Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1679-9275
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v38i3.28689

Modeling citrus huanglongbing data using a zero-inflated negative binomial distribution

Eudmar Paiva de Almeida[1*], Vanderly Janeiro[1], Terezinha Aparecida Guedes[1], Fabio Mulati[2], José Walter Pedroza Carneiro[2] and William Mario de Carvalho Nunes[2]

*[1]Departamento de Estatística, Universidade Estadual de Maringá, Av. Colombo, 5790, 87020-900, Maringá, Paraná, Brazil. [2]Departamento de Agronomia, Universidade Estadual de Maringá, Maringá, Paraná, Brazil. *Author for correspondence. E-mail: eudmar.almeida@gmail.com*

ABSTRACT. Zero-inflated data from field experiments can be problematic, as these data require the use of specific statistical models during the analysis process. This study utilized the zero-inflated negative binomial (ZINB) model with the log- and logistic-link functions to describe the incidence of plants with Huanglongbing (HLB, caused by *Candidatus liberibacter* spp.) in commercial citrus orchards in the Northwestern Parana State, Brazil. Each orchard was evaluated at different times. The ZINB model with random effects in both link functions provided the best fit, as the inclusion of these effects accounted for variations between orchards and the numbers of diseased plants. The results of this model show that older plants exhibit a lower probability of acquiring HLB. The application of insecticides on a calendar basis or during new foliage flushes resulted in a three times larger probability of developing HLB compared with applying insecticides only when the vector was detected.

Keywords: mixed model, random effect, BLUP method, EM algorithm.

Modelagem de dados huanglongbing citricos usando um modelo binomial negativo inflacionado de zero

RESUMO. Em diversas áreas do conhecimento, dados com excesso de zeros são encontrados com frequência. Para a análise de tais dados, é recomendado utilizar modelos que permitam uma contagem deste excesso de zero de forma adequada. Neste artigo, o modelo binomial negativo inflacionado de zero (ZINB) foi utilizado para descrever o número de plantas doentes, acometidas por Huanglongbing, em pomares comercias de laranjeiras na região noroeste do Estado do Paraná. Entretanto, deve-se levar em consideração que cada pomar foi avaliado ao longo do tempo, sendo assim, neste contexto, o modelo ZINB com efeito aleatório em ambas as funções de ligação, logarítmica e logística, apresentou melhor ajuste aos dados, pois a introdução destes efeitos consideraram as variações entre os pomares e a dependência entre número de plantas doentes. A partir deste modelo é possível perceber que as plantas mais velhas tendem a apresentar menor probabilidade de adquirir a doença. Todavia, a aplicação de inseticidas e o manejo por calendário apresentam três vezes mais chance de apresentar a doença do que o manejo somente pela presença do vetor.

Palavras-chave: modelo misto, efeito aleatório, método BLUP, algoritmo EM.

Introduction

Orange (*Citrus sinensis*) production encompasses the largest acreage of any *Citrus* species in Brazil. *Citrus sinensis* was introduced in Brazil during the colonial period, but did not become the largest orange production acreage in the world until the mid-1980s (Couto & Canniatti-Brazaca, 2010). Frozen orange juice, which is largely exported, is one of the most important agricultural commodities in Brazil. International trade could further improve if the current phyto-sanitary issues did not inhibit orchard productivity in Brazil (Paulillo, 2006). The crop is affected by numerous diseases, but only a few are of agricultural importance. The most destructive citrus disease in Brazil is Huanglongbing (HLB, caused by the bacteria *Candidatus liberibacter* spp.). The disease poses threat to citrus crops around the world. The first cases were detected in southern China in 1919. The disease has since spread to 40 countries in Asia, Africa, Oceania, South America and North America (Bové, 2006). In 2004, HLB was first detected in the Americas, in Araraquara County, Central São Paulo State, Brazil (Teixeira et al., 2005).

HLB does not immediately kill the tree, but decreases the tree health and productivity over a period of years. These health issues eventually compromise the economic viability of entire orchards over periods of seven to ten years, particularly when disease control decisions are not

expediently made (Gottwald, da Graça, & Bassanezi, 2007). In healthy orchards, the bacterium is transmitted by the vector *Diaphorina citri* Kuwayama (Hemiptera: Psyllidae). In some cases, orchards can contain large *D. citri* populations, but be completely HLB free. Thus, a data set encompassing the variable of incidence of diseased plants may contain many zeros.

