University of São Paulo
"Luiz de Queiroz" College of Agriculture

Statistical modeling for zero-inflated data using GAMs

Marcus Vinicius Silva Gurgel do Amaral

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2019

Marcus Vinicius Silva Gurgel do Amaral
Forest Engineer

Statistical modeling for zero-inflated data using GAMs

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Profa. Dra. **SÔNIA MARIA DE STEFANO PIEDADE**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Amaral, Marcus Vinicius Silva Gurgel do
Statistical modeling for zero-inflated data using GAMs / Marcus Vinicius Silva Gurgel do Amaral. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2019 .
60 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Modelos lineares generalizados 2. Modelos aditivos generalizados 3. Simulação de dados 4. Softare R 5. Modelagem ecológica . I. Título.

To Mum and Dad

ACKNOWLEDGMENTS

I would like to thank God for my existence and all the oportunities I've had until now. Mum and Dad, and all my family, for all the love and support thoughtout my life.

I would also like to thank my supervisor, Sonia Maria D'Estefano Piedade, for believing in me and helping me when it was most needed.

This work would not be possible without my friend Rafael Moral. He was not only a friend, he is a brother and a tutor. Thanks for Natalie Veronika for all the patience, Thiago for all the support acting like a brother sometimes, and all the friends I've made during the time studying, so many names that could be difficult to enumerate (Solange, Jorge, Eduardo, Wellington, Ricardo, Djair, ...).

Finnaly, many thanks to CAPES for providing financial support in Brazil not only to my work, but to all the research made along these years.

"If you have a positive attitude and strive to give your best,
eventually, you will overcome problems and find you are ready
for greater challenges."

Pat Riley

"Let us rise up and be thankful, for if we didn't learn a lot today,
at least we learned a little, and if we didn't learn a little,
at least we didn't get sick, and if we got sick, at least we didn't die;
so let us all be thankful."

Buddha

"What I thought possible, made my path difficult.
What could be a miracle showed me that I bleed.
Nothing, during this time, made me invincible but,
from now on, I can keep my way searching happiness.
And now I'm done of sleepless nights worrying about who,
or what, tried to steal the colors of my paradise."

Marcus Vinicius S. G. do Amaral

SUMÁRIO

Resumo . . . 7
Abstract . . . 8
List of Figures . . . 9
List of Tables . . . 10
List of Abbreviations and Acronyms . . . 11
Symbols List . . . 12
1 Introduction . . . 13
2 Statistical modeling of zero-inflated longitudinal count data in entomology . . . 15
2.1 Introduction . . . 15
2.2 Case Study . . . 16
2.3 Statistical methods . . . 19
2.3.1 Introduction to generalized linear models (GLMs) . . . 20
2.3.1.1 Poisson model . . . 21
2.3.1.2 Overdispersion . . . 21
2.3.1.3 Negative Binomial model . . . 21
2.3.1.4 Zero Inflation . . . 22
2.3.1.5 Zero Inflated Poisson model (ZIP) . . . 22
2.3.1.6 Zero Inflated Negative Binomial model (ZINB) . . . 22
2.3.2 Introduction to generalized additive mixed models (GAMs) . . . 23
2.3.2.1 Smoothing methods using splines . . . 23
2.3.2.2 Cubic splines . . . 23
2.4 Case study analysis . . . 24
2.4.1 Fitting Poisson models with polynomial functions . . . 24
2.4.2 Fitting negative binomial models . . . 29
2.4.3 Including splines in the linear predictor . . . 30
2.5 Discussion . . . 35
3 Simulation study for count data using GAMs . . . 37
3.1 Introduction . . . 37
3.2 Count data models . . . 38
3.2.1 Poisson and Negative binomial models . . . 38
3.2.2 Zero-inflated models: ZIP and ZINB . . . 39
3.3 Simulation design . . . 40
3.4 Results . . . 41
3.5 Discussion . . . 44
References . . . 47
Appendix . . . 51

