



## Zero-inflated beta regression model for leaf citrus canker incidence in orange genotypes grafted onto different rootstocks

Oilson Alberto Gonzatto Junior<sup>1\*</sup>, Terezinha Aparecida Guedes<sup>2</sup>, Aline Maria Orbolato Gonçalves-Zuliani<sup>4</sup> and William Mario de Carvalho Nunes<sup>3</sup>

<sup>1</sup>Programa de Pós-graduação em Bioestatística, Universidade Estadual de Maringá, Avenida Colombo, 5790, 87020-900, Maringá, Paraná, Brazil.

<sup>2</sup>Departamento de Estatística, Universidade Estadual de Maringá, Maringá, Paraná, Brazil. <sup>3</sup>Departamento de Agronomia, Universidade Estadual de Maringá, Maringá, Paraná, Brazil. <sup>4</sup>Centro Universitário de Maringá, Maringá, Paraná, Brazil. \*Author for correspondence.  
E-mail: oilson.agjr@gmail.com

**ABSTRACT.** Data with excess zeros are frequently found in practice, and the recommended analysis is to use models that adequately address the counting of zero observations. In this study, the Zero Inflated Beta Regression Model (BeZI) was used on experimental data to describe the mean incidence of leaf citrus canker in orange groves under the influence of genotype and rootstocks of origin. Based on the model, it was possible to quantify the odds that a null observation to mean incidence comes from a particular plant according to genotype and rootstock, and estimate its expected value according to this combination. *Laranja Caipira* rootstock proved to be the most resistant to leaf citrus canker as well as *Limão Cravo* proved to be the most fragile. The *Ipigúá IAC*, *Arapongas*, *EEL* and *Olímpia* genotypes have statistically equivalent chances.

**Keywords:** GAMLSS models, inflated models, odds ratio, *Citrus sinensis*, *Xanthomonas citri* subsp. *citri*.

## Modelo de regressão beta inflacionado de zeros para a incidência de cancro cítrico foliar em genótipos de laranja enxertados sobre diferentes porta-enxertos

**RESUMO.** Dados com excesso de zeros são encontrados muitas vezes na prática, e a análise recomendada é utilizar modelos que suportem adequadamente a contagem de observações nulas. Neste artigo, o Modelo de Regressão Beta Inflacionado de Zeros (BeZI) foi aplicado a dados experimentais para descrever a incidência média de cancro cítrico foliar em pomares de laranja sob a influência do genótipo e do porta-enxerto de origem. Com base no modelo, foi possível quantificar as chances de que uma observação nula para a incidência média seja proveniente de uma determinada planta, de acordo com o genótipo e o porta-enxerto, além de estimar o seu valor esperado conforme essa combinação. O porta-enxerto Laranja Caipira mostrou ser o mais resistente ao cancro cítrico foliar, assim como o Limão Cravo mostrou ser o mais suscetível. Os genótipos Ipigúá IAC, Arapongas, EEL e Olímpia apresentaram chances estatisticamente equivalentes.

**Palavras-chave:** Modelos GAMLSS, modelos inflacionados, razão de chances, *Citrus sinensis*, *Xanthomonas citri* subsp. *citri*.

### Introduction

#### The agronomic framework

Similar to other agricultural crops, the orange cultivation demands management and practices. Currently, the various diseases that affect the orange groves include citrus canker, caused by *Xanthomonas citri* subsp. *citri*. (Gonçalves-Zuliani et al., 2016). This disease does not have control, the recommendation is the elimination of infected plants, however, some studies have pointed possible means of control, indicating that in the future the disease will be controlled (Sauer et al., 2015). Current techniques to prevent citrus canker in orange groves suggest the use of grafting onto rootstocks resistant to disease (Reis et al., 2008).

Rootstock exerts significant influence on various cultural aspects of canopy variety, it is responsible for changes in growth, size, production precocity, productivity, time of ripening, fruit weight, coloring of peel, sugar and acid content, post-harvest conservation, drought and cold resistance, as well as resistance to pests and diseases (Leite Jr & Santos, 1988, Schafer, Bastianel, & Dornelles, 2001, Stuchi, Donadio, & Sempionato, 2002, Tazima, Neves, Yada, & Leite Jr, 2010).

Choosing a suitable rootstock should take many factors into account, and the possibility of establishing a resistance criterion to citrus canker for canopy variety is an important feature in the integrated management or prevention of disease,

especially in regions where citrus canker is endemic (Danos, Berger, & Stall, 1984, Agostini, Graham, & Timmer, 1985).

In the state of Paraná, there are few studies on the impact of citrus canker on specific genotypes under the influence of a rootstock. This study evaluated the resistance of Laranja Doce (*Citrus sinensis* (L.) Osbeck) var. Pera genotypes grafted onto different rootstocks to the action of the bacterium *Xanthomonas citri* subsp. *citri*. The modeling was performed using the Zero Inflated Beta Regression model.

### The statistical framework

Beta distribution is commonly used to understand the variability of a random variable  $Y$  supported on an open interval  $I = (a, b) \subset \mathbb{R}$ , with  $a, b \in \mathbb{R}$  and  $a < b$ . If, in particular,  $Y$  denotes a proportion, then  $a = 0$  and  $b = 1$ , i.e.  $U = (0, 1) \subset \mathbb{R}$ . A statistical approach that examines the existence and quantify any possible associations between  $Y$ , and the presence of a set of factors  $X$  that exert influence on their behavior, is the Beta Regression Model.

With extensive application, Beta Regression Models have been used to understand specific aspects of a multitude of phenomena, such as scores in tests conducted on dyslexics and non-dyslexics individuals (Smithson & Verkuilen, 2006), the canopy cover rate of trees in forests according to the basal area, tree height, local fertility, and others factors (Korhonen, Korhonen, Stenberg, Maltamo, & Rautiainen, 2007), the selective collection rate in some cities, considering socioeconomic, demographic factors and others (Ibáñez, Prades, & Simó, 2011), the relationship between trade and business cycles (Mendonça, Silvestre, & Passos, 2011), and many others areas such as health (Moraes, Rocha, & Machado, 2012) or meteorology and climatology (Mullen, Marshall, & McGlynn, 2013).

A severe restriction on the use of Beta distribution is its support, an open interval. This difficulty can be easily overcome with the use of a Two-Part Model. Although the model known as Zero Inflated Beta Model is relatively recent, the use of Inflated Models is well addressed in the literature. This research exposes the Zero Inflated Beta Model, and uses this theory to model a real agronomy database.

### Material and methods

#### Leaf citrus canker incidence

This study came from an experiment that used as plant material nine Laranja Doce var. Pera

genotypes. Each genotype: Arapongas, Bianchi, EEL, IAC, IAC 2000, Ipiruá IAC, N58, N59 and Olímpia, was grafted onto four rootstocks: Laranja Caipira (*Citrus sinensis* (L.) Osbeck), Limão Cravo (*Citrus limonia* (L.) Osbeck), Tangerina Cleópatra (*Citrus reshni* hort. ex Tanaka) and Tangerina Sunki (*Citrus sunki* (Hayata) hort. ex Tanaka), totaling 36 clones.

The sampling process consisted of quarterly evaluation of ten plants per clone in periods of 2010, 2011 and 2012. When the plants were about two years old, quarterly evaluations began. Four branches per plant were randomly sampled for technical diagnostics. All cultures received normal management for both citrus canker and other diseases, throughout the experiment.