Data sets with an excess number of zeros are commonly analyzed using zero-inflated models. Famoye and Singh (2006) fit a zero-inflated Poisson model (ZIP) to investigate domestic violence. Xie, Wei, and Lin (2008) investigated pharmaceutical data using a ZIP model with a random effect. Hall (2000) used the same model to investigate repeated measures in agriculture experiments. However, data are over dispersed, with or without random effects, these models are not appropriate. Thus, the use of a zero-inflated negative binomial (ZINB) model is more appropriate for analyzing these types of data sets. Garay, Hashimoto, Ortega, and Lachos (2011) effectively used a ZINB model for over dispersed data. Yau, Wang, and Lee (2003) used a ZINB model with random effects to analyze the hospitalized times of pancreatic patients in different healthcare facilities.

The HLB incidence data collected from multiple orchards over multiple time periods requires a similar approach. Thus, we investigated the applicability of the ZINB mixed regression model to analyze the HLB incidence in Brazilian orchards. The excessive zero count is included in the ZINB model, along with over dispersion in the negative binomial distribution. Furthermore, the response variable dependence on repeated prevention measures over time is accommodated by a random effect that is incorporated into the linear predictive model.

Material and methods

HLB incidence

Sweet orange orchards were analyzed in this study. The orchards are located in the Northwestern Parana State, an area where HLB has been monitored due to the prevalence of *D. citri*. The scale of citrus production in this area ranges from small to large orchards. However, various size orchards are managed using similar methods, including application of pest and disease control measures. The HLB incidence data were collected on four occasions at intervals of 90 days in 2011.

The orchards were planted with the "Pera", "Valência" and "Folha Murcha" sweet orange (*Citrus sinensis*) varieties. Three methods were used to make management decisions related to controlling the vector. First, insecticide was applied only after detecting the presence of the vector. Second, insecticide was applied based on the timing of new foliage flushes. Finally, a calendar-based application was used based on three different tree ages – orchards with trees from 0 to five, six to ten and greater than eleven years old. These data were collected by projects supported by the Núcleo de Biotecnologia Aplicada from the Centro de Ciências Agrárias in the Universidade Estadual de Maringá, Maringá County, Parana State, Brazil.

Zero-inflated negative binomial model

Suppose that $Y = (Y_1, \ldots, Y_m)'$ is a discrete random variable, where $Y_1, \ldots, Y_m$ are independent variables following the zero-inflated negative binomial distribution given by:

$$P(Y=y) = \begin{cases} p + (1-p)\left(\frac{k}{k+\lambda}\right)^k, & y = 0 \\ (1-p)\frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\left(\frac{k}{k+\lambda}\right)^k\left(\frac{\lambda}{k+\lambda}\right)^y, & y \geq 1 \end{cases} \quad (1)$$

where p is the probability of observing at least one diseased plant, λ is the mean number of diseased plants and k is the dispersion parameter. When $k \to \infty$, $1/k$ approaches 0 as the negative binomial approaches the Poisson distribution. Thus, the ZIP and ZINB are closely related distributions, and the ZINB distribution can be seen as a flexible extension of the ZIP (Minami, Lennert-Cody, Gao, & Román-Verdesoto, 2007). The expected estimate and variance of the ZINB distribution are, respectively given by:

$$\mathrm{E}[Y] = (1-p)\lambda$$

and

$$\mathrm{Var}(Y) = (1-p)\lambda\left(\lambda p + 1 + \frac{\lambda}{k}\right).$$

Maintaining generality, when Y_{ij} ($i = 1, \ldots, m$; $j = 1, \ldots, n_i$ and $\sum_{i=1}^{m} n_i = n$) is the response variable, the number of diseased plants in the *i*-th orchard during the *j*-th evaluation, which considers that Y_{ij} follows the ZINB distribution with the inclusion of the ZIP model covariables, is based on Lambert (1992). We investigated the logarithmic-link function (*log*-link) of the parameter λ, which

was used to linearize the mean from the negative binomial. In addition, the logistic link function (*logit*-link) of the parameter p, which represents the proportion of zeros, was also analyzed. Furthermore, the observations can be treated independently among the orchards, but correlations may exist among records from the same orchards that were recorded at different times. These instances were explicitly modeled by incorporating a random effect into the linear predictors. Thus, the functions that connect the model parameters and covariates (the link functions) are described by:

