, followed by genotypes *N58*, *Olimpia* and the others.
- All estimates for the main odds ratios of interest can be found in the matrix.

ROOTSTOCKS	IpigualAC	Arapongas	Bianchi	EEL	IAC	IAC2000	N58	N59	Olimpia	
15	0.0218	0.0218	0.0083	0.0218	0.0067	0.0074	0.0086	0.0218	0.0088	Tangerina Sunki
12	0.0189	0.0189	0.0073	0.0189	0.0059	0.0064	0.0075	0.0189	0.0077	
9	0.0206	0.0206	0.0079	0.0206	0.0064	0.0070	0.0081	0.0206	0.0083	
6	0.0279	0.0279	0.0107	0.0279	0.0086	0.0094	0.0110	0.0279	0.0113	
3	0.0470	0.0470	0.0180	0.0470	0.0145	0.0159	0.0186	0.0470	0.0191	
15	0.0270	0.0270	0.0103	0.0270	0.0083	0.0091	0.0107	0.0270	0.0109	Tangerina Cleopatra
12	0.0235	0.0235	0.0090	0.0235	0.0073	0.0080	0.0093	0.0235	0.0095	
9	0.0255	0.0255	0.0098	0.0255	0.0079	0.0086	0.0101	0.0255	0.0103	
6	0.0345	0.0345	0.0132	0.0345	0.0107	0.0117	0.0137	0.0345	0.0140	
3	0.0583	0.0583	0.0223	0.0583	0.0180	0.0197	0.0231	0.0583	0.0236	
15	0.0065	0.0065	0.0025	0.0065	0.0020	0.0022	0.0026	0.0065	0.0026	Limao Cravo
12	0.0057	0.0057	0.0022	0.0057	0.0018	0.0019	0.0022	0.0057	0.0023	
9	0.0062	0.0062	0.0024	0.0062	0.0019	0.0021	0.0024	0.0062	0.0025	
6	0.0083	0.0083	0.0032	0.0083	0.0026	0.0028	0.0033	0.0083	0.0034	
3	0.0141	0.0141	0.0054	0.0141	0.0043	0.0048	0.0056	0.0141	0.0057	
15	0.0836	0.0836	0.0320	0.0836	0.0258	0.0283	0.0331	0.0836	0.0339	Laranja Caipira
12	0.0728	0.0728	0.0279	0.0728	0.0225	0.0247	0.0288	0.0728	0.0295	
9	0.0790	0.0790	0.0303	0.0790	0.0244	0.0268	0.0313	0.0790	0.0320	
6	0.1070	0.1070	0.0410	0.1070	0.0331	0.0363	0.0424	0.1070	0.0433	
3	1,0000 (ref)	0.1808	0.0693	0.1808	0.0559	0.0612	0.0716	0.1808	0.0732	

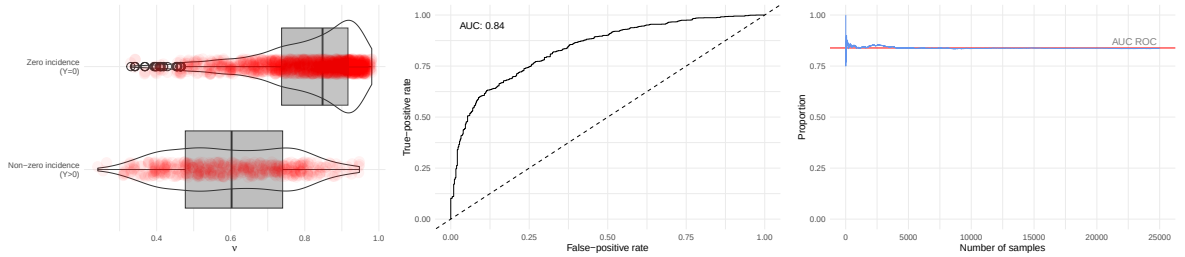
GENOTYPES

Figure 20 – Color matrix odds ratio estimates, adjusted by ZIBe(ν, μ, σ) regression model, with normal random effects, for the occurrence of a zero incidence according to the genotype, rootstocks, and evaluation after 3, 6, 9, 12 and 15 months.

Source: The authors

To gather some evidence regarding the quality of the fit, one can also observe its result from the perspective of the parameters ν and μ. In Figure 21a, it can be seen that the highest values of ν are concentrated in the case where there is no incidence ($Y = 0$), as expected, since

we defined $\Pr(Y = 0) = \nu$. In Figures 21b and 21c, a reasonable predictive power of the inflated part of the model is observed, based on ROC curve.



(a) (b) (c)

Figure 21 – (a) Predicted values of ν for each record in the dataset, compared with the actual observed condition of incidence, $Y = 0$ or $Y > 0$. (b) ROC curve associated with the estimated probabilities $\nu = \Pr(Y = 0)$ e $1 - \nu = \Pr(Y > 0)$. (c) Proportion of parametric bootstrap samples for which $\hat{\nu}$ is higher for $Y = 0$ case than for $Y > 0$ case.

Source: The authors

On the other hand, from the perspective of the model for $Y > 0$, Figure 22 shows, that the pattern observed in the collected data is reasonably captured by the proposed model, despite the occurrence of quite discrepant values in some combinations, such as the expressive "Arapongas-Limao Cravo" and "EEL-Tangerina Sunki".

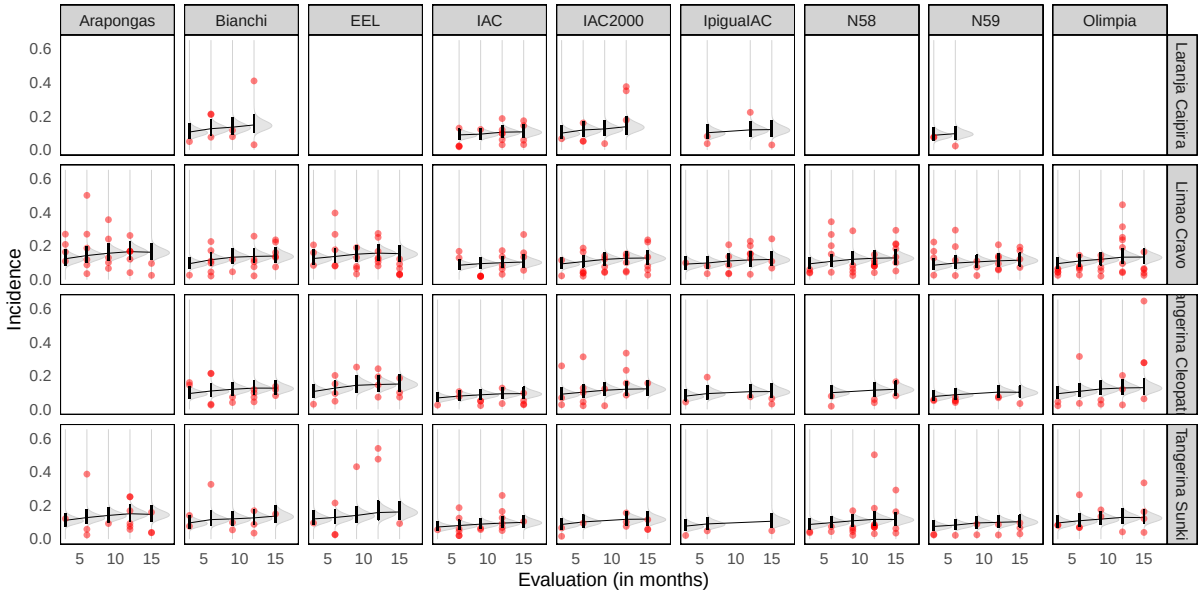


Figure 22 – For each genotype, rootstock, and evaluation, the observed values (in red dots) and the mean $\mu = \mathbb{E}(Y \mid Y > 0)$ with its respective credibility interval resulting from the model.