To obtain estimates for leaf citrus canker incidence, it was counted the number of infected leaves over total number of leaves. This process was repeated in five evaluations and, at the end of the experiment, the average of the observed incidences was calculated, resulting in 360 sampling leaf citrus canker mean incidences.

#### BeZI ( $v, \mu, \sigma$ ) distribution

The Beta distribution is a continuous probability distribution usually defined on the open unit interval  $U = (0, 1) \subset \mathbb{R}$ , parameterized by two parameters,  $p > 0$  and  $q > 0$  and, in this case, denoted by  $Be(p, q)$ . The probability density function of  $Be(p, q)$  is the real function in Equation 1:

$$f_{Be}(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1} \quad (1)$$

$$0 < y < 1,$$

where:

$p > 0, q > 0$  are shape parameters and the symbol  $\Gamma$  denotes the Gamma function.

Beta distribution is a very versatile probability distribution family, may take a wide variety of forms such constant format (uniform distribution, when  $p = q = 1$ ), symmetrical unimodal or bathtub (when  $p = q \neq 1$ ), negative asymmetrical unimodal or strictly increasing (when  $p > q$ ), positive asymmetric unimodal or strictly decreasing (when  $p < q$ ).

The expected value and variance of  $Be(p, q)$  distribution are expressed, respectively, by Equation 2:

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

In regression analysis is common, and reasonably convenient, to model the expected value of the response variable, that is, when possible, is it more interesting than the adjustment made on a parameter that represents the mean of random variable. In the case of  $\text{Be}(p, q)$ , note that the expected value of  $Y$  is a function of  $p$  and  $q$  parameters.

The idea is to work with a new parameterization of the Beta distribution, for this, consider  $\mu = p/(p+q)$ , so,  $0 < \mu < 1$  and  $\sigma = p+q$ , so,  $\sigma > 0$ , so that the reverse substitutions are uniquely determined by  $p = \sigma\mu$ , so,  $p > 0$  and  $q = \sigma(1-\mu)$ , so  $q > 0$ . Considering the new parameterization, the probability density function of  $\text{Be}(p, q)$  distribution will be denoted by  $\text{Be}(\mu, \sigma)$  and it is described as Equation 3:

$$f_{\text{Be}}(y; \mu, \sigma) = \frac{\Gamma[\sigma]}{\Gamma[\sigma\mu] \Gamma[\sigma(1-\mu)]} y^{\sigma\mu-1} (1-y)^{\sigma(1-\mu)-1}, \quad (3)$$

$$0 < y < 1,$$

where:

$\mu \in (0,1)$  is a position parameter,  $\sigma > 0$  is a scale parameter and  $\Gamma$  denotes the Gamma function.

Since the support of a random variable with Beta distribution is the unit interval  $U$ , it is a limitation for using this distribution in practical applications that have zero observations. One way to overcome this difficulty is to use Mixture Models theory (specifically, a Two-Part Model) and compose a new distribution, here denoted by  $\text{BeZI}(\nu, \mu, \sigma)$ , based on two distinct distributions: one to model the zeros observations and another to model the observations belonging to the  $U$  interval.

The  $\text{BeZI}(\nu, \mu, \sigma)$  probability density function, is expressed by Equation 4:

$$f_{\text{BeZI}}(y; \nu, \mu, \sigma) = [\nu]^{I_{[0]}(y)} [(1-\nu)f_{\text{Be}}(y; \mu, \sigma)]^{1-I_{[0]}(y)} \quad (4)$$

$$0 \leq y < 1,$$

where:

$\mu \in (0,1)$  is a position parameter,  $\sigma > 0$  is a scale parameter and  $\nu \in [0,1]$  is a shape parameter. The indicator function  $I_{[0]}(y)$  is a function such that  $I_{[0]}(y) = 1$ , if  $y = 0$  and  $I_{[0]}(y) = 0$ , if  $y \neq 0$ . The mean and variance of  $\text{BeZI}(\nu, \mu, \sigma)$  are expressed, respectively, by Equation 5:

$$E(Y) = (1-\nu)\mu \text{ and } \text{Var}(Y) = (1-\nu) \left[ \frac{\mu(1-\mu)}{\sigma+1} + \nu\mu^2 \right]. \quad (5)$$

### BeZI ( $\nu, \mu, \sigma$ ) regression model

The idea of modeling the expected value of  $\text{Be}(p, q)$  distribution was already under discussion for some time in the works of Jorgensen (1997), Paolino (2001) and Kieschnick and McCullough (2003), e.g., however, the regression model exposed by Ferrari and Cribari-Neto (2004) became popular for formulating more carefully the modeling of the expected value in  $\text{Be}(p, q)$  distribution, based on  $\text{Be}(\mu, \sigma)$  distribution parameterization, and to establish an association with GLM theory, a class well described in the literature by Nelder and Wedderburn (1972). These, and others works gave rise to a number of new studies and approaches, such as (Cepeda-Cuervo, 2015).

To make a more appropriate modeling in the context of small samples, the likelihood inference is discussed by Ferrari and Pinheiro (2011), whose proposed adjusted statistics were more reliable in simulation. A discussion about inference in a Beta Regression Model with non-constant precision, also in small samples cases, can be found in Cribari-Neto and Queiroz (2012).

The diagnostic analysis in Beta Regression Models was again discussed by Rocha and Simas (2011), Ferrari, Espinheira, and Cribari-Neto (2011), Chien (2011) and Anholetto, Sandoval, and Botter (2014), with some new influence measures, graphical tools and a new residual type for diagnosis. Variable selection methodology, in non-constant dispersion modeling, was proposed by Zhao, Zhang, Lv, and Liu (2014). Furthermore, Latif and Yab (2015) expose a way to determine the optimum design for an experiment with regression model  $\text{Be}(\mu, \sigma)$  involving a single predictor variable.

The  $\text{BeZI}(\nu, \mu, \sigma)$  regression model received a formal treatment in Ospina and Ferrari (2012). Because it is a more recent study, complementary theory that naturally comes is not yet so affordable, even though the whole development acquired for  $\text{Be}(\mu, \sigma)$  regression model is closely associated with it. One can find in the literature a new Likelihood Ratio Test for Beta Inflated Models (Pereira & Cribari-Neto, 2014b). A discussion of error detection in Beta Inflated modeling (Pereira & Cribari-Neto, 2014a).