$$\text{logit}(p_{ij}) = \xi_{ij} = G\gamma + u_i \quad (2)$$

and

$$\log(\lambda_{ij}) = \eta_{ij} = X\beta + v_i, \quad (3)$$

where X and G are the covariate matrices (m × n) considered in the models, which included management$_2$ (1 = insecticides applied to trees when producing new foliage flushes, 0 = otherwise), management$_3$ (1 = calendar application, 0 = otherwise), Pera (1 = Pera, 0 = otherwise), Valência (1 = Valência, 0 = otherwise), age$_2$ (1 = from 6 to 10 years old, 0 = otherwise), age$_3$ (1 = greater than 11 years old, 0 = otherwise), Evaluation$_2$ (1 = May, 0 = otherwise), Evaluation$_3$ (1 = August, 0 = otherwise) and Evaluation$_4$ (1 = December, 0 = otherwise). The management$_1$ (insecticide application only when a vector is present), Folha Murcha, age$_1$ (age from 0 to 5 years old) and Evaluation$_1$ (January) scenarios were adopted as references. β and γ represent the regression coefficients vectors, while $u = (u_1, \ldots, u_m)'$ and $v = (v_1, \ldots, v_m)'$ represent the unknown parameters with random effect vectors, respectively. We assumed that u and v were independent and distributed as $N(0, \sigma_u^2 I_n)$ and $N(0, \sigma_v^2 I_n)$, respectively. Based on Lambert (1992), the covariables affecting the mean in the ZIP model (the non-inflated component) may or may not be the same factors affecting the probability (p; the inflated component). Therefore, two data modeling approaches were developed, including the λ vector with a non-related p, or p as a function of λ. The same relationships were considered for the ZINB model. In addition, the odds ratio (OR) and relative risk (RR) of the link functions (*logit* and *log* with random effects) can be calculated.

Jiang (2007) applied BLUP (the best linear unbiased prediction, McGilchrist, 1994) procedures from the generalized linear mixed model (GLMM) method to maximize the sum of the components of the log-likelihood function $l = l_1 + l_2$. Through (1), l_1 is the log-likelihood function of y_{ij} given conditionally fixed u and v vectors based on the respective link functions (2 and 3), and l_2 is the log-likelihood for u and v. Thus, l_1 and l_2 are given by:

$$l_1 = \sum_{y_{ij}=0} \log\left(\frac{e^{\xi_{ij}} + t_{ij}^k}{1 + e^{\xi_{ij}}}\right) + \sum_{y_{ij}\geq 1}\left(\log\frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)} + k\log(t_{ij}) + y_{ij}\log(1-t_{ij}) - \log(1+e^{\xi_{ij}})\right)$$

and

$$l_2 = -\frac{1}{2}\left[m\log(2\pi\sigma_u^2) + u'u\sigma_u^{-2} + m\log(2\pi\sigma_v^2) + v'v\sigma_v^{-2}\right]$$

where $t_{ij} = \frac{k}{k + e^{\eta_{ij}}}$. The maximization of the log-likelihood function was achieved using the EM algorithm due to its stabilization ability (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 2007). Thus, we assumed a latent variable Z_{ij} following the Bernoulli distribution with the parameter p_{ij}, for which $z_{ij} = 1$ if the variable y_{ij} is derived from a class in the zeroes and $z_{ij} = 0$ otherwise. Hence, the log-likelihood function l_1 of Y and Z is:

$$l(y, z; \beta, \gamma, k) = \sum_{ij}[z_{ij}\xi_{ij} - \log(1 + e^{\xi_{ij}})] + \sum_{ij}(1 - z_{ij})\left[\log\frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)} + k\log(t_{ij}) + y_{ij}\log(1-t_{ij})\right].$$