Source: The authors

In light of the results presented, which aim to identify combinations of rootstocks resistant to citrus canker, this research also seeks to encourage the eradication of one of the most important agricultural diseases affecting citrus crops (citrus canker itself). That is, since the vast majority of farmers around the world no longer practice agriculture by "imitating nature" (which would

require a radical change in the way of life of the nearly 8 billion inhabitants of Earth), for now, in addition to the various genetic improvement methods known to science, the best course of action is to find plants that already possess resistance to potential diseases or to prevent these diseases from reaching the crops. Another approach, no less important, would be to combine certain plants that could produce orange orchards more resistant to citrus canker (and other diseases such as greening (Candidatus Liberibacter spp), for example).

It is important to remember that research of this nature (such as the one presented in this article) is extremely significant, given the cultivation methods that farmers around the world are encouraged to practice, primarily due to the financial support provided by their own governments to promote increasing monoculture.

However, we must remain vigilant and increasingly concerned with plant (and also animal) production methods worldwide. Achieving greater production in smaller land areas and with better quality is always a focus for farmers. Yet, extensive rural extension work is necessary so that farmers are well-guided and have access not only to the technologies developed (in a simple, direct manner with minimal guidance). It is essential that information reaches farmers (and that they put it into practice), particularly scientific knowledge and all necessary management practices for agricultural production aiming at the sustainability tripod: economic, social, and environmental.

In this way, the whole society benefits. By advancing the best scientific practices already discovered, technologies will last much longer, costs will decrease, production will be more sustainable, among many other benefits already identified by researchers in the field.

4.3 Conclusion

This approach has considered the information in the data set more clearly, taking into account issues not considered in the other analyses. The influence of the rootstock and genotype on the incidence among the already infected plants can be clarified.

The modeling based on ZIBe distribution was adequate to understand the behavior of leaf citrus canker incidence, highlighting aspects inherent to the intrinsic resistance of combination rootstock and genotype, as well as the possibility of infection as to the already infected plant resistance.

A model, such as this one, provides sufficient and relevant information for decision-making, indicating the best combination of rootstock and genotype when the priority criterion is resistant to leaf citrus canker. In addition, such an analysis can be applied to many other characteristics of interest.

While our initial choice aimed to use non-informative priors—assigning a flat normal prior $\mathrm{N}(0, 0.001)$ to each component of the regression vector $\boldsymbol{\beta}$ and a flat gamma prior

Gamma$(0.01, 0.01)$ for each component of the dispersion parameter vector $\boldsymbol{\lambda}$—we recognize that some "flat" priors can still introduce bias under certain conditions. In this context a sensitivity analysis would be valuable to assess the robustness of the conclusions.

CHAPTER

5

CONCLUSION AND FUTURE PROPOSAL

5.1 Conclusions of Chapter 2

After all the necessary meetings and studies, supported by theoretical foundations in statistics and agronomy, it was possible to develop a platform capable of generating sampling plans using georeferenced data from orange groves. Thus, all collaborators involved in inspecting the citrus greening disease in orange orchards will be able to perform their work much faster and more effectively, without the need to conduct a full census of citrus plantations.

5.2 Future Proposal of Chapter 2

Implement a third methodology on the platform (while maintaining the two initial ones presented in Chapter 2). Specifically, incorporate the methodology developed by Vila *et al.* (2024), where the authors modified the beta distribution to capture bimodality in proportion data. This enhancement may lead to more robust results for the platform.

Integrate the results obtained and the platform developed in Chapter 2 with drones, ensuring that 100% of the monitoring and data collection activities become feasible in an automated and remote manner.

5.3 Conclusions of Chapter 3 (part 1)

The generalized nonlinear model proved a reasonable alternative approach. Using the characteristic, from the adjusted growth curves, would endure the interpretability of the parameters allowed to characterize interesting aspects, like the exposure of several orange genotypes to citrus canker disease. Statistical evidence points out that the most resistance to citrus canker disease were Irradiada, Valencia and Valencia 2, respectively. In contrast, the Hamlin genotype

was the most susceptible to the disease.

5.4 Future Proposals for Chapter 3 (Part 1)

Seek partnerships for further work addressing similar problems, so that we can propose new types of solutions for cases where the data exhibit characteristics identical or similar to those in this chapter. I believe that countless researchers in the field of agricultural sciences have problems to be solved in this same context.

5.5 Conclusions of Chapter 3 (part 2)

The proposed method demonstrates a strong capability to evaluate the statistical repeatability of results in agronomic trials. By leveraging posterior distributions and comparing them across multiple data collections, the method provides a robust for assessing consistency in experimental outcomes. This allows for the detection of whether key parameters, such as growth rates or disease resistance measures, remain stable across repeated trials. As a result, the method reduces the subjectivity associated with visual comparisons and enhances the reliability of interpretations, particularly for genotypes under study.

Furthermore, the statistical repeatability of the method ensures that any observed variability can be attributed to true biological differences rather than experimental errors or random fluctuations. This is particularly valuable in agronomic studies, where environmental factors and natural variability can complicate the interpretation of results. By confirming that core parameters remain consistent across trials, the method provides stronger evidence for the validity and robustness of agronomic findings, allowing for more confident decision-making in breeding programs or agricultural management strategies.

5.6 Future Proposals for Chapter 3 (Part 2)

Seek partnerships for further work addressing similar problems, so that we can propose new types of solutions for cases where the data exhibit characteristics identical or similar to those in this chapter. I believe that countless researchers in the field of agricultural sciences have problems to be solved in this same context.

5.7 Conclusions of Chapter 4

This approach has considered the information in the data set more clearly, taking into account issues not considered in the other analyses. The influence of the rootstock and genotype on the incidence among the already infected plants can be clarified.

The modeling based on ZIBe distribution was adequate to understand the behavior of leaf citrus canker incidence, highlighting aspects inherent to the intrinsic resistance of combination rootstock and genotype, as well as the possibility of infection as to the already infected plant resistance.

A model, such as this one, provides sufficient and relevant information for decision-making, indicating the best combination of rootstock and genotype when the priority criterion is resistant to leaf citrus canker. In addition, such an analysis can be applied to many other characteristics of interest.

While our initial choice aimed to use non-informative priors—assigning a flat normal prior $\mathrm{N}(0, 0.001)$ to each component of the regression vector $\boldsymbol{\beta}$ and a flat gamma prior $\mathrm{Gamma}(0.01, 0.01)$ for each component of the dispersion parameter vector $\boldsymbol{\lambda}$—we recognize that some "flat" priors can still introduce bias under certain conditions. In this context a sensitivity analysis would be valuable to assess the robustness of the conclusions.

5.8 Future Proposals for Chapter 4

As soon as possible, seek more modern and even adapted methodologies to conduct work aimed at solving problems similar to the one presented in Chapter 4.

BIBLIOGRAPHY

ABDEL-KARIM, A. H. Extended zero-one inflated beta and adjusted three-part regression models for proportional data analysis. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 46, n. 8, p. 6155–6172, 2017. Citations on pages 24 and 25.

ADVANCES in Weed Science. Sociedade Brasileira da Ciência das Plantas Daninhas (SBCPD), 2024. Acesso em: 25 nov. 2024. Available: <https://awsjournal.org/>. Citation on page 21.

AGHAYERASHTI, M.; SAMANI, E. B.; GANJALI, M. Bayesian latent variable model of mixed correlated rank and beta-binomial responses with missing data for the international statistical literacy project poster competition. **Sankhya B**, Springer, v. 85, n. 1, p. 210–250, 2023. Citation on page 26.

ALVES[1], G. R.; BELOTI[1], V. H. *et al.* Manejo e controle do huanglongbing (hlb) dos cítricos. **Investigación Agraria**, Facultad de Ciencias Agrarias, UNA, v. 16, n. 2, p. 69–82, 2014. Citations on pages 31 and 36.

AMORIM, L.; REZENDE, J. A. M.; CAMARGO, L. F. A. **Manual de Fitopatologia: Volume 2 - Doenças das Plantas Cultivadas**. 3ª edição. ed. Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo: Editora Agronômica Ceres, 1997. ISBN 9788531800535. Citations on pages 31 and 36.