### Model definition

Consider a set of independent and identically distributed random observations  $Y = (Y_1, \dots, Y_n)^T$ ,

where each component  $Y_i$ , with  $i = 1, \dots, n$ , such  $Y_i \sim \text{BeZI}(\nu_i, \mu_i, \sigma_i)$ . Keep in mind the formulation of a Generalized Additive Model for Location, Scale and Shape (GAMLSS) exposed in Rigby and Stasinopoulos (2005), each parameter  $\nu_i$ ,  $\mu_i$  and  $\sigma_i$  of  $\text{BeZI}(\nu_i, \mu_i, \sigma_i)$  distribution can be associated with the random variable  $Y_i$  by a link function. The Zero Inflated Beta Regression Model can be defined by Equation 6:

$$Y_i \sim \text{BeZI}(\nu_i, \mu_i, \sigma_i), \quad (6)$$

where:

$i = 1, \dots, n$ . Furthermore, the parameters  $\nu_i$ ,  $\mu_i$  and  $\sigma_i$  satisfy the functional relationships, according Equation 7:

$$\begin{aligned} g_\nu(\nu_i) &= \eta_{iv} = x_{iv}^\top \beta_\nu, \\ g_\mu(\mu_i) &= \eta_{i\mu} = x_{i\mu}^\top \beta_\mu, \\ g_\sigma(\sigma_i) &= \eta_{i\sigma} = x_{i\sigma}^\top \beta_\sigma, \end{aligned} \quad (7)$$

where:

$\beta_\nu = (\beta_{1\nu}, \dots, \beta_{J'_\nu\nu})^\top$ ,  $\beta_\mu = (\beta_{1\mu}, \dots, \beta_{J'_\mu\mu})^\top$  and  $\beta_\sigma = (\beta_{1\sigma}, \dots, \beta_{J'_\sigma\sigma})^\top$  are parameters vectors with order  $J'_\nu$ ,  $J'_\mu$  and  $J'_\sigma$ , respectively, associated with the three distribution parameters and the vectors  $x_{iv} = (x_{iv1}, \dots, x_{ivJ'_\nu})^\top$ ,  $x_{i\mu} = (x_{i\mu1}, \dots, x_{i\mu J'_\mu})^\top$  and  $x_{i\sigma} = (x_{i\sigma1}, \dots, x_{i\sigma J'_\sigma})^\top$  are vectors with explanatory variables observations associated with  $\nu$ ,  $\mu$  and  $\sigma$  to  $i$ -th individual.

#### Expected value and variance

Considering the link functions  $g_\nu$ ,  $g_\mu$  and  $g_\sigma$  for distribution parameters  $\nu$ ,  $\mu$  and  $\sigma$ , respectively, the expected value and variance of  $Y_i$  are expressed by Equations 8 and 9:

$$\begin{aligned} E(Y_i) &= [1 - \nu_i] \mu_i = [1 - g_\nu^{-1}(\eta_{iv})] g_\mu^{-1}(\eta_{i\mu}) \\ &= [1 - g_\nu^{-1}(x_{iv}^\top \beta_\nu)] g_\mu^{-1}(x_{i\mu}^\top \beta_\mu), \end{aligned} \quad (8)$$

$$\begin{aligned} \text{Var}(Y_i) &= (1 - \nu_i) \left[ \frac{\mu_i(1 - \mu_i)}{\sigma_i + 1} + \nu_i \mu_i^2 \right] \\ &= [1 - g_\nu^{-1}(\eta_{iv})] \left\{ \frac{g_\mu^{-1}(\eta_{i\mu})[1 - g_\mu^{-1}(\eta_{i\mu})]}{g_\sigma^{-1}(\eta_{i\sigma}) + 1} + \right. \\ &\quad \left. g_\nu^{-1}(\eta_{iv}) [g_\mu^{-1}(\eta_{i\mu})]^2 \right\} \\ &= [1 - g_\nu^{-1}(x_{iv}^\top \beta_\nu)] \left\{ \frac{g_\mu^{-1}(x_{i\mu}^\top \beta_\mu)[1 - g_\mu^{-1}(x_{i\mu}^\top \beta_\mu)]}{g_\sigma^{-1}(x_{i\sigma}^\top \beta_\sigma) + 1} + \right. \\ &\quad \left. g_\nu^{-1}(x_{iv}^\top \beta_\nu) [g_\mu^{-1}(x_{i\mu}^\top \beta_\mu)]^2 \right\} \end{aligned} \quad (9)$$

The interval estimates for  $E(Y_i)$  and  $\text{Var}(Y_i)$ , are conditioned by obtaining the respective standard errors, which can be determined with Delta Method.

## Results and discussion

A  $\text{BeZI}(\nu, \mu, \sigma)$  regression model was adjusted to understand the behavior of the mean incidence of leaf citrus canker (denoted by  $Y$ ), influenced by two categorical covariates, the rootstock (denoted by PE and the categories *Laranja Caipira*, *Limão Cravo*, *Tangerina Cleópatra* and *Tangerina Sunki*) and another to represent the genotype (denoted by GE and the categories *Arapongas*, *Bianchi*, *ELL*, *IAC*, *IAC 2000*, *Ipiguá IAC*, *N58*, *N59* and *Olímpia*).

### Exploratory analysis

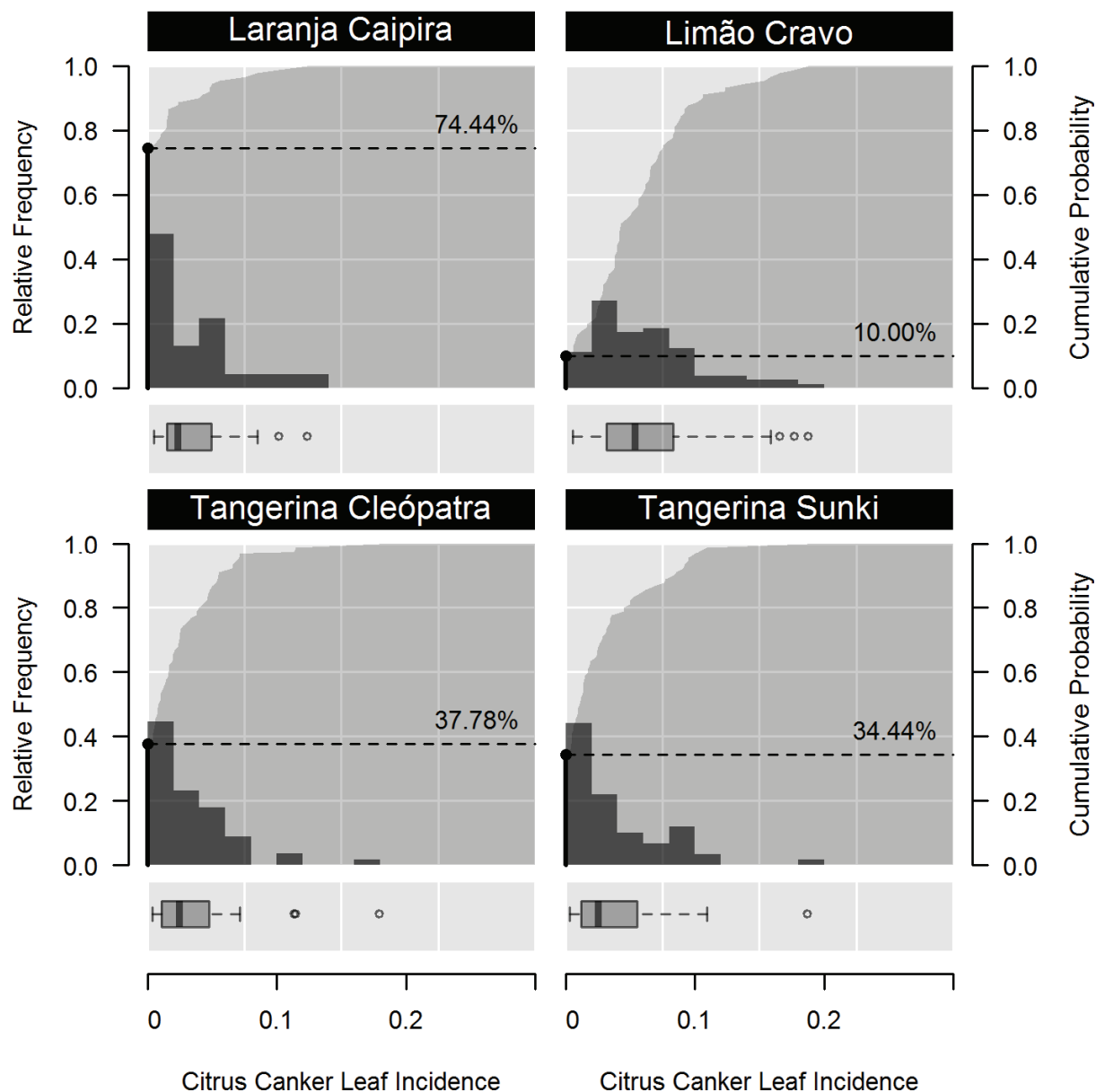
Figure 1 illustrates the histogram, box plot and empirical cumulative probability function, for the mean incidence considering each rootstock. The highest zero proportion corresponds to rootstock *Laranja Caipira* (74.44%) while the lowest corresponds to the rootstock *Limão Cravo* (10.00%). Considering the dispersion of observations, visually, the less dispersed one correspond to rootstock *Laranja Caipira*, followed by rootstock *Tangerina Cleópatra*, the two intermediaries in zero proportion observations.