Writing $l(y, z; \beta, \gamma, k)$ as $l_c = l_\xi + l_\eta$, the log-likelihood function for the complete data set with ξ_{ij} and η_{ij} separated to facilitate the parameter estimates is given by:

$$l_{\xi} = \sum_{ij}[z_{ij}\xi_{ij} - \log(1+e^{\xi_{ij}})] - \frac{1}{2}[m\log(2\pi\sigma_u^2) + u'u\sigma_u^{-2}]$$

$$l_{\eta} = \sum_{ij}(1-z_{ij})\left[\log\frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)} + k\log(t_{ij}) + y_{ij}\log(1-t_{ij})\right] -$$

$$-\frac{1}{2}[m\log(2\pi\sigma_v^2) + v'v\sigma_v^{-2}].$$

The EM algorithm replaces z_{ij} based on the conditional expected value $z_{ij}^{(l)}$ under the current estimates of $\hat{\gamma}^{(l)}$, $\hat{\beta}^{(l)}$, $\hat{u}^{(l)}$ and $\hat{v}^{(l)}$. Yau, Wang, and Lee (2003) reported that the expected value of z_{ij} is given by:

$$Z_{ij}^{(l)} = \begin{cases} (1+t_{ij}^{\hat{k}^{(l)}} e^{-(G\hat{\gamma}^{(l)}+\hat{u}^{(l)})})^{-1}, \text{if } y_{ij} = 0; \\ 0, \text{if } y_{ij} = 1. \end{cases}$$

The ($\hat{\gamma}^{(l+1)}$, $\hat{u}^{(l+1)}$) and ($\hat{\beta}^{(l+1)}$, $\hat{k}^{(l+1)}$, $\hat{v}^{(l+1)}$) estimates can be separated by setting the z_{ij} values in $z_{ij}^{(l)}$. The maximization of $l_c = l_{\xi} + l_{\eta}$ is achieved using two sets of Newton-Raphson algorithms, as given by:

$$\begin{bmatrix}\hat{\gamma}\\ \hat{u}\end{bmatrix} = \begin{bmatrix}\gamma_0\\ u_0\end{bmatrix} + I_{\gamma,u}^{-1}\begin{bmatrix}\frac{\partial l_{\xi}}{\partial \gamma}\\ \frac{\partial l_{\xi}}{\partial u}\end{bmatrix}$$

and

$$\begin{bmatrix}\hat{\beta}\\ \hat{v}\end{bmatrix} = \begin{bmatrix}\beta_0\\ v_0\end{bmatrix} + I_{\beta,v}^{-1}\begin{bmatrix}\frac{\partial l_{\eta}}{\partial \beta}\\ \frac{\partial l_{\eta}}{\partial v}\end{bmatrix}$$

where γ_0, u_0, β_0 and v_0 represent the initial values of γ, u, β and v, respectively. $I_{\gamma,u}$ is the second negative derivative of l_{ξ}, which accounts for β and u. $I_{\beta,v}$ is the second negative derivative of l_{η}, which accounts for β and v.

The maximization of the log-likelihood function $l_c = l_{\xi} + l_{\eta}$ assumes a dispersion parameter k and variance components σ_u^2 and σ_v^2. However, these parameters are generally unknown and must be estimated.

Based on the current estimates of $\hat{\beta}^{(l)}$, $\hat{v}^{(l)}$ and $z_{ij}^{(l-1)}$, l_{η} is maximized to estimate $\hat{k}^{(l)}$ from the dispersion parameter. Similarly, the Newton-Raphson algorithm estimates the elements of the linear predictor based on the initial σ_u^2 and σ_v^2 values. The components σ_u^2 and σ_v^2 are determined by the most recent values of $\hat{u}$ and $\hat{v}$ and the corresponding elements of the information matrix upon convergence. The residual maximum likelihood (REML) corrects the bias from the maximum likelihood by estimating the variance components (Yau & Lee, 2001; McGilchrist, 1994; Wood, 2011).

The method used in this article was developed using the R version 3.0.2. statistical software package, in which regression models with and without random effects were adjusted based on the program created by Andy Lee and Kelvin Yau in SPlus and adapted for R by Dave Atkins.

Results and discussion

The incidence of HLB diseased trees in orchards was frequently low, resulting in data sets with large numbers of zeros, justifying the use of zero-inflated models (Figure 1).