ANHOLETO, T.; SANDOVAL, M. C.; BOTTER, D. A. Adjusted Pearson Residuals in Beta Regression Models. **Journal of Statistical Computation and Simulation**, Informa UK Limited, v. 84, n. 5, p. 999–1014, out 2012. Citations on pages 79 and 80.

BAREIKIS, G.; MANSTAVIČIUS, E. Construction of the beta distributions using the random permutation divisors. **Nonlinear Analysis: Modelling and Control**, p. 1–16, 2024. Citation on page 24.

BARRERO, E. C. P. D. N. M. L.; (VICE), P. D. J. L. G. **Revista Semina: Ciências Agrárias**. 2023. Acesso em: 22/12/2023. MIAR: ICDS 2021: 11.0. Fatores de impacto: JCR -2021 - 0,595 - SJR - 0,5 - H index - 18. ISSN - JRC: 1676-546X. Sobre a Revista: A Revista Semina: Ciências Agrárias é uma publicação bimestral associada à Universidade Estadual de Londrina, focada em Ciências Agrárias, Zootecnia, Ciências Alimentares e Medicina Veterinária. Available: <https://ojs.uel.br/revistas/uel/index.php/semagrarias/about/submissions>. Citation on page 32.

BASSANEZI, R. **Greening, que devastou laranjais da Flórida, se dissemina em São Paulo**. 2023. Autor: Renato Bassanezi, pesquisador da Fundecitrus. Acessado em: data de acesso (insira a data de acesso aqui). Available: <https://globorural.globo.com/agricultura/noticia/2023/09/greening-que-devastou-laranjais-da-florida-se-dissemina-em-sao-paulo.ghtml>. Citation on page 32.

BASSANEZI, R. B.; BEHLAU, F.; LOPES, S. A.; WULFF, N. A.; MIRANDA, M. P. de; BARBOSA, J. C. **Manual Técnico sobre Defesa da Citricultura**. Araraquara, SP: Fundo de Defesa da Citricultura (Fundecitrus), 2024. Capa e diagramação: Juliana Retamero. Revisão linguística e

final: Viviane Moura e Rafael de Paula. Ficha catalográfica elaborada pela Biblioteca Fundecitrus. Todos os direitos reservados. Available: <https://www.fundecitrus.com.br>. Citation on page 21.

BEHLAU, F.; AMORIM, L.; BELASQUE, J. J.; FILHO, A. B.; JR, R. P. L.; GRAHAM, J. H.; GOTTWALD, T. R. Annual and polyetic progression of citrus canker on trees protected with copper sprays. **Plant Pathology**, v. 59, p. 1031–1036, 2010. Citation on page 34.

BEHLAU, F.; BELASQUE, J. R.; FILHO, A. B.; GRAHAM, J. H.; LEITE, J. R. P.; GOTTWALD, T. R. Copper sprays and windbreaks for control of citrus canker on young orange trees in southern brazil. **Crop Protection**, v. 27, p. 807–813, 2008. Citation on page 34.

BERNARDO, J. M.; SMITH, A. F. M. **Bayesian Theory**. [S.l.]: John Wiley & Sons, 1994. Citation on page 60.

BIEDERMANN, A.; KOTSOGLOU, K. N. Commentary on "three-way rocs for forensic decision making" by nicholas scurich and richard s. john (in: Statistics and public policy). **Statistics and Public Policy**, Taylor & Francis, v. 11, n. 1, p. 2288166, 2024. Citation on page 19.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006. Citation on page 61.

BOCK, C. H.; PARKER, P. E.; GOTTWALD, T. R. Effect of simulated wind-driven rain on duration and distance of dispersal of xanthomonas axonopodis pv. citri from canker-infected citrus trees. **Plant Disease**, v. 89, p. 71–80, 2005. Citation on page 34.

BOTEON, M. **HF BRASIL/CEPEA: greening avança em SP e eleva gastos na citricultura**. Piracicaba, SP, 2023. Pesquisadora: Margarete Boteon. Acessado em: 12/12/2023. Available: <https://www.cepea.esalq.usp.br/br/releases/hf-brasil-cepea-greening-avanca-em-sp-e-eleva-gastos-na-citricultura.aspx>. Citation on page 32.

BOX, G. E. P.; HUNTER, J. S.; G, H. W. **Statistics for Experimenters**. 2ª. ed. New York: Wiley, 2005. 633 p. Citations on pages 19 and 70.

BRADLEY, P. C.; THOMAS, A. L. **Bayesian Methods for Data Analysis**. 1. ed. [S.l.]: A Chapman and Hall Book, 2008. 529 p. Citations on pages 64 and 82.

BRASWELL, W.; PARK, J.; STANSLY, P.; KOSTYK, B.; LOUZADA, E.; GRAÇA, J. da; KUNTA, M. Root samples provide early and improved detection of candidatus liberibacter asiaticus in. **Citrus. Sci. Rep.**, v. 10, n. 1, p. 16982, 2020. Citation on page 20.

BRIEN, C. J.; DEMÉTRIO, C. G. B. Formulating mixed models for experiments, including longitudinal experiments. **Journal of agricultural, biological, and environmental statistics**, Springer, v. 14, p. 253–280, 2009. Citation on page 19.

BROOKS, S.; GELMAN, A.; JONES, G.; MENG, X.-L. (Ed.). **Handbook of Markov Chain Monte Carlo**. [S.l.]: Chapman and Hall/CRC, 2011. Citation on page 61.

BROOKS, S. P.; GELMAN, A. General methods for monitoring convergence of iterative simulations. **Journal of Computational and Graphical Statistics**, Informa UK Limited, v. 7, n. 4, p. 434–455, dec 1998. Citations on pages 64 and 83.

BROOKS, S. P.; ROBERTS, G. O. Assessing convergence of markov chain monte carlo algorithms. **Statistics and Computing**, v. 8, p. 319–335, 1998. Citation on page 65.

BROOKS, S. P.; SMITH, J.; VEHTARI, A.; PLUMMER, M.; STONE, M.; ROBERT, C. P.; TITTERINGTON, D. M.; NELDER, J. A.; ATKINSON, A.; DAWID, A. P.; LAWSON, A.; CLARK, A.; BERNARDO, J. M.; SAHU, S. K.; RICHARDSON, S.; GREEN, P.; BURNHAM, K. P.; DEIORIO, M.; ROBERT, C. P.; DRAPER, D. **Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde**. Blackwell Publ Ltda, 2002. 616–639 p. Available: <http://wrap.warwick.ac.uk/10409/>. Citations on pages 64 and 82.

BROWN, R.; MAYER, D. G. Representing cumulative germination. 2. the use of the weibull function and other empirically derived curves. **Annals of Botany**, Oxford University Press, v. 61, n. 2, p. 127–138, 1988. Citation on page 28.

BUCHWALD, P. A general bilinear model to describe growth or decline time profiles. **Mathematical biosciences**, Elsevier, v. 205, n. 1, p. 108–136, 2007. Citation on page 29.

BüRKNER, P.-C. Bayesian item response modeling in R with brms and Stan. **Journal of Statistical Software**, v. 100, n. 5, p. 1–54, 2021. Citation on page 83.

CASELLA, G.; BERGER, R. L. **Inferência estatística**. [S.l.]: Cengage Learning, 2010. Citation on page 22.

CEPEDA-CUERVO, E. **Variability Modeling in Generalized Linear Models**. Phd Thesis (doutorado) — Instituto de Matemática – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2001. Citation on page 79.

CEPEDA-CUERVO, E. Beta Regression Models: Joint Mean and Variance Modeling. **Journal of Statistical Theory and Practice**, Informa UK Limited, v. 9, n. 1, p. 134–145, jul 2014. Citation on page 79.

CHAN, S. N. S.; AFUECHETA, E. An r package for value at risk and expected shortfall. **Communications in Statistics - Simulation and Computation**, Taylor Francis, v. 45, n. 9, p. 3416–3434, 2016. Available: <https://doi.org/10.1080/03610918.2014.944658>. Citation on page 30.