RESUMO

Modelagem estatística para dados zero-inflacionados usando GAMs

Dados de contagem são comuns em estudos biológicos, como em entomologia, em que são observados números de indivíduos ou proporções. Geralmente esses experimentos ou processos de amostragem apresentam superdispersão e excesso de zeros. O emprego de modelos zero-inflacionados auxilia o processo de estudo para conjuntos de dados com esses comportamentos. O primeiro trabalho aqui apresentado trata de um experimento conduzido com uma praga de algodão. O objetivo principal é comparar os modelos para dados de contagem superdispersos zero-inflacionados. Os modelos lineares generalizados e os modelos aditivos generalizados são comparados em termos de ajuste de acordo com a inclusão de funções de suavização. Após a seleção do modelo, foram incluídos covariáveis de inimigos naturais. Modelos que utilizaram funções de suavização permitiram uma melhor avaliação das interações biológicas ao longo do tempo entre a praga e os seus inimigos naturais. O segundo trabalho trata de um estudo de simulações que visa comparar a eficiência da inclusão de funções de suavização para conjuntos de dados simulados zero-inflacionados longitudinais. A inclusão de funções de suavização não apresentou melhora no ajuste dos modelos para os cenários de simulação criados.

Palavras-chave: Modelos lineares generalizados; Modelos aditivos generalizados; Simulação de dados; Softare R; Modelagem ecológica

ABSTRACT

Statistical modeling for zero-inflated data using GAMs

Count data are common in biological studies, such as in entomology, where numbers of individuals or proportions are observed. Generally, these experiments or sampling processes have overdispersion and excess of zeros. The use of zero-inflated models supports the study process for data sets with these behaviors. The first paper presented here deals with an experiment conducted with a cotton pest. The main objective is to compare the models for zero-inflated overdispersed count data. Generalized linear models and generalized additive models are compared in terms of fit according to the inclusion of smoothing functions. After the selection of the model, covariates of natural enemies were included. Models that used smoothing functions allowed a better evaluation of the biological interactions over time between the pest and its natural enemies. The second work deals with a simulation study that aims to compare the efficiency of the inclusion of smoothing functions for simulated data sets of zero-inflated longitudinal data. The inclusion of smoothing functions did not show improvement in the fit of the models for the created simulation scenarios.

Keywords: Generalized linear models, Generalized additive models, Data simulation, R Software, Ecological modelling

LIST OF FIGURES

2.1 Representation of the randomization scheme of the experiment with the definition of the distinct sampling processes used to observe *Aphis gossypii* and natural enemies. 16
2.2 Graph of average trends for the different spacings considering winged aphids throughout the observation weeks for the three spacings (0.4m, 0.8m, 1.6m) and the three Sections (1,2,3). 17
2.3 Graph of average trends for the different spacings considering wingless aphids throughout the observation weeks for the three spacings (0.4m, 0.8m, 1.6m) and the three Sections (1,2,3). 17
2.4 Graphic representation for the mean values of the main natural enemies in each spacings, along 16 observation weeks. 18
2.5 Mean Vs Variance for: (a) wingless and (b) winged. 18
2.6 Proportion of zeros presented in the observations during the weeks of conduction of the experiment considering spacing between plants of the same planting line. 19
2.7 Half-normal plot for the different degrees of polynomial functions for winged and wingless aphids. 26
2.8 Half-normal plot to compare the different models for winged aphids and 5-th degree of the polynom. 27
2.9 Half-normal plot to compare the different models for wingless aphids (a) 5-th degree of the polynom, (b) interaction, (c) ZIP (constant) and ZIP (regression). 28
2.10 Residuals versus dependent variables for wingless aphids (a) and winged aphids (b). 29
2.11 Residuals versus predicted variables for wingless aphids (a) and winged aphids (b). 29
2.12 Half-normal plot to compare the different models for winged aphids and 5-th degree of the polynom. 31
2.13 Half-normal plot to compare the different models for wingless aphids and 5-th degree of the polynom. 32
2.14 Wormplot of the selected models for wingless aphids and winged aphids respectively. 33
2.15 Wormplot of the selected models for wingless aphids and winged aphids respectively. 34

LIST OF TABLES

2.1 Likelihood ratio test for nested models to choose the polynomial degree of the Poisson model for winged aphids. . . . 25
2.2 Likelihood ratio test for nested models to choose the polynomial degree of the model for wingless aphids. . . . 25
2.3 Likelihood ratio test for nested models to evaluate the inclusion of natural enemies for wingless aphids with the ZIP 5-th degree polynomial model. . . . 27
2.4 Likelihood ratio test for nested models to evaluate the inclusion of natural enemies for wingled aphidswith the ZIP 5-th degree polynomial model. . . . 28
2.5 Likelihood ratio test for nested models to evaluate the inclusion of natural enemies for winged aphids with the ZINB 5-th degree polynomial model. . . . 30
2.6 Likelihood ratio test for nested models to evaluate the inclusion of natural enemies for wingless aphids with the ZINB 5-th degree polynomial model. . . . 30
2.7 Values of AIC for model selection using splines. . . . 33
2.8 Values of AIC for model selection using splines including each natural enemy. . . . 34
2.9 Values of AIC for model selection using splines 34
2.10 Values of AIC for model selection using splines including each natural enemy. . . . 35