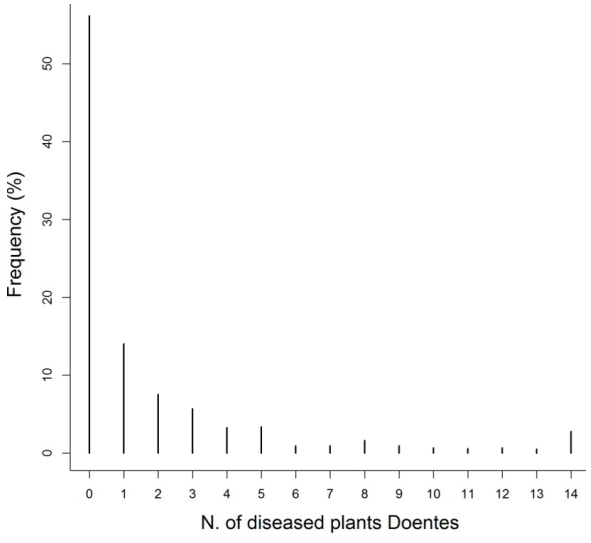

Figure 1. Incidence of HLB-diseased sweet orange trees (%).

When the ZINB model was fit to these data, the variables had no influence on the proportion of zeros (healthy trees) (Table 1).

The non-inflated component, negative binomial, management method (methods 2 and 3), the Pera variety and evaluation time (times 3 and 4) were significant factors affecting the incidence of diseased plants. The application of insecticides to trees when producing new foliage flushes ($management_2$) and the calendar-based approach ($management_3$) exhibited a higher incidence of HLB-diseased trees compared with orchards receiving insecticides only when the (*D. citri*) vector was observed in the orchard ($management_1$), where RR = 4.059 and CI (95%) = (3.025; 5.447), and RR = 3.158 and CI (95%) = (2.193; 4.547), respectively. The Pera variety also exhibited a higher incidence of HLB-diseased trees compared with "Folha Murcha", with RR = 1.459 and CI (95%) = (1.120; 1.901). Evaluation times 3 and 4 were displayed significantly higher incidences of HLB-diseased trees compared with evaluation time 1, with RR = 2.219 and CI (95%) = (1.615; 3.048), and RR = 1.872 and CI (95%) = (1.349; 2.597), respectively.

The estimate of $1/\hat{k}$ (1.977) indicates overdispersion in these data, providing compelling evidence to fit the ZINB model.

Table 1. Estimates, standard errors and confidence intervals (CI 95%) for the odds ratio and relative risk associated with the fitted ZINB model.

Variable	Estimate (SE)	
Inflated		*OR (CI 95%)*
Intercept	-5.479 (2.703)	0.004 (0.000; 0.834)
$Management_2$	1.634 (1.222)	5.124 (0.467; 56.211)
$Management_3$	0.649 (1.369)	1.914 (0.131; 28.001)
Pera	0.403 (0.487)	1.496 (0.576; 3.887)
Valência	0.100 (0.588)	1.105 (0.349; 3.499)
Age_2	-3.521 (5.697)	0.030 (0.000; 2090.419)
Age_3	3.519 (2.323)	33.751 (0.356; 3203.759)
$Evaluation_2$	-3.316 (2.917)	0.036 (0.000; 11.038)
$Evaluation_3$	-0.029 (0.534)	0.971 (0.341; 2.767)
$Evaluation_4$	0.483 (0.552)	1.621 (0.549; 4.782)
Non-Inflated		*RR (CI 95%)*
Intercept	-1.129 (0.224)	0.323 (0.208; 0.502)
$Management_2$	1.401 (0.150)⋆	4.059 (3.025; 5.447)
$Management_3$	1.150 (0.186)⋆	3.158 (2.193; 4.547)
Pera	0.378 (0.135)⋆	1.459 (1.120; 1.901)
Valência	-0.094 (0.145)	0.910 (0.685; 1.209)
Age_2	0.414 (0.141)⋆	1.513 (1.148; 1.994)
Age_3	0.183 (0.149)	1.201 (0.897; 1.608)
$Evaluation_2$	0.069 (0.163)	1.071 (0.778; 1.475)
$Evaluation_3$	0.797 (0.162)⋆	2.219 (1.615; 3.048)
$Evaluation_4$	0.627 (0.167)⋆	1.872 (1.349; 2.597)
$1/\hat{k}$		1.977 (0.148)⋆
Log-likelihood		-129.548
Deviance		169.116
Pearson residual		1185.451

⋆P-value > 0.05

This ZINB model (Table 1) assumes the independence of the response variable (incidence of HLB-diseased trees). However, the origin of the response variable is the incidence of HLB-diseased plants in every orchard, but at different times (repeated measures). Thus, the assumed correlation between the number of diseased plants and the lack of data independence is typical in this sort of analysis. Therefore, the ZINB model with random effects represents the preferable model for analyzing these repeat measurement data sets.