Table 1 presents the summary measures for each rootstock, note that the lower mean value corresponds to *Laranja Caipira* rootstock (0.0095), also with the lowest standard deviation (0.0230); on the other hand, the highest mean corresponds to *Limão Cravo* rootstock (0.0544), also with the highest standard deviation (0.0428).

**Table 1.** Summary measures for the mean incidence of citrus canker leaf, analyzed per plant, taking into account the rootstock of origin.

| Rootstock                  | Min. Value | Max. Value | Mean Value | Standard Deviation | Coefficient of Variation |
|----------------------------|------------|------------|------------|--------------------|--------------------------|
| <i>Laranja Caipira</i>     | 0.0000     | 0.1236     | 0.0095     | 0.0230             | 2.4281                   |
| <i>Tangerina Cleópatra</i> | 0.0000     | 0.1794     | 0.0210     | 0.0303             | 1.4435                   |
| <i>Tangerina Sunki</i>     | 0.0000     | 0.1872     | 0.0252     | 0.0350             | 1.3900                   |
| <i>Limão Cravo</i>         | 0.0000     | 0.1876     | 0.0544     | 0.0428             | 0.7874                   |

In the genotypes cases, the summary for mean incidence of measures in each genotype is shown in Table 2, it is possible to observe that the lowest means correspond to *Ipiguá IAC* (0.0144) and *N59* (0.0187) genotypes, whose standard deviations are among the three lowest, while the highest means are observed for *EEL* (0.0354) and *Olímpia* (0.0331).



**Figure 1.** Histograms, box plots and empirical cumulative probability functions for the mean incidence of citrus canker leaf per plant analyzed taking into account the genotype of origin.

**Table 2.** Summary measures for the mean incidence of citrus canker leaf, analyzed per plant, taking into account the genotype of origin.

| Genotype          | Minimum Value | Maximum Value | Mean Value | Standard Deviation | Coefficient of Variation |
|-------------------|---------------|---------------|------------|--------------------|--------------------------|
| <i>Ipigua</i> LAC | 0.0000        | 0.1062        | 0.0144     | 0.0282             | 1.9567                   |
| <i>N59</i>        | 0.0000        | 0.0860        | 0.0187     | 0.0257             | 1.3697                   |
| <i>IAC</i>        | 0.0000        | 0.0918        | 0.0245     | 0.0249             | 1.0130                   |
| <i>Arapongas</i>  | 0.0000        | 0.1876        | 0.0283     | 0.0483             | 1.7054                   |
| <i>N58</i>        | 0.0000        | 0.1390        | 0.0303     | 0.0369             | 1.2151                   |
| <i>Bianchi</i>    | 0.0000        | 0.1536        | 0.0311     | 0.0367             | 1.1798                   |
| <i>IAC 2000</i>   | 0.0000        | 0.1136        | 0.0316     | 0.0305             | 0.9654                   |
| <i>Olimpia</i>    | 0.0000        | 0.1794        | 0.0331     | 0.0403             | 1.2204                   |
| <i>EEL</i>        | 0.0000        | 0.1872        | 0.0354     | 0.0523             | 1.4768                   |

Figure 2 shows the histogram, box plot and empirical cumulative probability function, for the

mean incidence considering each genotype. The highest zero proportion observations correspond to *Arapongas* and *Ipigua* LAC genotypes (60.00 and 57.50%) while the lowest correspond to *IAC* and *IAC 2000* genotypes (22.50% both), however, in general, genotypes present proportions similar to each other, i.e. the minimum and maximum are not so different from the others. Also note that, based on box plots, the set of observations visually less dispersed corresponds to *IAC* and *N59* genotypes, which although have low proportions of null observations, concentrate their observations very close to zero.

### Adjustment and considerations

Taken as reference the combination *Laranja Caipira* + *Ipiguá* IAC, the adopted model after the selection between candidate models has the Equation 10.

Where the response variable  $Y_i \in (0, 1)$ , its expected value  $\mu_i \in (0, 1)$ , the mixture parameter  $\nu_i \in [0, 1]$ , the precision parameter  $\sigma > 0$  and the identity  $\beta_{0\sigma} > 0$ , the vectors of parameters  $\beta_\nu \in \mathbb{R}^{12}$  and  $\beta_\mu \in \mathbb{R}^{12}$ . Furthermore, the categorical variable  $GE_j$ , with  $j = 1, \dots, 8$ , refers to eight genotypes compared to the reference combination and the categorical variable  $PE_j$ , with  $j = 1, 2, 3$ , refers to

three rootstocks compared to reference combination.