3.1 Parameters of the respective distributions used in the simulation scenarios. . . . 41
3.2 Proportion of selected models for ZINB simulation with respective values of the parameters. 42
3.3 Proportion of selected models for ZIP simulation with respective values of the parameters. . 42
3.4 Proportion of selected models for ZINB simulation with respective values of the parameters. 42
3.5 Mean estimates of the parameters for the ZINB simulation and respective values of the parameters. . . . 43
3.6 Mean estimates of the parameters for the ZIP simulation and respective values of the parameters. . . . 44
3.7 Mean estimates of the parameters for the non-linear ZINB simulation and respective values of the parameters. . . . 44

LIST OF ABBREVIATIONS AND ACRONYMS

GAM	Generalized additive models
GAMLSS	Generalized additive models for location shape and scale
GLM	Generalized linear models
hnp	Half-normal plot
N.B.	Negative binomial
PO	Poisson
ZINB	Zero-inflated negative binomial
ZIP	Zero-inflated Poisson

SYMBOLS LIST

α	alpha greek
β	beta
χ^2	chi ditribution
Γ	gamma distribution
μ	mean or location parameter
ν	scale parameter
σ	variance or scale parameter

1 INTRODUCTION

There are many factors that can contribute to the production of a given crop. The presence of large numbers of pests, such as insects, is one of the main variables that can lower productivity. To maximize productivity we must learn about the biological relationships of the pest species with the crop before carrying out appropriate management practices (Kidd and Amarasekare, 2012). This may comprise not only characteristics inherent to the population but also trophic relationships with other species, such as predation and parasitism, which must be considered when developing biological control strategies.

Different forms of statistical modeling can be employed for the evaluation of ecological processes considering characteristics of the data. When working with generalized linear models (GLMs) (Nelder and Wedderburn, 1972), different types of distributions that belong to the exponential family of distributions can be used. We may also include smoothing functions for the covariates present in the linear predictor the model resulting in a generalized additive model (GAM) or a generalized additive mixed model (GAMM) (Hastie and Tibshirani, 1986).

Normally for count data, the relationship between observations and explanatory variables are treated using models involving Poisson or negative binomial distributions. Once the Poisson model assumes equality of the mean and variance, negative binomial models have greater flexibility modeling the relationships between the mean and the variance (Colin and Pravin, 2013). Although they are present in many packages of statistical software, negative binomial models are limited to modeling over-dispersed data, and unable to deal with under-dispersed data (McCullagh and Nelder, 1989). But it is not uncommon to find zero-inflation in count data in addition to overdispersion. Zero inflated Poisson models (ZIP) (Lambert, 1992) and Zero inflated negative binomial models (ZINB) (Greene, 1994) are alternatives used to deal with these characteristics. Although used in many areas such as psychology (Atkins and Gallop, 2007) and computer science (Souza *et al.*, 2016), if the non-zero part of data is over-dispersed, the parameter estimates of a ZIP can be biased just as standard errors may be underestimated. It is also possible to use smoothing functions in these models for prediction of the nonparametric regression or longitudinal effect, estimating a nonparametric function that minimizes the penalized least squares criterion (Aydin *et al.*, 2013).

Here we have the GAM theory applied initially to explain the biological dynamics of a species of agricultural pest, in which smoothing theory is applied to improve model fit. In the second case, a simulation study is made to compare the efficiency of model fit with the inclusion of smoothing functions.

2 STATISTICAL MODELING OF ZERO-INFLATED LONGITUDINAL COUNT DATA IN ENTOMOLOGY

Abstract

Count data are common in biological studies, such as in entomology, in which numbers of individuals or proportions are observed. In these experiments the occurrence of overdispersed and zero-inflated data is common. This work studies longitudinal data of a cotton plague with the purpose of fitting and comparing models for counting data. The generalized linear models and generalized additive models are compared in goodnes of fit according to the inclusion of smoothing functions. After the model selection, covariates of natural enemies were included. Models that used smoothing functions allowed a better evaluation of biological interactions over time between the pest and its main natural enemies.