The ZINB with random effects estimates for both link functions (logistic and logarithm) are shown in Table 2. Both the magnitudes of the estimates and the standard errors decreased compared with the model described in Table 1. Furthermore, the relatively high variance component estimates ($\hat{\sigma}_u^2$ = 0.903) and odds ratio (OR) confidence intervals reinforce the need to incorporate the random effects. Plants older than eleven years (age_3) and at later assessment dates (evaluations 2, 3 and 4) yielded significantly different results. Thus, orchards with trees greater than eleven years old exhibited a lower HLB-disease incidence compared to orchards with 0 to 5 year old (age_1) trees (OR = 2.314 with CI (95%) = (1.426; 3.755)). Significant reductions in the number of zeros in the data set were observed during the later assessments dates (evaluations 2, 3 and 4), indicating an increase in the number of HLB-diseased plants over time.

In the non-inflated component of the model, the estimates and their standard errors displayed acceptable differences when compared with the ZINB model (Table 1). The variance component estimate ($\hat{\sigma}_v^2$ = 0.434) is relatively small compared with that observed in the zero-inflated component. Both the application of insecticides to trees producing new foliage flushes ($management_2$) and the calendar based approach ($management_3$) exhibited a significantly higher relative risks (RR = 3.304 and RR = 3.187) than orchards receiving insecticide only when the (*D. citri*) vector was observed in the orchard ($management_1$). Thus, the application of insecticides to trees with new foliage flushes and via the calendar based method correlated with a higher number of diseased plants than in orchards where insecticide is only applied when the vector is detected, with RR = 3.304 and CI (95%) = (2.281; 4.785), and RR = 3.187 and CI (95%) = (2.046; 4.963), respectively.

The highest *D. citri* vector populations generally occurred during the vegetative flush of young foliage. Increased infection rates are related to the highest populations of the vector, with outbreaks during the spring and summer periods (Yamamoto, Paiva, & Gravena, 2001; Bassanezi et al., 2010). The expected ratio of symptomatic to infected plants displayed high variability throughout the year, but was lower in the autumn and winter and higher in the spring and summer. Based on the current model (Table 2), a similar pattern was observed in the Northwestern Parana State, as the number of HLB-diseased plants increased during the year. The values calculated during evaluations 3 (August) and 4 (December) include RR = 1.471 and CI (95%) = (1.183; 1.829), and RR = 1.317, CI (95%) = (1.045; 1.659), respectively.

Table 2. Estimates, standard errors and confidence intervals (CI 95%) for the odds ratio and relative risk associated with the fitted ZINB model with random effects.

Variable	Estimate (SE)	
Inflated		*OR (CI 95%)*
Intercept	0.257 (0.438)	1.293 (0.548; 3.051)
$Management_2$	-0.256 (0.332)	0.774 (0.404; 1.484)
$Management_3$	0.215 (0.389)	1.240 (0.578; 2.658)
Pera	-0.071 (0.250)	0.931 (0.571; 1.520)
Valência	0.088 (0.276)	1.092 (0.636; 1.876)
Age_2	-0.411 (0.289)	0.663 (0.376; 1.168)
Age_3	0.839 (0.247)⋆	2.314 (1.426; 3.755)
$Evaluation_2$	-0.836 (0.257)⋆	0.433 (0.262; 0.717)
$Evaluation_3$	-0.738 (0.222)⋆	0.478 (0.309; 0.739)
$Evaluation_4$	-0.636 (0.224)⋆	0.529 (0.341; 0.821)
$\hat{\sigma}_u^2$		0.903
Non-Inflated		*RR (CI 95%)*
Intercept	-0.289 (0.252)	0.749 (0.457; 1.227)
$Management_2$	1.195 (0.189)⋆	3.304 (2.281; 4.785)
$Management_3$	1.159 (0.226)⋆	3.187 (2.046; 4.963)
Pera	0.109 (0.149)	1.115 (0.833; 1.493)
Valência	-0.145 (0.166)	0.865 (0.625; 1.198)
Age_2	0.230 (0.157)	1.259 (0.925; 1.712)
Age_3	0.067 (0.156)	1.069 (0.788; 1.452)
$Evaluation_2$	-0.143 (0.121)	0.867 (0.684; 1.099)
$Evaluation_3$	0.386 (0.111)⋆	1.471 (1.183; 1.829)
$Evaluation_4$	0.275 (0.118)⋆	1.317 (1.045; 1.659)
$\hat{\sigma}_v^2$		0.434
$1/\hat{k}$		0.232 (0.035)⋆
Log-likelihood		-1578.754
Deviance		140.016
Pearson residual		813.033