CHATFIELD, C.; GOODHARDT, G. J. The beta-binomial model for consumer purchasing behaviour. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 19, n. 3, p. 240–250, 1970. Citation on page 27.

CHEN, H. Use of linear, weibull, and log-logistic functions to model pressure inactivation of seven foodborne pathogens in milk. **Food Microbiology**, Elsevier, v. 24, n. 3, p. 197–204, 2007. Citation on page 28.

CHEN, H.; HOOVER, D. G. Pressure inactivation kinetics of yersinia enterocolitica atcc 35669. **International Journal of Food Microbiology**, Elsevier, v. 87, n. 1-2, p. 161–171, 2003. Citation on page 28.

CHEN, H.; LI, Q.; ZHU, F. A covariate-driven beta-binomial integer-valued garch model for bounded counts with an application. **Metrika**, Springer, p. 1–22, 2023. Citation on page 26.

CHICHARRO, J. L.; VAQUERO, A. F. **Fisiología del Ejercicio**. 3. ed. [S.l.]: Editora Médica Panamericana, 2006. 1008 p. Citation on page 63.

CHIEN, L.-C. Diagnostic Plots in Beta-Regression Models. **Journal of Applied Statistics**, Informa UK Limited, v. 38, n. 8, p. 1607–1622, ago 2011. Citations on pages 79 and 80.

CHOI, T. J. A rotationally invariant stochastic opposition-based learning using a beta distribution in differential evolution. **Expert Systems with Applications**, Elsevier, p. 120658, 2023. Citation on page 24.

CHRISTENSEN, R.; JOHNSON, W.; BRANSCUM, A.; HANSON, T. E. **Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians**. [S.l.]: CRC Press, 2010. Citation on page 60.

Christopher H. Jackson. Multi-state models for panel data: The msm package for R. **Journal of Statistical Software**, v. 38, n. 8, p. 1–29, 2011. Citation on page 82.

CITRICULTURA, F. C. e sustentabilidade para a (Ed.). **Levantamento da incidência das doenças dos citros: greening, cvc e cancro cítrico no cinturão citrícola de São Paulo e triângulo/sudoeste mineiro**. Araraquara - São Paulo: [s.n.], 2021. 77 p. Capa e diagramação por Valmir Aparecido Campos. Revisão linguística e final por Beatriz Flório. Editado pelo Fundo de Defesa da Citricultura. Responsáveis: Pesquisadores - Renato Beozzo Bassanezi (Fundecitrus), Franklin Behlau (Fundecitrus), Silvio Aparecido Lopes (Fundecitrus), Nelson Arno Wulff (Fundecitrus), Marcelo Pedreira de Miranda (Fundecitrus), José Carlos Barbosa (UNESP/FCAV). Coordenação - Antonio Juliano Ayres (Fundecitrus), Ivaldo Sala (Fundecitrus), Vinicius Gustavo Trombin (Markestrat). Available: <http://www.fundecitrus.com.br>. Citation on page 31.

CLEARY, R. J. **Statistics: Informed decisions using data and an introduction to statistical analysis for business and industry: A problem solving approach**. [S.l.]: Taylor & Francis, 2008. Citation on page 19.

ĆMIEL, B.; NAWAŁA, J.; JANOWSKI, L.; RUSEK, K. Generalised score distribution: under-dispersed continuation of the beta-binomial distribution. **Statistical Papers**, Springer, p. 1–33, 2023. Citation on page 26.

COMULADA, W. S.; WEISS, R. E. On models for binomial data with random numbers of trials. **Biometrics**, Oxford University Press, v. 63, n. 2, p. 610–617, 2007. Citation on page 23.

CORDEIRO, G. M.; PESCIM, R. R.; DEMÉTRIO, C. G.; ORTEGA, E. M. The kummer beta generalized gamma distribution. **Journal of Data Science**, , v. 12, n. 4, p. 661–697, 2014. Citation on page 28.

CRESPO, A. A. **Estatística (Série em foco)**. [S.l.]: Saraiva Educação SA, 2020. Citation on page 19.

CRISÓSTOMO, L. A.; NAUMOV, A. (Ed.). **Adubando para Alta Produtividade e Qualidade: Fruteiras Tropicais do Brasil**. Fortaleza, CE: Embrapa Agroindústria Tropical, 2009. 238 p. Citation on page 35.

CROP Science. Wiley, 2024. Acesso em: 25 nov. 2024. Available: <https://onlinelibrary.wiley.com/journal/1439037X>. Citation on page 21.

DAGOGO, J.; CYNTHIA, O. U.; OYINEBIFUN, B. *et al.* Comparative analysis of additive and multiplicative error terms of weibull, logistic gompertz, hills and richards models with four parameters. **Journal of Advances in Mathematics and Computer Science**, v. 38, n. 5, p. 1–34, 2023. Citation on page 28.

DEWDNEY, M. M.; GRAHAM, J. H. **Florida Citrus Production Guide. Citrus Canker**. 2018. 181 p. Available: <http://ccms.farmjournal.com/sites/default/files/inline-files/2017_FLCitrusPestGuide.pdf={06nov.2018}>. Citation on page 34.

DUMELLE, M.; HIGHAM, M.; HOEF, J. M. V.; OLSEN, A. R.; MADSEN, L. A comparison of design-based and model-based approaches for finite population spatial sampling and inference. **Methods in ecology and evolution**, Wiley Online Library, v. 13, n. 9, p. 2018–2029, 2022. Citation on page 20.

DURÁN, C. E. A.; WIVES, D. G. Decision making and agriculture: a recent review of organic farming. **Desenvolvimento em questão: revista do programa de pós-graduação em desenvolvimento [recurso eletrônico]. Ijuí, RS. Vol. 16, n. 43 (abr./jun. 2018), p. 175-199.**, 2018. Citation on page 19.

ENNIS, D. M.; BI, J. The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. **Journal of Sensory Studies**, Wiley Online Library, v. 13, n. 4, p. 389–412, 1998. Citation on page 42.

FABIO, L. C.; PAULA, G. A.; CASTRO, M. de. A poisson mixed model with nonnormal random effect distribution. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 6, p. 1499–1510, 2012. Citation on page 24.

FAWCETT, H. S.; JENKINS, A. E. Records of citrus canker from herbarium specimens of the genus citrus in england and the united states. **Phytopathology**, v. 23, p. 820–824, 1933. Citation on page 34.

FERRARI, S.; CRIBARI-NETO, F. Beta Regression for Modelling Rates and Proportions. **Journal of Applied Statistics**, Informa UK Limited, v. 31, n. 7, p. 799–815, ago 2004. Citation on page 79.

FERRARI, S. L. P.; ESPINHEIRA, P. L.; CRIBARI-NETO, F. Diagnostic Tools in Beta Regression with Varying Dispersion. **Statistica Neerlandica**, John Wiley and Sons, v. 65, n. 3, p. 337–351, 2011. Citations on pages 79 and 80.

FERRARI, S. L. P.; PINHEIRO, E. C. Improved Likelihood Inference in Beta Regression. **Journal of Statistical Computation and Simulation**, Taylor and Francis Group, v. 81, n. 4, p. 431–443, abr 2011. Citation on page 79.

FISHER, R. A. **The Design of Experiments**. Edinburgh, Scotland: Oliver and Boyd, 1935. 272 p. Citation on page 19.

FOSGATE, G. A cluster-adjusted sample size algorithm for proportions was developed using a beta-binomial model. **Journal of clinical epidemiology**, Elsevier, v. 60, n. 3, p. 250–255, 2007. Citations on pages 27 and 51.