Keywords: Count data; entomology; generalized linear models; generalized additive models; smothing functions.

2.1 Introduction

There are many factors that can contribute to the production of a given crop. The presence of large numbers of pests, such as insects, is one of the main variables that can lower productivity. To maximize productivity we must learn about the biological relationships of the pest species with the crop before carrying out appropriate management practices (Kidd and Amarasekare, 2012). This may comprise not only characteristics inherent to the population but also trophic relationships with other species, such as predation and parasitism, which must be considered when developing biological control strategies.

There are many specific deterministic models in the area of entomology for the explanation of interactions among species whose main purpose is to describe the essence of biological processes (Kot, 2001). Models that include interactions between predator and prey, competitors and hosts, as well as parasitoids are common in the context of biological control (Battel *et al.*, 2012).

Different forms of statistical modeling can be employed for the evaluation of ecological processes considering characteristics of the data. When working with generalized linear models (GLMs) (Nelder and Wedderburn, 1972), different types of distributions that belong to the exponetial family of distributions can be used. They have statistical properties that aid in modeling, and the main objective is to establish a relationship between the response variable and the explanatory variable. To evaluate the population dynamics of a given individual, the population size should be observed over time, which characterizes a longitudinal study whose main characteristic is the correlation between observations throughout the study period. An alternative to model the correlation structure between observations is the use of the generalized estimation equations (GEE) approach that allows for the modeling of the correlation between observations (Liang and Zeger, 1986). We can also account for the correlation between repeated measures by including random effects, characterizing a generalized linear mixed model (GLMM).

We may also include smoothing functions for the covariates present in the linear predictor the model resulting in a generalized additive model (GAM) or a generalized additive mixed model (GAMM) (Hastie and Tibshirani, 1986).

The data set that will be explored in this paper refers to an entomological study that aims to study the population development of an insect pest from the cotton crop with the objective of biological

control. The data from these experiments consist of discrete variables, numbers of aphids and their main natural enemies, observed throughout time in cotton.

By carrying out exploratory analyses, we expect to obtain insight on how to model the data properly. Starting from the exploratory analysis we have the elaboration of simple models to more complex models that involve overdispersion, zero-inflation and the use of smoothing functions.

2.2 Case Study

Considering the context of entomological experiments, which are carried out for different purposes, the case study here is an experiment set up in a randomized complete block design. It was carried out to test the influence of diferent spacings between plants in the same planting line, on the population dynamics is a pest of cotton, the aphid *Aphis gossypii* conducted at EMBRAPA Algodão, Campina Grande - PB - Brazil. Three diferent levels of spacing (0,4 m, 0,8 m e 1,6m) were randomized into four blocks, totaling 12 plots (Fig. 2.1). Using simple random sampling, five plants within each plot were marked and observed in regular intervals of seven days over sixteen weeks, each one in three diferent regions of the plant: basal, median and apical region. The observed variables were the number of wingless aphids and the number of winged aphids. Six different natural enemies were also observed, namely, *Lysipheblus testaceips*, *Chrysopidae* (Green Lacewings), *Scymnus* (Ladybug), *Cycloneda* (Ladybug), *Syrphidae* (flies) and spiders.