⋆P-value > 0.05.

The inclusion of random effects in the model reduces the estimate of the dispersion parameter from 1.997 to 0.232. However, these numbers are still significant, suggesting substantial overdispersion. The random effects model estimates exhibit greater precision, as illustrated by the OR values from ages 2 and 3 in the zero-inflated component (Table 1).

The log-likelihood, deviance and Pearson residual results verify that the zero-inflated negative binomial model with random effects in both link functions provides a better fit for the sampled data.

The quantile-quantile plots of the random effects u and v illustrate that the estimates possess a near-normal distribution, which can be partially used to validate the ZINB model with random effects (Figure 2).

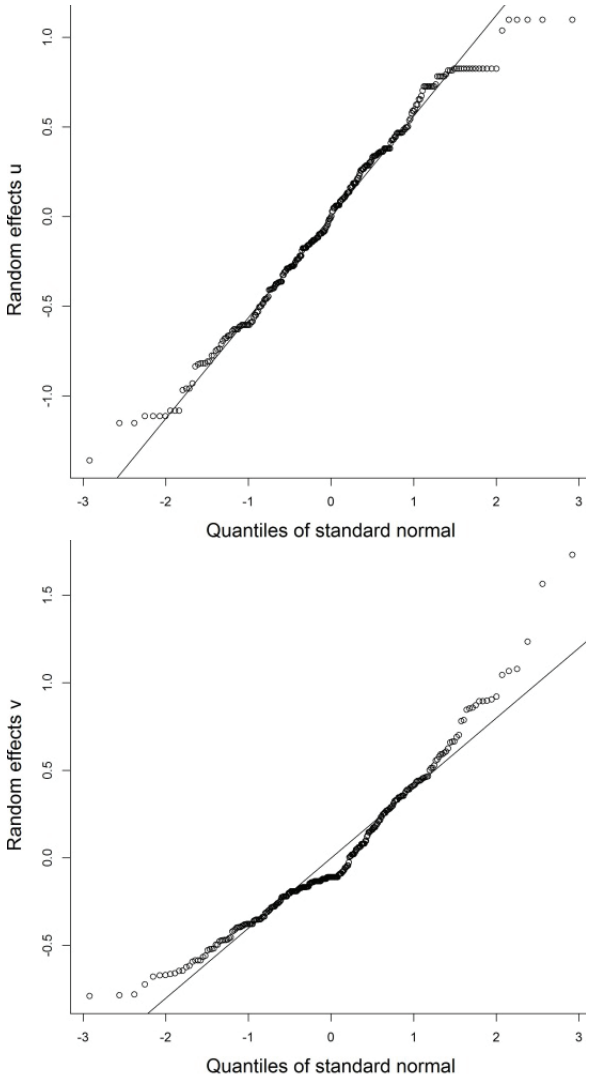

Figure 2. Quantile-quantile plots of the random effects *u* and *v* based on the ZINB model with random effects.

Furthermore, the half-normal plot, in which the residuals lie within the simulation envelope, suggests that the model provides a good fit, despite the overdispersion effect (Figure 3).