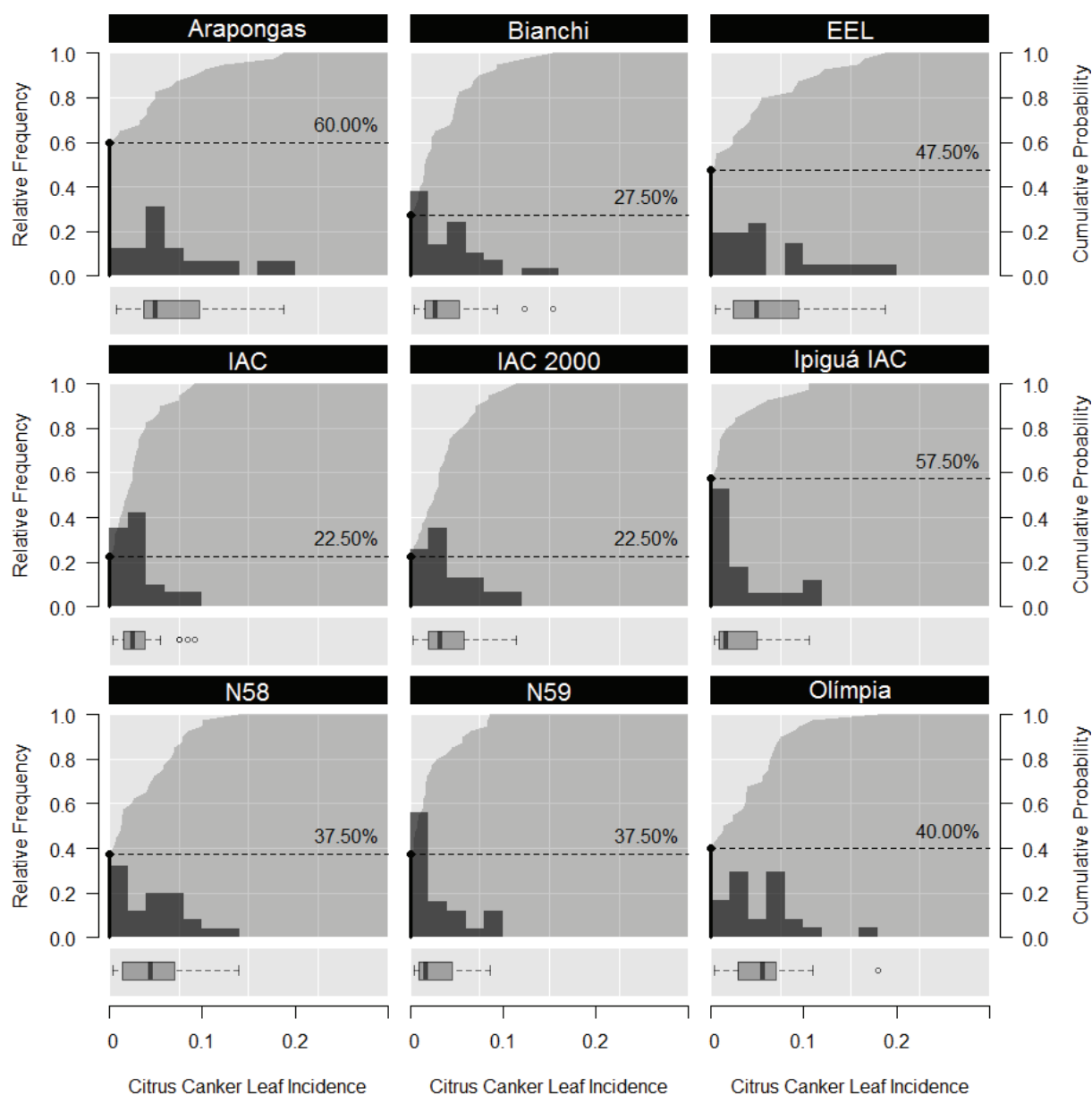
$$Y_i \sim \text{BeZI}(\nu_i, \mu_i, \sigma)$$

$$\log\left(\frac{\nu_i}{1 - \nu_i}\right) = \beta_{0\nu} + \sum_{j=1}^8 \beta_{j\nu} GE_{ij} + \sum_{j=1}^3 \beta_{(j+8)\nu} PE_{ij} \quad (10)$$

$$\mu_i = \beta_{0\mu} + \sum_{j=1}^8 \beta_{j\mu} GE_{ij} + \sum_{j=1}^3 \beta_{(j+8)\mu} PE_{ij}$$

$$\sigma = \beta_{0\sigma},$$

The results adjusted for  $\nu$  coefficients of BeZI ( $\nu, \mu, \sigma$ ) distribution can be observed in Table 3.



**Figure 2.** Histograms, box plots and empirical cumulative probability functions for the mean incidence of citrus canker leaf per plant analyzed taking into account the genotype of origin.

There is evidence that *Arapongas*, *EEL* and *Olímpia* genotypes influence on zero proportion for leaf citrus canker mean incidence seems to not differ statistically from that influence observed for reference combination (*Laranja Caipira* + *Ipigua IAC*). On the other hand, estimates indicate that influence exerted by the other three rootstocks (*Limão Cravo*, *Tangerina Cleópatra* and *Tangerina Sunki*) is statistically significant compared to reference combination.

In the case of  $\mu$  and  $\sigma$  coefficients, results are in Table 4. There are indications that the influence of *Arapongas*, *EEL* and *Olímpia* genotypes on the mean incidence of leaf citrus canker, in this case, seems statistically different from the reference combination. Furthermore, only the *Limão Cravo* rootstock differs statistically from reference combination.

**Table 3.** Point and interval estimates for  $\nu$  regression coefficients of BeZI ( $\nu, \mu, \sigma$ ) distribution.

|                            | $\beta_\nu$     | Estimative | CI (95%)             |
|----------------------------|-----------------|------------|----------------------|
| (Intercept)                | $\beta_{0\nu}$  | 2.24917    | (1.34704; 3.15129)   |
| <i>Arapongas</i>           | $\beta_{1\nu}$  | 0.13855    | (-0.89352; 1.17061)  |
| <i>Bianchi</i>             | $\beta_{2\nu}$  | -1.70849   | (-2.80083; -0.61615) |
| <i>EEL</i>                 | $\beta_{3\nu}$  | -0.54300   | (-1.56907; 0.48307)  |
| <i>IAC</i>                 | $\beta_{4\nu}$  | -2.05454   | (-3.18594; -0.92315) |
| <i>IAC 2000</i>            | $\beta_{5\nu}$  | -2.05454   | (-3.18594; -0.92315) |
| <i>N58</i>                 | $\beta_{6\nu}$  | -1.09784   | (-2.14350; -0.05218) |
| <i>N59</i>                 | $\beta_{7\nu}$  | -1.09784   | (-2.14350; -0.05218) |
| <i>Olímpia</i>             | $\beta_{8\nu}$  | -0.95581   | (-1.99441; 0.08280)  |
| <i>Limão Cravo</i>         | $\beta_{9\nu}$  | -3.63153   | (-4.53108; -2.73199) |
| <i>Tangerina Cleópatra</i> | $\beta_{10\nu}$ | -1.77199   | (-2.46022; -1.08376) |
| <i>Tangerina Sunki</i>     | $\beta_{11\nu}$ | -1.93578   | (-2.63218; -1.23938) |

**Table 4.** Point and interval estimates for  $\mu$  and  $\sigma$  regression coefficients of BeZI ( $\nu, \mu, \sigma$ ) distribution.

|                            | Coefficients $\mu$ |            |                     | Coefficients $\sigma$ |            |                    |
|----------------------------|--------------------|------------|---------------------|-----------------------|------------|--------------------|
|                            | $\beta_\mu$        | Estimative | CI (95%)            | $\beta_\sigma$        | Estimative | CI (95%)           |
| (Intercept)                | $\beta_0$          | 0.03057    | (0.01601; 0.04514)  | $\beta_{0\sigma}$     | 37.0943    | (37.0797; 37.1088) |
| <i>Arapongas</i>           | $\beta_{1\mu}$     | 0.02978    | (0.00865; 0.05092)  |                       |            |                    |
| <i>Bianchi</i>             | $\beta_{2\mu}$     | 0.01205    | (-0.00270; 0.02680) |                       |            |                    |
| <i>EEL</i>                 | $\beta_{3\mu}$     | 0.02351    | (0.00555; 0.04147)  |                       |            |                    |
| <i>IAC</i>                 | $\beta_{4\mu}$     | 0.00602    | (-0.00781; 0.01984) |                       |            |                    |
| <i>IAC 2000</i>            | $\beta_{5\mu}$     | 0.01170    | (-0.00282; 0.02623) |                       |            |                    |
| <i>N58</i>                 | $\beta_{6\mu}$     | 0.01459    | (-0.00130; 0.03048) |                       |            |                    |
| <i>N59</i>                 | $\beta_{7\mu}$     | 0.00124    | (-0.01276; 0.01524) |                       |            |                    |
| <i>Olímpia</i>             | $\beta_{8\mu}$     | 0.02182    | (0.00493; 0.03872)  |                       |            |                    |
| <i>Limão Cravo</i>         | $\beta_{9\mu}$     | 0.01516    | (0.00169; 0.02862)  |                       |            |                    |
| <i>Tangerina Cleópatra</i> | $\beta_{10\mu}$    | -0.00730   | (-0.01988; 0.00528) |                       |            |                    |
| <i>Tangerina Sunki</i>     | $\beta_{11\mu}$    | -0.00645   | (-0.01922; 0.00631) |                       |            |                    |