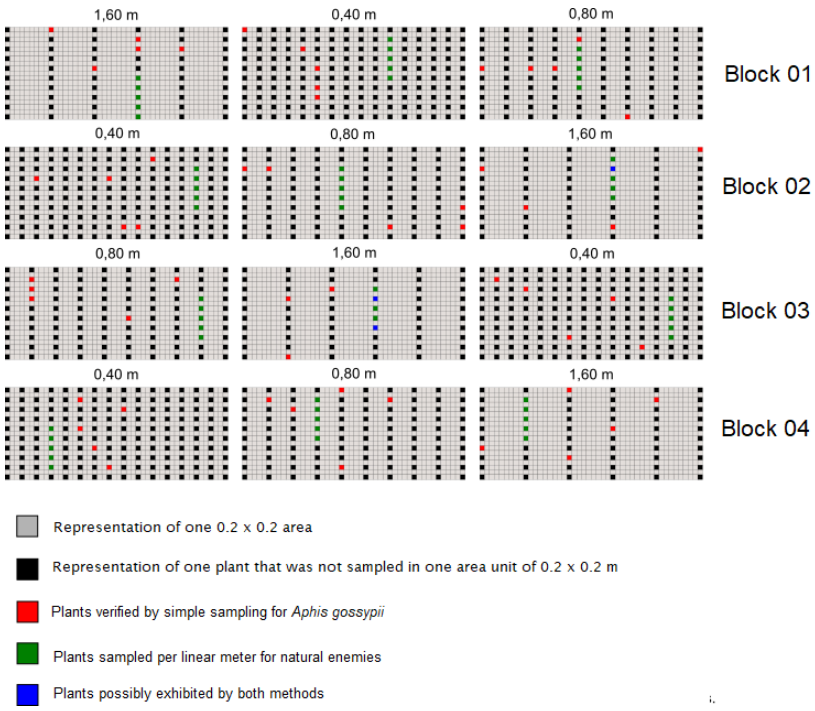

Figure 2.1. Representation of the randomization scheme of the experiment with the definition of the distinct sampling processes used to observe *Aphis gossypii* and natural enemies.

To see the behavior of the data, a graph of average trends was elaborated.

The exploratory analysis reveals indications that the means of observed variables are not constant over time Figure 2.2 and 2.3. The same type of graphical evaluation was performed for natural enemies with the objective of observing the behavior of each one during the experiment. Figure 2.4

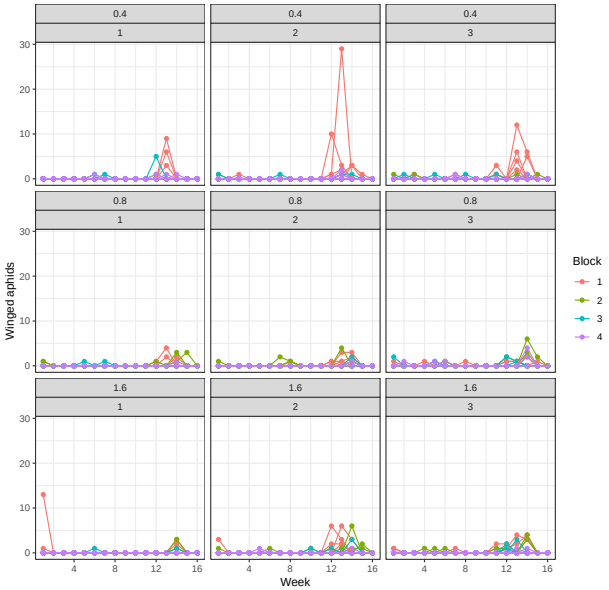

Figure 2.2. Graph of average trends for the different spacings considering winged aphids throughout the observation weeks for the three spacings (0.4m, 0.8m, 1.6m) and the three Sections (1,2,3).

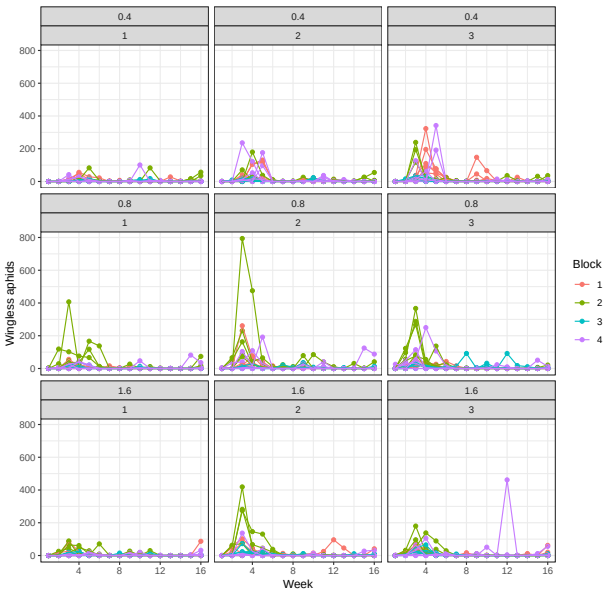

Figure 2.3. Graph of average trends for the different spacings considering wingless aphids throughout the observation weeks for the three spacings (0.4m, 0.8m, 1.6m) and the three Sections (1,2,3).

presents the values of the variable natural enemies observed throughout the experiment that, apparently, vary over time.

Because this is an experiment whose observed variables are counts, that is, discrete variables, the most appropriate approach is the use of generalized linear models (Nelder and Wedderburn, 1972), and the initial models may comprise distributions such as the Poisson for counts.