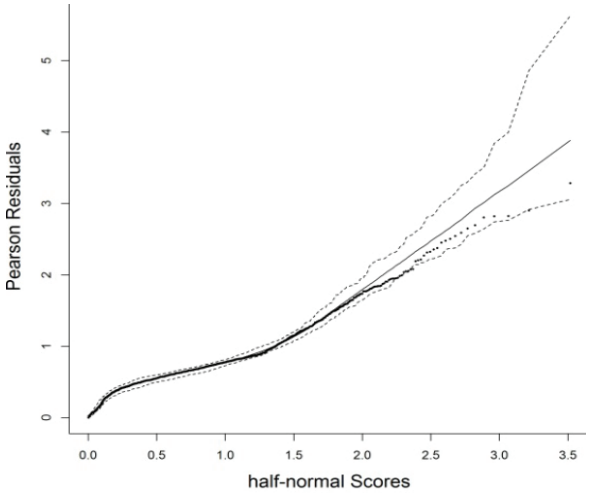

Figure 3. Half-normal plot for the ZINB model with random effects.

Conclusion

The ZINB model with the random effects provides a more appropriate fit for the HLB-diseased tree incidence over time compared to models that do not incorporate these random effects. Furthermore, we increased the parameter estimate precision by considering intercepts on an orchard by orchard basis.

We were able to detect that age_3 exhibited lower probabilities for contracting HLB-disease compared with trees less than five years old. In addition, the orchards exhibited a higher probability for possessing HLB-diseased trees as the season progressed. The application of $management_2$ and $management_3$ led to three times more diseased plants than the insecticide application method based on the presence of the vector.

Acknowledgements

We thank Kelvin Yau, a professor at City University of Hong Kong, for helping to analyze the data.

References

Bassanezi, R. B., Lopes, S. A., Belasque Júnior, J., Spósito, M. B., Yamamoto, P. T., Miranda, M. P., ... Wulff, N. A. (2010). Epidemiologia do *huanglongbing* e suas implicações para o manejo da doença. *Citrus Research & Technology, 31*(1), 11-23.

Bové, J. M. (2006). Huanglongbing: a destructive, newly-emerging, century-old disease of citrus. *Journal of Plant Pathology*, *88*(1), 7-37.

Couto, M. A., & Canniatti-Brazaca, S. G. (2010). Quantificação de vitamina C e capacidade antioxidante de variedades cítricas. *Ciência e Tecnologia de Alimentos, 30*(1), 15-19.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1-38.

Famoye, F., & Singh, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science, 4*(1), 117-130.

Garay, A. M., Hashimoto, E. M., Ortega, E. M., & Lachos, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis, 55*(3), 1304-1318.

Gottwald, T. R., Da Graça, J. V., & Bassanezi, R. B. (2007). Citrus huanglongbing: the pathogen and its impact. *Plant Health Progress*, *6*(1), 1-18.

Hall, D. B. (2000). Zero-inflated Poisson and Binomial regression with random effects: a case study. *Biometrics*, *56*(4), 1030-1039.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. New York, NY: Springer Science & Business Media.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1-14.

McGilchrist, C. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(1), 61-69.

McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions*. Hoboken, NJ: John Wiley & Sons.

Minami, M., Lennert-Cody, C. E., Gao, W., & Román-Verdesoto, M. (2007). Modeling shark by catch: the zero-inflated negative Binomial regression model with smoothing. *Fisheries Research, 84*(2), 210-221.

Paulillo, L. F. (2006). *Agroindústria e citricultura no Brasil: diferenças e dominâncias*. Rio de janeiro, RJ: Editora E-papers.

Teixeira, D. C., Saillard, C., Eveillard, S., Danet, J. L., Da Costa, P. I., Ayres, A. J., & Bové, J. (2005). "*Candidatus liberibacter americanus*", associated with citrus Huanglongbing (greening disease) in São Paulo State, Brazil. *International Journal of Systematic and Evolutionary Microbiology*, *55*(5), 1857-1862.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(1), 3-36.

Xie, F. C., Wei, B. C., & Lin, J. G. (2008). Assessing influence for pharmaceutical data in zero-inflated generalized Poisson mixed models. *Statistics in Medicine, 27*(18), 3656-3673.

Yamamoto, P. T., Paiva, P. E. B., & Gravena, S. (2001). Flutuação Populacional de *Diaphorina citri* Kuwayama

(Hemíptera: Psyllidae) em Pomares de Citros na Região Norte do Estado de São Paulo. *Neotropical Entomology, 30*(1), 165-170.

Yau, K. K., & Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, *20*(19), 2907-2920.

Yau, K. K., Wang, K., & Lee, A. H. (2003). Zero-inflated negative Binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, *45*(4), 437-452.

Received on July 29, 2015.
Accepted on January 19, 2016.