GASQUES, J. G.; BASTOS, E. T.; TUBINO, M. A. A.; MIRANDA, M. C.; GOMES, E. G.; SOUZA, G. da Silva e. **Projeções do Agronegócio (Projeções de Longo Prazo) - Brasil 2021/22 a 2031/2032**. Brasília, 2022. Acesso em 01/11/2023. Disponível em: <https://www.gov.br/agricultura/pt-br>. Available: <https://www.gov.br/agricultura/pt-br>. Citation on page 35.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian Data Analysis**. 3rd. ed. [S.l.]: Chapman and Hall/CRC, 2014. Citations on pages 60 and 61.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Institute of Mathematical Statistics, v. 7, n. 4, p. 457–472, nov 1992. Citations on pages 64 and 83.

GERALDELLO, C. S.; SOARES, S. A.; MATHIAS, S. K.; MELLO, F. de C.; CRUZ, S. V. e. **Medidas antidumping e política doméstica: o caso da citricultura estadunidense**. 1. ed. São Paulo, SP: Cultura Acadêmica (Unesp), 2015. ISBN 978-85-7983-665-7. Citation on page 31.

GEWEKE, J. F. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: PRESS, U. (Ed.). **Bayesian Statistics**. [S.l.]: F. R. Minneapolis, 1992. p. 169–193. Citations on pages 64 and 83.

GOLDSTEIN, H. **Multilevel statistical models. London: Institute of Education, Multilevel Models project**. 1999. Citation on page 22.

GOMPERTZ, B. On the nature of the function expressive of the law of human mortality and on a new model of determining life contingencies. In: **Philosophical Transactions of the Royal Society of London**. [S.l.]: doi:10.1098/rstl.1825.0026, 1825. p. 513–585. Citation on page 63.

GONÇALVES-ZULIANI, A. M. O. Resistência de genótipos de laranja doce (citrus sinensis) ao cancro cítrico e diversidade genética de xanthomonas citri subsp. citri. 2014. Citation on page 21.

GONÇALVES-ZULIANI, A. M. O. **Resistência de Genótipos de Laranja Doce *Citrus sinensis* ao Cancro Cítrico e Diversidade de *Xanthomonas Citri* subsp. *citri.*** Phd Thesis (doutorado) — Departamento de Agronomia – Universidade de Estadual de Maringá, Maringá, 2014. Citation on page 78.

GOTTWALD, T. R.; GRAHAM, J. H.; SCHUBERT, T. S. Na epidemiological analysis of the spread of citrus canker in urban miami, florida, an synergistic interaction with the asian leaf miner. **Fruits**, v. 52, p. 383–390, 1997. Citation on page 34.

GOTTWALD, T. R.; SUN, X.; RILEY, T.; GRAHAM, J. H.; FERRANDINO, F.; TAYLOR, E. L. Geo-referenced spatio temporal analysis of the urban citrus canker epidemic in flórida. **Phytopathology**, v. 92, p. 361–377, 2002. Citation on page 34.

GRIFFITHS, D. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. **Biometrics**, JSTOR, p. 637–648, 1973. Citation on page 23.

HEIDELBERGER, P.; WELCH, P. D. Simulation run length control in the presence of an initial transient. **Operations Research**, v. 31, p. 1109–1144, 1983. Citations on pages 64 and 83.

HUGHES, G.; MADDEN, L.; MUNKVOLD, G. Cluster sampling for disease incidence data. **Phytopathology**, Saint Paul, v. 86, n. 2, p. 132–137, 1996. Citations on pages 41, 42, and 48.

HUNT, D. L.; CHENG, C.; POUNDS, S. The beta-binomial distribution for estimating the number of false rejections in microarray gene expression studies. **Computational statistics & data analysis**, Elsevier, v. 53, n. 5, p. 1688–1700, 2009. Citation on page 27.

IBÁÑEZ, M. V.; PRADES, M.; SIMÓ, A. Modelling Municipal Waste Separation Rates Using Generalized Linear Models and Beta Regression. **Resources, Conservation and Recycling**, Elsevier Science, v. 55, n. 12, p. 1129–1138, 2011. Citation on page 77.

JABBAR, A. K.; AL-SAEDI, H. M. Mathematical modeling of tumors growth: Competition based on gompertz model in two dimensions. **Baghdad Science Journal**, January 2024. ISSN 2078-8665. Published Online First: January 2024. Citation on page 28.

JORGENSEN, B. **The Theory of Dispersion Models**. London: Chapman & Hall, 1997. 247 p. Citation on page 79.

JUNIOR, O. A. G.; GUEDES, T. A.; GONçALVES-ZULIANE, A. M. O.; NUNES, W. M. d. C. Zero-inflated beta regression model for leaf citrus canker incidence in orange genotypes grafted onto different rootstocks. **Acta Scientiarum. Biological Sciences**, v. 39, n. 2, p. 161–171, 2017. Available: <https://doi.org/10.4025/actascibiolsci.v39i2.33063>. Citation on page 31.

JúNIOR, G. S. **Seminário Internacional - Desafios da citricultura europeia e aspectos da fisiologia da frutificação**. Av. Adhemar Pereira de Barros, 201, Araraquara-SP, Brasil. Seminário realizado na Fundecitrus. Mais informações disponíveis em <https://www.fundecitrus.com.br/comunicacao/noticias/integra/seminario-internacional-discute-citricultura-na-europa-e-fisiologia-da-frutificacao/1595>. Available: <https://www.fundecitrus.com.br/evento/210>. Citation on page 31.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression Analysis of Variates Observed on (0,1): Percentages, Proportions and Fractions. **Statistical Modelling**, v. 3, n. 3, p. 193–213, 2003. Citation on page 79.

KIMANI, P. K. Asymmetrical fixed-bed breakthrough curve modelling: Comparing simplistic, log-modified, fractal-like, and probability distribution function models. **Chemical Engineering Research and Design**, Elsevier, v. 201, p. 446–456, 2024. Citation on page 28.

KONG, F. L. with contributions from Y. **zoib: Bayesian Inference for Beta Regression and Zero-or-One Inflated Beta Regression**. [S.l.], 2023. R package version 1.6. Available: <https://CRAN.R-project.org/package=zoib>. Citation on page 83.

KORHONEN, L.; KORHONEN, K. T.; STENBERG, P.; MALTAMO, M.; RAUTIAINEN, M. Local Models for Forest Canopy Cover with Beta Regression. **Silva Fennica**, v. 41, n. 4, p. 671–685, 2007. Citation on page 77.

LATIF, S.; YAB, M. Z. D-Optimal Designs for Beta Regression Models with Single Predictor. **Journal of Statistical Computation and Simulation**, Taylor and Francis Group, v. 85, n. 9, p. 1709–1724, jun 2015. Citation on page 80.

LECOUTERE, E.; CHU, L. Supporting women's empowerment by changing intra-household decision-making: A mixed-methods analysis of a field experiment in rural south-west tanzania. **Development Policy Review**, Wiley Online Library, v. 42, n. 3, p. e12758, 2024. Citation on page 19.

LIN, C.-S.; POUSHINSKY, G.; MAUER, M. An examination of five sampling methods under random and clustered disease distributions using simulation. **Canadian Journal of Plant Science**, NRC Research Press Ottawa, Canada, v. 59, n. 1, p. 121–130, 1979. Citation on page 56.

LORA, M. I.; SINGER, J. M. Beta-binomial/poisson regression models for repeated bivariate counts. **Statistics in medicine**, Wiley Online Library, v. 27, n. 17, p. 3366–3381, 2008. Citation on page 23.

_____. Beta-binomial/gamma-Poisson regression models for repeated counts with random parameters. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 25, n. 2, p. 218 – 235, 2011. Available: <https://doi.org/10.1214/10-BJPS118>. Citation on page 23.

MALUF, Y. S.; FERRARI, S. L. P.; QUEIROZ, F. F. Robust beta regression through the logit transformation. **Metrika**, Springer, v. 87, n. 1, p. 1–21, 2024. Citation on page 80.

MARÍN, X. F. i. ggmcmc: Analysis of MCMC samples and Bayesian inference. **Journal of Statistical Software**, v. 70, n. 9, p. 1–20, 2016. Citation on page 83.