The model suitability evaluation, if positive, allows drawing conclusions based on the adjustment. This assessment should be performed by diagnostic criteria, such as graphical analysis of residuals from the optimization process.

The standardized residuals for discrete and continuous components are observed in Figure 3a and b, respectively. For discrete component, observations 11, 15, 21, 22, 66, 272, 273 and 274 are further away from the other observations. In continuous component, distance between observations is not so much expressive.

The withdrawal of observations one by one, and altogether, did not result insignificant differences for estimates or different interpretations of the obtained model in the presence of all observations, in this way we decided to keep all observations.

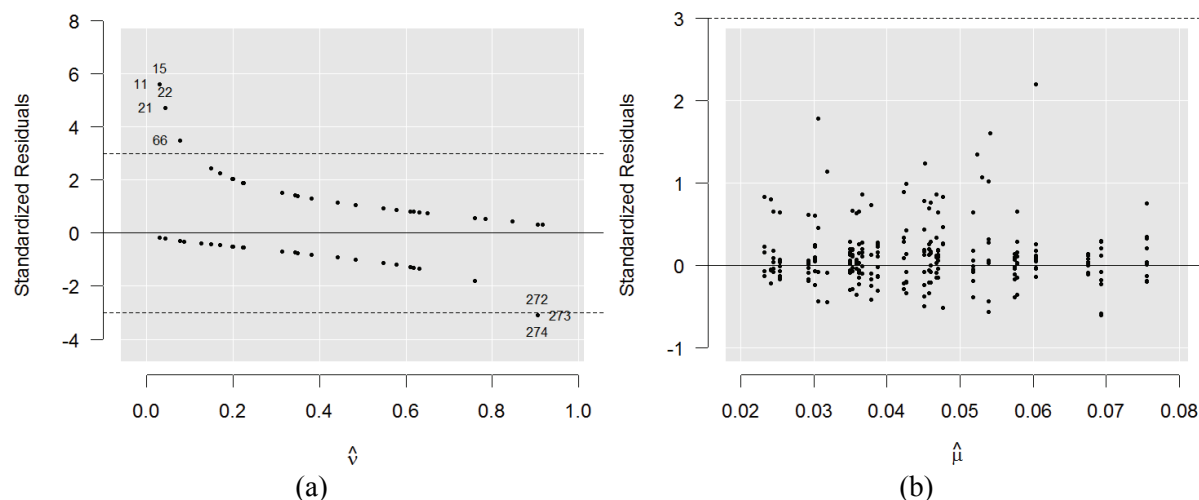
Figure 4 shows the simulated envelopes to assess the adequacy of the standardized (a), randomized (b) and weighted (c) residuals. In any graphics, very few observations fall outside the confidence bands, which strengthens the assumption that the Zero Inflated Beta Regression model accommodates the inherent variability of the response variable.

Finally, in Figure 5 are exposed worm plots built with a model individually for each component, discrete (a) and continuous (b). None of the observations lies beyond the confidence bands, indicating a good fit.

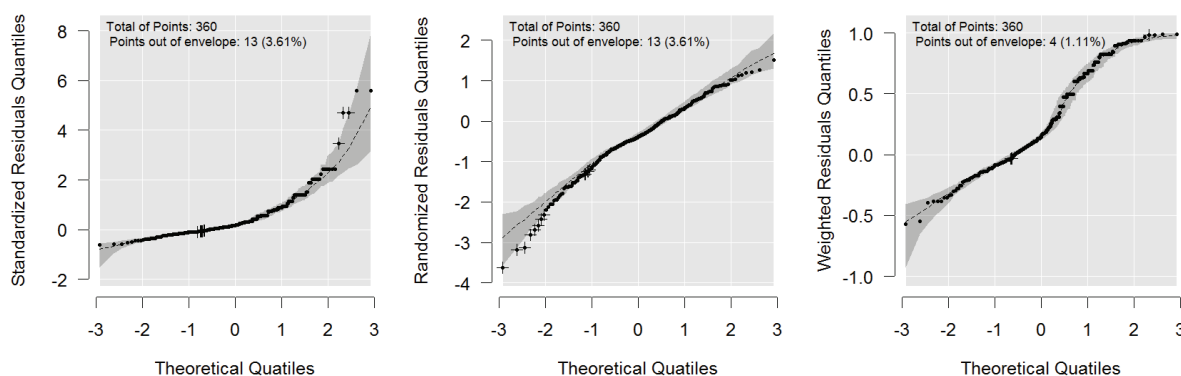
Based on graphical analysis, it is assumed that the model has the ability to briefly represent the observed reality in the experiment. A pseudo- $R$  equal to 0.49989, with a confidence interval (0.41813; 0.57361), allows us to understand that the adjustment can account for approximately 49.99% of the variability of the studied incidence.

Since the graphical analysis indicate good fit for the proposed model, it is possible to interpret the results and make conclusions based on modeling. Estimates obtained for  $\nu$  coefficients, shown in Table 3, allow to compute the odds compared to observe a null mean incidence of leaf citrus canker between each genotype or rootstock with the reference combination. The point and interval estimates can be found in Table 5.

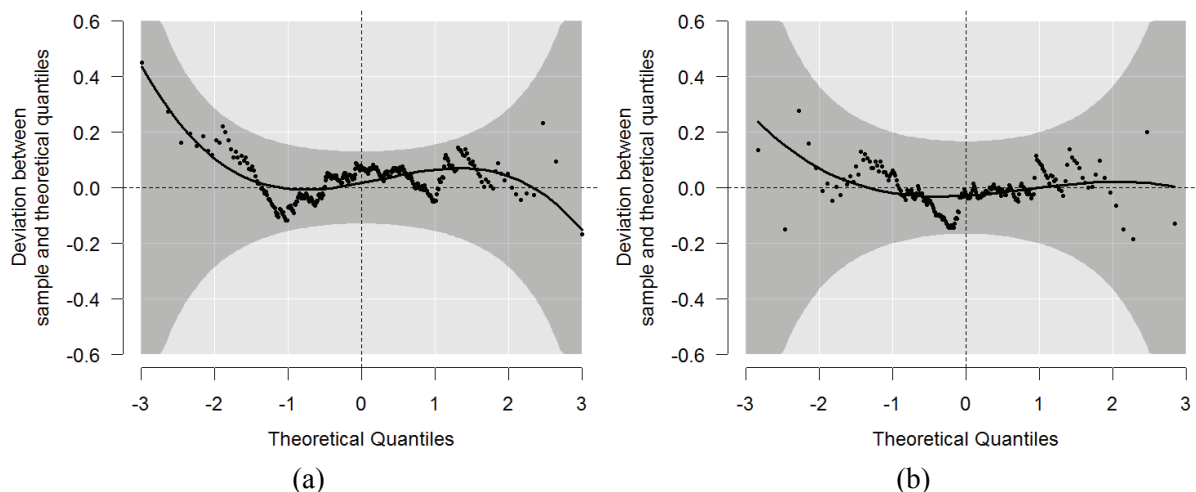
Considering a chance equals one for an observation of reference combination (and also of genotype and/or rootstock classified as statistically equivalent) be equals zero, Figure 6 illustrates a color matrix for estimated chance of a particular plant presents a null mean incidence of leaf citrus canker, compared to the reference combination genotype + rootstock.