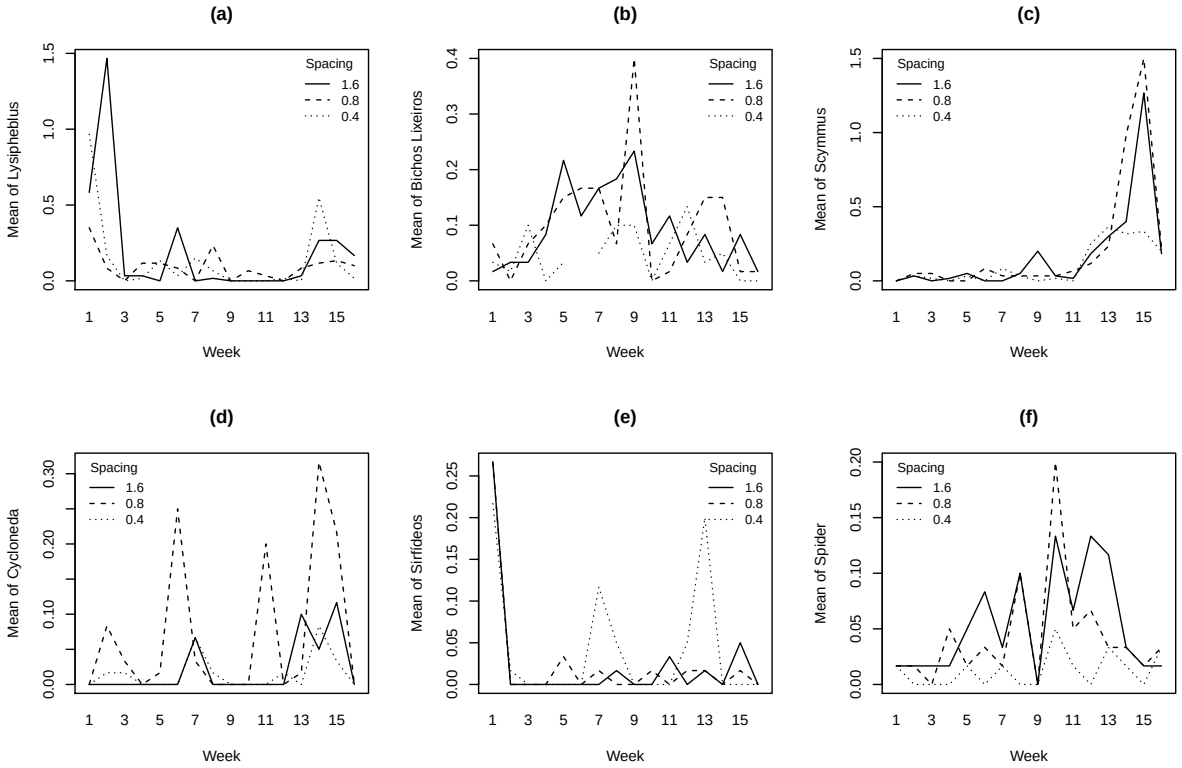

Figure 2.4. Graphic representation for the mean values of the main natural enemies in each spacings, along 16 observation weeks.