MARQUES, P. J. **Doença da laranja preocupa fruticultores do noroeste do Paraná**. 2023. Autor: Paulo Jorge Marques - Coordenador de vigilância da fruticultura da Adapar. Acessado em: 30/11/2023. Available: <https://g1.globo.com/pr/parana/caminhos-do-campo/noticia/2023/06/18/doenca-da-laranja-preocupa-fruticultores-do-noroeste-do-parana.ghtml>. Citation on page 31.

MARTIN, O. Da estatística política à sociologia estatística. desenvolvimento e transformações da análise estatística da sociedade (séculos xvii-xix). **Revista brasileira de História**, SciELO Brasil, v. 21, p. 13–34, 2001. Citation on page 19.

MARTINEZ, R. O. **Modelos de regressão beta inflacionados**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2008. Citation on page 24.

MASOUDI, A.; AZARFAR, A. Comparison of nonlinear models describing growth curves of broiler chickens fed on different levels of corn bran. **International Journal of Avian and Wildlife Biology**, v. 2, n. 1, p. 1–7, 2017. Citation on page 29.

MAZUCHELI, J.; EMANUELLI, I. P. Aplicação da distribuição nakagami na análise de dados de precipitação. **Revista Brasileira de Meteorologia**, rbmet.org.br, v. 34, n. 1, p. 1–7, 2019. Citation on page 30.

MCKELLAR, R. C.; LU, X. **Modeling microbial responses in food**. [S.l.]: CRC press, 2003. Citation on page 28.

MENDONÇA, A.; SILVESTRE, J.; PASSOS, J. The Shrinking Endogeneity of Optimum Currency Areas Criteria: Evidence from the European Monetary Union – A Beta Regression Approach. **Economics Letters**, Elsevier Science, v. 113, n. 1, p. 65–69, 2011. Citation on page 77.

MERWE, A. Van der. Chromium and manganese toxicity. is it important in transvaal citrus greening? **Farming S. Afr.**, v. 12, p. 439–440, 1937. Citation on page 31.

MOALA, F. A.; RAMOS, P. L.; ACHCAR, J. A. Bayesian inference for two-parameter gamma distribution assuming different noninformative priors. **Revista Colombiana de Estadística**, Universidad Nacional de Colombia, v. 36, n. 2, p. 319–336, 2013. Citation on page 28.

MOLENBERGHS, G.; VERBEKE, G.; EFENDI, A.; BRAEKERS, R.; DEMÉTRIO, C. G. A combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event data. **Statistical Methods in Medical Research**, SAGE Publications Sage UK: London, England, v. 24, n. 4, p. 434–452, 2015. Citation on page 28.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 7ª. ed. New York: Wiley, 2008. 656 p. Citations on pages 19 and 70.

MORAES, R. M.; ROCHA, A. V.; MACHADO, L. S. Intelligent Assessment Based on Beta Regression for Realistic Training on Medical Simulators. **Knowledge-Based Systems**, Elsevier Science, v. 32, n. 1, p. 3–8, 2012. Citation on page 77.

MULLEN, R.; MARSHALL, L.; MCGLYNN, B. A Beta Regression Model for Improved Solar Radiation Predictions. **Journal of Applied Meteorology and Climatology**, v. 52, n. 8, p. 1923–1938, 2013. Citation on page 77.

NADARAJAH, S. C. S.; AFUECHETA, E. Tabulations for value at risk and expected shortfall. **Communications in Statistics - Theory and Methods**, Taylor Francis, v. 46, n. 12, p. 5956–5984, 2017. Available: <https://doi.org/10.1080/03610926.2015.1116572>. Citation on page 30.

NAJERA-ZULOAGA, J.; LEE, D.-J.; ESTEBAN, C.; AROSTEGUI, I. Multidimensional beta-binomial regression model: A joint analysis of patient-reported outcomes. **Statistical Modelling**, SAGE Publications Sage India: New Delhi, India, p. 1471082X231151311, 2023. Citation on page 26.

NARASIMHAN, B.; JOHNSON, S. G.; HAHN, T.; BOUVIER, A.; KIêU, K. **cubature: Adaptive Multivariate Integration over Hypercubes**. [S.l.], 2023. R package version 2.1.0. Available: <https://CRAN.R-project.org/package=cubature>. Citation on page 82.

NASCIMENTO, D. C.; JUNIOR, O. A. G.; ELAL-OLIVERO, D.; BONNAIL, E.; FERREIRA, P. H. Statistical process control (spc) for double-bounded information: Choosing wisely the parametric family for unit data. **Quality Engineering**, Taylor & Francis, p. 1–19, 2023. Citations on pages 42 and 43.

NAVARRO, D.; PERFORS, A. **An introduction to the Beta-Binomial model**. Adelaide - Austrália, 2012. 1-9 p. Notas de aula. Citation on page 28.

NAWA, V. M.; NADARAJAH, S. New closed form estimators for the beta distribution. **Mathematics**, Multidisciplinary Digital Publishing Institute, v. 11, n. 13, p. 2799, 2023. Citation on page 24.

____. Closed form estimators for a bivariate beta distribution. **Journal of Computational and Applied Mathematics**, Elsevier, v. 439, p. 115603, 2024. Citation on page 24.

NEGRAO, I. d. O.; AQUINO, L. H. d.; BEARZOTI, E. Avaliação de quatro métodos de estimação dos parâmetros da distribuição beta-binomial pela simulação monte carlo. **Ciências Agropecuárias**, v. 25, n. 6, p. 1370–1381, nov 2001. Citation on page 27.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, v. 135, n. 3, p. 370–384, Aug. 1972. Citation on page 79.

NETTO, I.; REGINA, C. **A produção de citros está em alta**. 2023. Disponível em: <https://citrusbr.com/noticias/a-producao-de-citros-esta-em-alta/>. Available: <https://citrusbr.com/noticias/a-producao-de-citros-esta-em-alta/>. Citation on page 35.

ORTEGA, E. M.; RIZZATO, F. B.; DEMÉTRIO, C. G. The generalized log-gamma mixture model with covariates: local influence and residual analysis. **Statistical Methods and Applications**, Springer, v. 18, p. 305–331, 2009. Citation on page 28.

OSPINA, R.; FERRARI, S. L. A General Class of Zero-or-One Inflated Beta Regression Models. **Computational Statistics & Data Analysis**, Elsevier BV, v. 56, n. 6, p. 1609–1623, jun 2012. Citation on page 80.

____. A general class of zero-or-one inflated beta regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 6, p. 1609–1623, 2012. Citation on page 24.

PADMANABHAM, D.; VIDHYASEKARAN, P.; RAJAGOPALAN, C. K. S. Changes in photosynthesis and carbohydrate content in canker and halo regions in xanthomonas citri infected citrus leaves. **Indian Journal Phytopathology**, v. 26, p. 215–217, 1973. Citation on page 34.

PAOLINO, P. Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables. **Political Analysis**, v. 9, n. 4, p. 325–346, 2001. Citation on page 79.

PENG, X.; LI, G.; LIU, Z. Zero-inflated beta regression for differential abundance analysis with metagenomics data. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 23, n. 2, p. 102–110, 2016. Citations on pages 24 and 25.

PEREIRA, G. H.; BOTTER, D. A.; SANDOVAL, M. C. The truncated inflated beta distribution. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 41, n. 5, p. 907–919, 2012. Citations on pages 24 and 25.

PEREIRA, T. L.; CRIBARI-NETO, F. Modified Likelihood Ratio Statistics for Inflated Beta Regressions. **Journal of Statistical Computation and Simulation**, Taylor and Francis Group, v. 84, n. 5, p. 982–998, mai 2012. Citation on page 80.

PLUMMER, M. **rjags: Bayesian Graphical Models using MCMC**. [S.l.], 2018. R package version 4-8. Available: <https://CRAN.R-project.org/package=rjags>. Citations on pages 82 and 83.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for mcmc. **R News**, v. 6, n. 1, p. 7–11, 2006. Available: <https://journal.r-project.org/archive/>. Citations on pages 65, 82, and 83.