**Figure 3.** Discrete component (a): Standardized residuals versus Adjusted Values for  $v$ . Continuous component (b): Standardized residuals versus Adjusted Values for  $\mu$ .



**Figure 4.** Simulated envelopes based on standardized (a), randomized (b) and weighted (c) residuals, resulting from the regression model BeZI ( $v, \mu, \sigma$ ) adjustment.



**Figure 5.** Worm plots built on the model set individually for the discrete (a) and continuous (b) components.

Clearly, *Laranja Caipira* rootstock presents a resistance to leaf citrus canker superior to the others. The chances to observe a null mean incidence with *Limão Cravo*, *Tangerina Cleópatra* and *Tangerina Sunki*

rootstocks are, respectively, 0.0265, 0.1700 and 0.1443 times the chance of *Laranja Caipira* rootstock. Based on these results, *Limão Cravo* rootstock has the highest susceptibility to leaf citrus canker.

**Table 5.** Odds ratios estimated for BeZI ( $v, \mu, \sigma$ ) model. Comparisons between the chance to present a zero mean incidence of different rootstocks and genotypes compared to the combination of rootstock *Laranja Caipira* and genotype *Ipigua IAC*.

|                            | OR      | CI (95%)           |
|----------------------------|---------|--------------------|
| <i>Arapongas</i>           | 1.14860 | (0.40921; 3.22397) |
| <i>Bianchi</i>             | 0.18114 | (0.06076; 0.54002) |
| <i>EEL</i>                 | 0.58100 | (0.20824; 1.62104) |
| <i>IAC</i>                 | 0.12815 | (0.04134; 0.39727) |
| <i>IAC 2000</i>            | 0.12815 | (0.04134; 0.39727) |
| <i>N58</i>                 | 0.33359 | (0.11724; 0.94916) |
| <i>N59</i>                 | 0.33359 | (0.11724; 0.94916) |
| <i>Olímpia</i>             | 0.38450 | (0.13609; 1.08632) |
| <i>Limão Cravo</i>         | 0.02648 | (0.01077; 0.06509) |
| <i>Tangerina Cleópatra</i> | 0.16999 | (0.08542; 0.33832) |
| <i>Tangerina Sunki</i>     | 0.14431 | (0.07192; 0.28956) |

Regarding genotypes and the resistance to leaf citrus canker, four genotypes are statistically equivalent, *Ipigua IAC* (reference) *Arapongas*, *EEL* and *Olímpia*. The chance that a null mean incidence is found in individuals of *N58* and *N59* genotypes is indistinguishable from each other (and equal to 0.3336 times the chance of *Ipigua IAC* genotype), as well as the *IAC* and *IAC 2000* (equivalent to 0.1282 times the chance of *Ipigua IAC* genotype). Moreover, the chance of *Bianchi* genotype presents a null mean incidence is equal to 0.1811 times the chance of *Ipigua IAC* genotype.

Since the observed mean incidence is different from zero, that is, considering that the plant is infected, the results that include this situation are

listed in Table 4 and based on them it can be concluded that:

- Once the plant is infected, *Bianchi*, *IAC*, *IAC 2000*, *N58* and *N59* genotypes are indistinguishable from reference combination, as well *Tangerina Cleópatra* and *Tangerina Sunki* rootstock;
- *Arapongas*, *EEL* and *Olímpia* genotypes are statistically different from reference combination, but, note that the contribution of the estimated mean incidence of leaf citrus canker of each is positive in any case, indicating that the estimated mean incidence is higher in these genotypes;
- Only *Limão Cravo* rootstock is statistically different from the reference combination, and also contributes to increase the mean incidence of leaf citrus canker.

In Figure 7, the color matrix is showed in accordance with each combination rootstock + genotype for mean incidence of leaf citrus canker estimated by the model. The combination *Limão Cravo* + *Arapongas* is more susceptible to the disease, with an expected mean incidence equal to 6.04%, followed by combinations *Limão Cravo* + *EEL* (5.54%) and *Limão Cravo* + *Olímpia* (5.40%). On the other hand, the reference combination had the highest resistance, with expected mean incidence equal to 0.29%, followed by *Laranja Caipira* + *Olímpia* (0.50%) and *Laranja Caipira* + *N58* or *N59* (both 0.73%).

|           | Ipigua IAC         | Arapongas | Bianchi | EEL    | IAC    | IAC 2000 | N58    | N59    | Olímpia |                     |
|-----------|--------------------|-----------|---------|--------|--------|----------|--------|--------|---------|---------------------|
| ROOTSTOCK | 0.1443             | 0.1443    | 0.0261  | 0.1443 | 0.0185 | 0.0185   | 0.0481 | 0.0481 | 0.1443  | Tangerina Sunki     |
|           | 0.1700             | 0.1700    | 0.0308  | 0.1700 | 0.0218 | 0.0218   | 0.0567 | 0.0567 | 0.1700  | Tangerina Cleopatra |
|           | 0.0265             | 0.0265    | 0.0048  | 0.0265 | 0.0034 | 0.0034   | 0.0088 | 0.0088 | 0.0265  | Limão Cravo         |
|           | 1.0000 (reference) | 1.0000    | 0.1811  | 1.0000 | 0.1282 | 0.1282   | 0.3336 | 0.3336 | 1.0000  | Laranja Caipira     |
| GENOTYPE  |                    |           |         |        |        |          |        |        |         |                     |

**Figure 6.** Color matrix for estimates odds ratios, based on the regression model for the occurrence of a null mean incidence according to genotype and rootstock.

|           | Ipiruá IAC        | Arapongas | Bianchi | EEL   | IAC   | IAC 2000 | N58   | N59   | Olímpia |                     |
|-----------|-------------------|-----------|---------|-------|-------|----------|-------|-------|---------|---------------------|
| ROOTSTOCK | 1.29%             | 2.55%     | 2.45%   | 2.28% | 2.60% | 2.60%    | 2.10% | 2.10% | 2.21%   | Tangerina Sunki     |
|           | 1.17%             | 2.31%     | 2.37%   | 2.07% | 2.53% | 2.53%    | 1.99% | 1.99% | 2.01%   | Tangerina Cleopatra |
|           | 3.66%             | 6.04%     | 4.37%   | 5.54% | 4.43% | 4.43%    | 4.22% | 4.22% | 5.40%   | Limão Cravo         |
|           | 0.29% (reference) | 0.58%     | 1.13%   | 0.52% | 1.38% | 1.38%    | 0.73% | 0.73% | 0.50%   | Laranja Calpira     |
| GENOTYPE  |                   |           |         |       |       |          |       |       |         |                     |

**Figure 7.** Color matrix for estimates of expected values for the mean incidence of citrus canker leaf, adjusted based on the regression model according to genotype and rootstock.