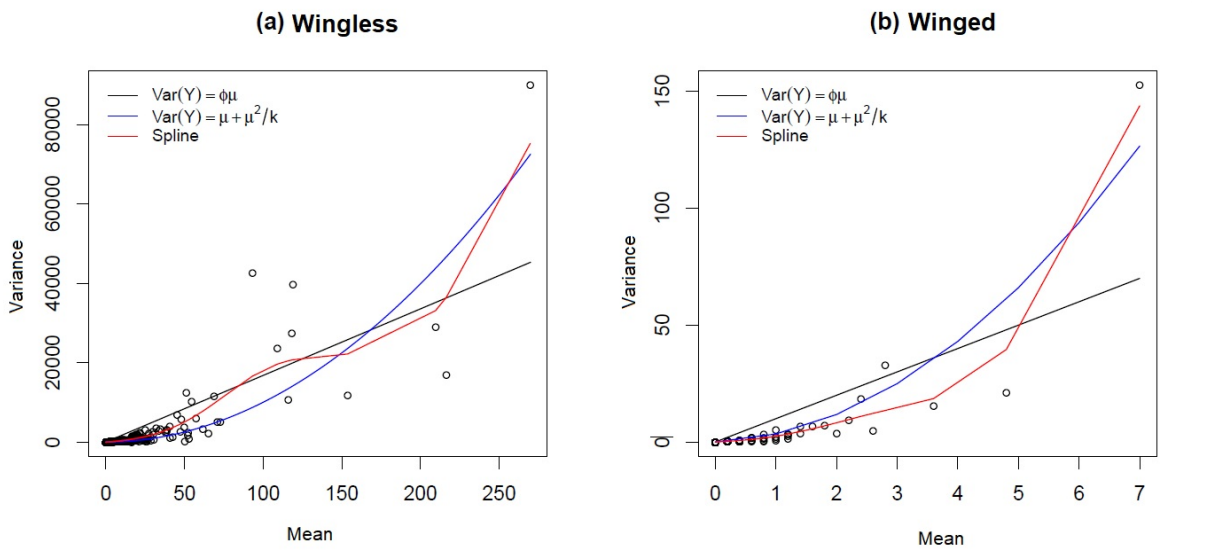

Figure 2.5. Mean Vs Variance for: (a) wingless and (b) winged.

As part of the exploratory analysis we can verify indicatives of overdispertion by using mean versus variance plots. We can evaluate indications of overdispersion by using mean vs variance plots. As for counting data, the behavior of the data is initially compared, for mean and variance, with Poisson

and binomial distributions, if the variation is much larger than the mean, we have evidence that there is overdispersion. Here there is evidence that the data are overdispersed (Figure 2.5), because the variances are much larger than the means. The behavior of the variances in relation to the means is apparently quadratic, suggesting the use of a negative binomial model, for example, to accommodate the overdispersion of the data.

However, part of the overdispersion can be due to the zero inflation as verified in Figure 2.6. Considering the verification of the proportion of zeros in relation to the spacings (Figure 2.6), the periods of smaller occurrences of zeros coincide with the periods of greatest number of observations where both wingless and winged aphids were recorded.

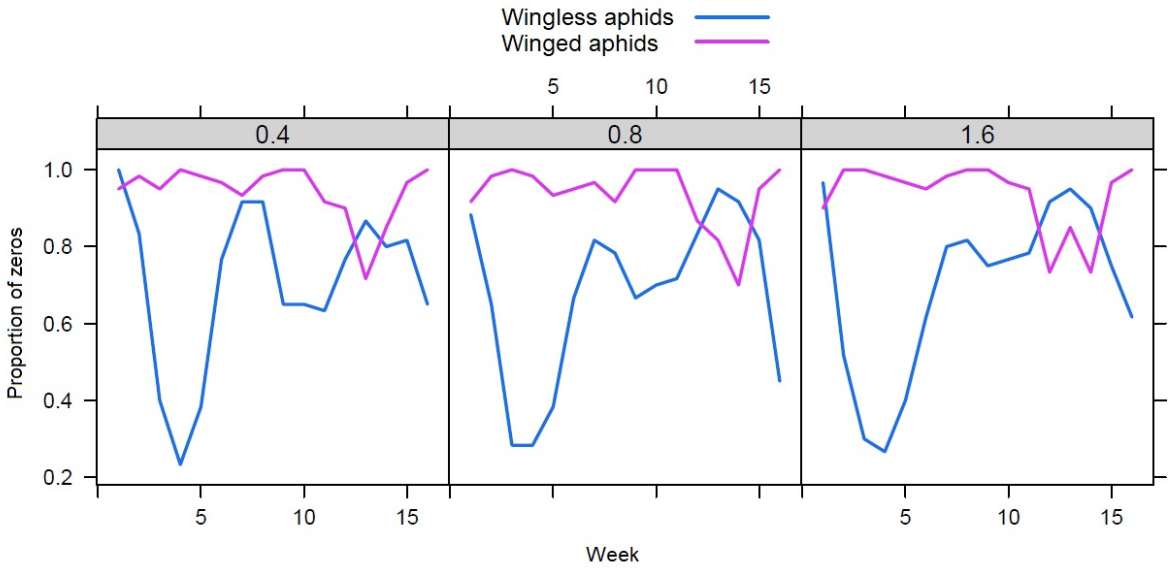

Figure 2.6. Proportion of zeros presented in the observations during the weeks of conduction of the experiment considering spacing between plants of the same planting line.