POMPEU, J. J.; BLUMER, S. A introdução de germolpasma: uma contribuição ao melhoramento dos citros. **Laranja**, v. 27, p. 341–354, 2006. Citation on page 35.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Available: <https://www.R-project.org/>. Citations on pages 65 and 83.

RAMOS, E.; EGBON, O. A.; RAMOS, P. L.; RODRIGUES, F. A.; LOUZADA, F. Objective bayesian analysis for the differential entropy of the gamma distribution. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 38, n. 1, p. 53–73, 2024. Citation on page 28.

RAMOS, E. M. L. S. **Estatística: poderosa ciência ao alcance de todos**. Belém - PA: [s.n.], 2016. 2 p. Jornal Beira do Rio. Available: <http://www.jornalbeiradorio.ufpa.br/novo/index.php/2004/61-edicao-21/691-opiniao-estatistica-poderosa-ciencia-ao-alcance-de-todos->. Accessed: 13 jul. 2016. Citation on page 19.

RAMOS, P. L.; DEY, D. K.; LOUZADA, F.; RAMOS, E. On posterior properties of the two parameter gamma family of distributions. **Anais da Academia Brasileira de Ciências**, SciELO Brasil, v. 93, p. e20190826, 2021. Citation on page 28.

RAMOS, P. L.; NASCIMENTO, D. C.; FERREIRA, P. H.; WEBER, K. T.; SANTOS, T. E.; LOUZADA, F. Modeling traumatic brain injury lifetime data: Improved estimators for the generalized gamma distribution under small samples. **PLoS one**, Public Library of Science San Francisco, CA USA, v. 14, n. 8, p. e0221332, 2019. Citation on page 28.

RAUDENBUSH, S. W.; BRYK, A. S. **Hierarchical linear models: Applications and data analysis methods**. [S.l.]: sage, 2002. Citation on page 22.

REINKING, O. A. *et al.* Diseases of economic plants in southern china. **Philippine Agriculturist**, Los Baños., v. 8, n. 4, 1919. Citation on page 31.

RIBEIRO, G. D. **Algumas Espécies de Plantas Reunidas por Familias e suas Propriedades**. Porto Velho, RO: Embrapa Rondônia, 2010. 179 p. Citation on page 35.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized Additive Models for Location, Scale and Shape (with discussion). **Applied Statistics**, v. 54, n. 3, p. 507–554, 2005. Citations on pages 62 and 81.

ROBERT, C. P. **The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation**. 2nd. ed. [S.l.]: Springer, 2007. Citation on page 60.

ROCHA, A. V.; SIMAS, A. B. Influence Diagnostics in a General Class of Beta Regression Models. **Test**, v. 20, n. 1, p. 95–119, 2011. Citations on pages 79 and 80.

RStudio Team. **RStudio: Integrated Development Environment for R**. Boston, MA, 2015. Available: <http://www.rstudio.com/>. Citation on page 83.

SALSBURG, D. S. **Uma senhora toma chá... como a estatística revolucionou a ciência no século XX**. [S.l.]: Zahar Rio de Janeiro, 2009. Citation on page 19.

SANGIN, E.; PATIL, P. R.; MISHRA, S.; SEN, S. Proficiency of probability distributions in unit hydrograph derivation. **Hydrology Research**, IWA Publishing, p. 1–21, 2024. Citations on pages 28 and 30.

SANTOS, D. P. dos; SERMARINI, R. A.; SANTOS, A. dos; DEMÉTRIO, C. G. B. Optimal designs in plant breeding experiments: A simulation study comparing grid-plot and partially replicated (p-rep) design. **Sugar Tech**, Springer, v. 26, n. 2, p. 387–395, 2024. Citation on page 19.

SARAIVA, E. F.; LOUZADA, F.; MILAN, L. A.; MEIRA, S.; COBRE, J. A bayesian approach for decision making on the identification of genes with different expression levels: An application to escherichia coli bacterium data. **Computational and Mathematical Methods in Medicine**, Wiley Online Library, v. 2012, n. 1, p. 953086, 2012. Citation on page 19.

SCAPIM, C. A.; CARVALHO, C. G. P. de; CRUZ, C. D. Uma proposta de classificação dos coeficientes de variação para a cultura do milho. **Pesquisa agropecuária brasileira**, v. 30, n. 5, p. 683–686, 1995. Citation on page 20.

SCHAAD, N. W.; POSTNIKOVA, E.; LACY, G.; SECHLER, A.; AGARKOVA i.; STROMBERG, V. K.; VIDAVER, A. K. Emended classification of xanthomonad pathogens on citrus. **Systematic and Applied Microbiology**, v. 29, p. 690–695, 2006. Citation on page 34.

SEMINA: Ciências Agrárias. Universidade Estadual de Londrina (UEL), 2024. Acesso em: 25 nov. 2024. Available: <https://ojs.uel.br/revistas/uel/index.php/semagrarias>. Citation on page 21.

SENRA, N. de C. As instituições estatísticas como centros de ciência, uma (r) evolução necessária. **Estatística e Sociedade**, n. 1, 2011. Citation on page 19.

SERMARINI, R. A.; BRIEN, C.; DEMÉTRIO, C. G. B.; SANTOS, A. dos. Impact on genetic gain from using misspecified statistical models in generating p-rep designs for early generation plant-breeding experiments. **Crop Science**, Wiley Online Library, v. 60, n. 6, p. 3083–3095, 2020. Citation on page 19.

SHARMIN, A. A.; ZULKAFLI, H. S.; ALI, N. M. Establishing cut-off points for consistency in reporting hypoglycemia symptoms among diabetes patients. **JP Journal of Biostatistics**, v. 24, n. 1, p. 31–46, 2024. Citation on page 24.

SHKEDY, Z.; MOLENBERGHS, G.; CRAENENDONCK, H. V.; STECKLER, T.; BIJNENS, L. A hierarchical binomial-poisson model for the analysis of a crossover design for correlated binary data when the number of trials is dose-dependent. **Journal of biopharmaceutical statistics**, Taylor & Francis, v. 15, n. 2, p. 225–239, 2005. Citations on pages 22, 23, and 24.

SHKEDY, Z.; VANDERSMISSEN, V.; MOLENBERGHS, G.; CRAENENDONCK, H. V.; AERTS, N.; STECKLER, T.; BIJNENS, L. Behavioral testing of antidepressant compounds: an analysis of crossover design for correlated binary data. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, Wiley Online Library, v. 47, n. 3, p. 286–298, 2004. Citation on page 24.

____. Behavioral testing of antidepressant compounds: an analysis of crossover design for correlated binary data. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, Wiley Online Library, v. 47, n. 3, p. 286–298, 2005. Citation on page 22.

SMITH, D. Algorithm as 189: Maximum likelihood estimation of the parameters of the beta binomial distribution. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, JSTOR, v. 32, n. 2, p. 196–204, 1983. Citation on page 28.

SMITHSON, M.; VERKUILEN, J. A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. **Psychological Methods**, American Psychological Association, v. 11, n. 1, p. 54–71, 2006. Citation on page 77.

SNEE, R. D. Statistics in industry. **Encyclopedia of Statistical Sciences S. Kotz and NL John Editors**, v. 4, p. 69–73, 1983. Citation on page 19.

SØBJERG, L. M.; TAYLOR, B. J.; PRZEPERSKI, J.; HORVAT, S.; NOUMAN, H.; HARVEY, D. Using risk factor statistics in decision-making: prospects and challenges. **European Journal of Social Work**, Taylor & Francis, v. 24, n. 5, p. 788–801, 2021. Citation on page 19.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. van der. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley, v. 64, n. 4, p. 583–639, oct 2002. Citations on pages 64 and 82.

Stan Development Team. **RStan: the R interface to Stan**. 2024. R package version 2.32.6. Available: <https://mc-stan.org/>. Citation on page 65.

STEPHANES, R. **Instrução Normativa MAPA - Nº 53, de 16/10/2008**. 2008. Acessado em: 30/06/2022. Available: <https://www.defesa.agricultura.sp.gov.br/legislacoes/instrucao-normativa-mapa-53-de-16-10-2008,830.html>. Citation on page 32.