## Conclusion

The BeZI ( $\nu, \mu, \sigma$ ) regression model was adequate to represent the mean incidence of leaf citrus canker. The approach based on a regression model allows to isolate and quantify influence effects on the response variable.

The discrete component enabled to understand the odds of observing a zero mean incidence and quantitatively explained that the *Laranja Calpira* rootstock is the most resistant to leaf citrus canker as well *Limão Cravo* is more susceptible to the same disease. In addition, the *Ipiruá IAC*, *Arapongas*, *EEL* and *Olímpia* genotypes have odds of observing a zero mean incidence statistically equivalent.

The combination between the discrete and continuous components allowed to quantify the expected mean incidence and demonstrated that *Limão Cravo* rootstock and *Arapongas*, *EEL* and *Olímpia* genotype tend to positively increase the expected mean incidence. The highlight for lowest incidence expected average is the reference combination.

## Acknowledgements

The authors thank the scholar ships provided by Capes and Fundação Araucária.

## References

Agostini, J. P., Graham, J. H., & Timmer, W. L. (1985). Relation ship between development of citrus canker and rootstock cultivar for young 'Valencia' orange trees

in Misiones, Argentina. *Proceedings Florida State Horticultural Society*, 98(1), 19-22.

Anholeto, T., Sandoval, M. C., & Botter, D. A. (2014). Adjusted pearson residuals in beta regression models. *Journal of Statistical Computation and Simulation*, 84(5), 999-1014.

Cepeda-Cuervo, E. (2015). Beta regression models: joint mean and variance modeling. *Journal of Statistical Theory and Practice*, 9(1), 135-145.

Chien, L.-C. (2011). Diagnostic plots in beta-regression models. *Journal of Applied Statistics*, 38(8), 1607-1622.

Cribari-Neto, F., & Queiroz, M. P. (2012). On testing inference in beta regressions. *Journal of Statistical Computation and Simulation*, 84(1), 186-203.

Danos, E., Berger, R. D., & Stall, R. E. (1984). Temporal and spatial spread of citrus canker within groves. *Phytopathology*, 74(8), 904-908.

Ferrari, S. L., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7), 799-815.

Ferrari, S. L., Espinheira, P. L., & Cribari-Neto, F. (2011). Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica*, 65(3), 337-351.

Ferrari, S. L., & Pinheiro, E. C. (2011). Improved likelihood inference in beta regression. *Journal of Statistical Computation and Simulation*, 81(4), 431-443.

Gonçalves-Zuliani, A. M., Cardoso, K. A., Belasque Junior, J., Zanutto, C. A., Hashiguti, H. T., Bock, C. H., ... Nunes, W. M. (2016). Reaction of detached leaves from different varieties of sweet orange to inoculation with *Xanthomonas citri* subsp. *citri*. *Summa Phytopathologica*, 42(2), 125-133.

Ibáñez, M. V., Prades, M., & Simó, A. (2011). Modelling municipal waste separation rates using generalized

- linear models and beta regression. *Resources, Conservation and Recycling*, 55(12), 1129–1138.
- Jorgensen, B. (1997). *The theory of dispersion models*. London, UK: Chapman & Hall.
- Kieschnick, R., & McCullough, B. D. (2003). Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling*, 3(3), 193–213.
- Korhonen, L., Korhonen, K. T., Stenberg, P., Maltamo, M., & Rautiainen, M. (2007). Local models for forest canopy cover with beta regression. *Silva Fennica*, 41(4), 671–685.
- Latif, S., & Yab, M. Z. (2014). D-optimal designs for beta regression models with single predictor. *Journal of Statistical Computation and Simulation*, 85(9), 1709–1724.
- Leite Jr, R. P., & Santos, S. D. (1988). Suscetibilidade do limão Siciliano (*Citrus lemon*), enxertados sobre diferentes porta enxertos, ao cancro cítrico causado por *Xanthomonas campestris* pv. *citri*. *Fitopatologia Brasileira*, 13(4), 353–358.
- Mendonça, A., Silvestre, J., & Passos, J. (2011). The shrinking endogeneity of optimum currency areas criteria: Evidence from the European Monetary Union – A beta regression approach. *Economics Letters*, 113(1), 65–69.
- Moraes, R. M., Rocha, A. V., & Machado, L. S. (2012). Intelligent assessment based on Beta Regression for realistic training on medical simulators. *Knowledge-Based Systems*, 32(1), 3–8.
- Mullen, R., Marshall, L., & McGlynn, B. (2013). A beta regression model for improved solar radiation predictions. *Journal of Applied Meteorology and Climatology*, 52(8), 1923–1938.
- Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135(3), 370–384.
- Ospina, R., & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis*, 56(6), 1609–1623.
- Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9(4), 325–346.
- Pereira, T. L., & Cribari-Neto, F. (2014a). Detecting model misspecification in inflated beta regressions. *Communications in Statistics Computation and Simulation*, 43(3), 631–656.
- Pereira, T. L., & Cribari-Neto, F. (2014b). Modified likelihood ratio statistics for inflated beta regressions. *Journal of Statistical Computation and Simulation*, 84(5), 982–998.
- Reis, B., Koller, O. C., Schwarz, S. F., Theisen, S., Sartori, I. A., Nichele, F. S., ... Petry, H. B. (2008). Produção de frutos e incidência de cancro cítrico em laranjeiras “Monte Parnaso” enxertadas sobre sete porta-enxertos. *Ciência Rural*, 38(3), 672–678.
- Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, 54(3), 507–554.
- Rocha, A. V., & Simas, A. B. (2011). Influence diagnostics in a general class of beta regression models. *Test*, 20(1), 95–119.
- Sauer, A. V., Santos, E. M., Gonçalves-Zuliani, A. M., Nocchi, P. T., Nunes, W. M., & Bonato, C. M. (2015). Bacteriostatic and bactericidal activity *in vitro* of different essential oils as alternative treatments to control *Xanthomonas citri* subsp. *citri*. *Acta Horticulturae*, 1065(3), 931–936.
- Schafer, G., Bastianel, M., & Dornelles, A. L. (2001). Porta-enxertos utilizados na Citricultura. *Ciência Rural*, 31(4), 723–733.
- Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1), 54–71.
- Stuchi, E. S., Donadio, L. C., & Sempionato, O. R. (2002). Qualidade industrial e produção de frutos de laranja ‘Valencia’ enxertada sobre sete porta-enxertos. *Laranja*, 23(2), 453–471.
- Tazima, Z. H., Neves, C. S., Yada, I. F., & Leite Jr, R. P. (2010). Produção e qualidade dos frutos de clones de Laranja Pera no norte do Paraná. *Revista Brasileira Fruticultura*, 32(1), 189–195.
- Zhao, W., Zhang, R., Lv, Y., & Liu, J. (2014). Variable selection for varying dispersion beta regression model. *Journal of Applied Statistics*, 41(1), 95–108.

Received on August 11, 2016.

Accepted on January 7, 2017.

License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.