SUNWASIYA, D. K.; CHANDOLIA, L. K.; UTTAM, V. Evaluating genetic variability in goats considering von bertalanffy growth curve parameters. **International Journal of Advanced Biochemistry Research**, v. 8, n. 1, p. 334–336, 2024. Citation on page 28.

TAMURA, R. N.; YOUNG, S. S. A stabilized moment estimator for the beta-binomial distribution. **Biometrics**, [Wiley, International Biometric Society], v. 43, n. 4, p. 813–824, 1987. ISSN 0006341X, 15410420. Available: <http://www.jstor.org/stable/2531535>. Citation on page 27.

TANG, B.; FRYE, H. A.; GELFAND, A. E.; SILANDER, J. A. Zero-inflated beta distribution regression modeling. **Journal of Agricultural, Biological and Environmental Statistics**, Springer, v. 28, n. 1, p. 117–137, 2023. Citation on page 80.

TAUFINA, T.; CHANDRA, C.; FAUZAN, A.; SYARIF, M. I. Development of statistics in elementary school based rme approach with problem solving for revolution industry 4.0. In: ATLANTIS PRESS. **5th International Conference on Education and Technology (ICET 2019)**. [S.l.], 2019. p. 716–721. Citation on page 19.

TEIXEIRA, D.; SAILLARD, C.; EVEILLARD, S.; DANET, J.; AYRES, A.; BOVÉ, J. A new liberibacter species, candidatus liberibacter americanus sp. nov., is associated with citrus huanglongbing (greening disease) in são paulo state, brazil. In: **International Organization of Citrus Virologists Conference Proceedings (1957-2010)**. [S.l.: s.n.], 2005. v. 16, n. 16. Citation on page 31.

TEXEIRA, D.; AYRES, J.; KITAJIMA, E.; DANET, L.; JAGOUEIX-EVEILLARD, S.; SAILLARD, C.; BOVÉ, J. First report of a huanglongbing-like disease of citrus in são paulo state, brazil and association of a new liberibacter species,"candidatus liberibacter americanus", with the disease. **Plant disease**, Am Phytopath Society, v. 89, n. 1, p. 107–107, 2005. Citation on page 31.

TJøRVE, E. Shapes and functions of species-area curves: A review of possible models. **Journal of Biogeography**, v. 30, p. 827 – 835, 06 2003. Citation on page 30.

TOMAZELLA, V.; LOUZADA-NETO, F.; SILVA, G. Bayesian modeling of recurrent events data with an additive gamma frailty distribution and a homogeneous poisson process. **Journal of Statistical Theory and Applications**, v. 5, n. 4, p. 417–429, 2006. Citation on page 28.

TRIPATHI, R. C.; GUPTA, R. C.; GURLAND, J. Estimation of parameters in the beta binomial model. **Annals of the Institute of Statistical Mathematics**, Springer, v. 46, p. 317–331, 1994. Citation on page 27.

TROYO, A.; FULLER, D. O.; CALDERÓN-ARGUEDAS, O.; BEIER, J. C. A geographical sampling method for surveys of mosquito larvae in an urban area using high-resolution satellite imagery. **Journal of vector ecology: journal of the Society for Vector Ecology**, NIH Public Access, v. 33, n. 1, p. 1, 2008. Citation on page 54.

TSOULARIS, A. Analysis of logistic growth models. **Res. Lett. Inf. Math. Sci**, v. 2, p. 23–46, 2001. Disponível online em <http://www.massey.ac.nz/wwiims/~rlims>. Citation on page 30.

TUFFÉRY, S. **Data mining and statistics for decision making**. [S.l.]: John Wiley & Sons, 2011. Citation on page 19.

UDOUMOH, E. F.; EBONG, D. W.; IWOK, I. A. Simulation of project completion time with burr xii activity distribution. **Asian Res J Math**, v. 6, n. 4, p. 1–14, 2017. Citation on page 30.

VERHULST, P. F. Notice sur la loi que la population poursuit dans son accroissement. In: GARNIER, J. G.; QUETELET, A. (Ed.). **Correspondance Mathémathique et Physique**. [S.l.]: Observatoire Royal de Belgique, 1838. p. 113–121. Citation on page 63.

VERONESI, J. A.; CRUZ, C. D.; CORRÊA, L. A.; SCAPIM, C. A. Comparação de métodos de ajuste do rendimento de parcelas com estandes variados. **Pesquisa Agropecuária Brasileira**, v. 30, n. 2, p. 169–174, 1995. Citation on page 20.

VIAGEM, A. **grupo da Cocamar viu que a Flórida perdeu a luta contra o greening**. [S.l.]: Ilustrada, 2014. Citation on page 31.

VIDAL, M. de F. **Produção de laranja na área de atuação do BNB**. 2021. Caderno Setorial ETENE, Ano 6, Nº 198. Pesquisa realizada em 20/12/2023. Disponível em: <https://www.bnb.gov.br/s482-dspace/bitstream/123456789/1041/1/2021_CDS_198.pdf>. Available: <https://www.bnb.gov.br/s482-dspace/bitstream/123456789/1041/1/2021_CDS_198.pdf>. Citation on page 35.

VILA, R.; ALFAIA, L.; MENEZES, A. F.; ÇANKAYA, M. N.; BOURGUIGNON, M. A model for bimodal rates and proportions. **Journal of Applied Statistics**, Taylor & Francis, v. 51, n. 4, p. 664–681, 2024. Citations on pages 25 and 95.

VILORIA, Z.; DROUILLARD, D. L.; GRAHAM., J. H.; GROSSER, J. W. Screening triploid hybrids of lakeland limequat for resistance to citrus canker. **Plant Disease**, v. 88, p. 1056–1060, 2004. Citations on pages 34 and 35.

WAIZ, H. A.; GAUTAM, L.; WAIZ, S. A. Appraisal of growth curve in sirohi goat using non-linear growth curve models. **Tropical animal health and production**, Springer, v. 51, p. 1135–1140, 2019. Citation on page 28.

WEIBULL, W. A statistical distribution function of wide applicability. **Journal of applied mechanics**, v. 18, p. 293–297, september 1951. Citation on page 63.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Available: <http://ggplot2.org>. Citation on page 83.

XIA, Y.; SUN, J. **Bioinformatic and statistical analysis of microbiome data: from raw sequences to advanced modeling with QIIME 2 and R**. [S.l.]: Springer Nature, 2023. Citation on page 80.

YAMAMOTO, E.; YANAGIMOTO, T. Moment estimators for the beta-binomial distribution. **Journal of applied statistics**, Taylor & Francis, v. 19, n. 2, p. 273–283, 1992. Citation on page 27.

YANG, R. C.; KOZAK, A.; SMITH, J. H. G. The potential of weibull-type functions as flexible growth curves. **Canadian Journal of Forest Research**, NRC Research Press Ottawa, Canada, v. 8, n. 4, p. 424–431, 1978. Citation on page 30.

ZEIDE, B. Analysis of growth equations. **Forest science**, Oxford University Press, v. 39, n. 3, p. 594–616, 1993. Citation on page 30.

ZHAO, W.; ZHANG, R.; LV, Y.; LIU, J. Variable Selection for Varying Dispersion Beta Regression Model. **Journal of Applied Statistics**, Informa UK Limited, v. 41, n. 1, p. 95–108, ago 2014. Citation on page 80.

ZHOU, G.; LIN, Z. Improved beta-binomial estimation for reliability of healthcare quality measures. **medRxiv**, Cold Spring Harbor Laboratory Press, p. 2023–01, 2023. Citation on page 26.

ZHU, J.; EICKHOFF, J. C.; KAISER, M. S. Modeling the dependence between number of trials and success probability in beta-binomial–poisson mixture distributions. **Biometrics**, Oxford University Press, v. 59, n. 4, p. 955–961, 2003. Citation on page 23.


ICMC
USP
SÃO CARLOS