By observing the dynamics of the occurrence of zeros considering both spacing between plants (applied treatment) and section of the plants, there is a decrease in the number of zeros near the beginning of the observations for wingless aphids whereas, for winged individuals, apparently the occurrence of zeros is closer to the end of the experimental period.

2.3 Statistical methods

The main variable observed in this experiment is the number of aphids, both winged and wingless, represented by the random variable Y_{ijkt} with, $i = 1, 2, 3, 4; j = 1, 2, 3; k = 1, 2, 3; t = 1, ..., 16$ If we assume each observational unit has a random number os aphids, it's reasonable to assume a Poisson model and a Negative Binomial as the starting points to analyze the data, $Y_{ijkt} \sim PO(\mu)$ and $Y_{ijkt} \sim NB(\mu, \sigma)$, and their extentions, such as the ones discussed in later sections. These particular ones are examples of generalized linear models (Nelder and Wedderburn, 1972).

2.3.1 Introduction to generalized linear models (GLMs)

The class of models known as generalized linear models (GLMs) is defined by three distinct components: The first one, called the random component, corresponds to the random variables $Y_1, \ldots, Y_n$ that belong to the exponential family of distributions each in terms of a distinct parameter θ_i. The exponential family in the canonical form has a density or probability density function (pdf) expressed as:

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i; \phi)\}, \tag{2.1}$$

where $b(\cdot)$ and $c(\cdot)$ are known functions, $\phi > 0$ is a dispersion parameter and θ_i is called canonical parameter. As members of the exponential family, the normal, Poisson, binomial, gamma, inverse normal and binomial negative distributions (each with its appropriate dispersion parameters) can be expressed in the canonical form (2.1).

Another characteristic as exponential family members is the definition of the expectation given by $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$ and $\mathrm{Var}(Y_i) = \phi b''(\theta_i) = \phi V_i$, where $V_i = V(\mu_i) = d\mu_i/d\theta_i$ is called the variance function and depends only on the mean μ_i.

The second component is linear predictor related to the explanatory variables defined as

$$\eta = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x},$$

where $\boldsymbol{\beta}$ is a vector of p unknown parameters and $\mathbf{x} = [x_1, \ldots, x_n]'$ is the i-th column of the $n \times p$ design matrix. The last component is a link function $g(\mu_i) = \eta_i$, relating the systematic to the random component (Hinde and Demétrio, 1998).

To assess the significance of the effects in the linear predictor (Nelder and Wedderburn, 1972) proposed the analysis of deviance, a measure that compares a fitted model to the saturated model (i.e. a complete model with one parameter per observation). For a known ϕ, it can be used as a measurement of goodness-of-fit for the fitted models. For the Poisson model, the residual deviance can be written as

$$D_P = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)\right]$$

where $\hat{\mu_i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p)$ with $i = 1, 2, \ldots, n$, on the fitted values for the current model. Asymptotically (i.e. for a large sample size), D_P has an approximate χ^2 distribution with $n-p$ degrees of freedom (df). To compare nested models, that are models containing the same terms and one has at least one additional term, by writing D_p for the residual deviance of the full model and D_q for the reduced model, the statistic $D_q - D_p \sim \chi^2_{q-p}$ can be used to test the hypothesis that true coefficient values of omitted terms are zero. This test corresponds to a likelihood ratio test, and if $D_q - D_p > \chi^2_{q-p;(1-\alpha)}$, the upper $100 \times \alpha$ percentile of the χ^2_{q-p} distribution, we reject the null hypothesis that the additional parameter is zero at a significance level of α, which means that the parameters tested are important to describe the data and should remain in the model (Demétrio, C. G. B., Hinde, J., & Moral, 2014).

As a possible diagnostic tool, we can use plots to detect failure on the model fitting, comparing observed and fitted values using a chosen residual, the deviance residuals. It is also possible to check the goodness-of-fit of a model by contructing a half-normal plot of the residuals (Atkinson, 1985). The R package *"hnp"* (Moral *et al.*, 2017) produces half-normal plots with a simulades envelope that should include where most residuals if the observed data were a plausible realisation of the fitted model, making it possible to detect overdispersion in the data. A satisfactory fit should show a maximum 5% of the points outside the envelope, usually.

2.3.1.1 Poisson model

Assuming Y a random variable with a Poisson distribution, that is, $Y \sim \text{PO}(\mu)$, where $\mu > 0$, its probability function may be written as

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \quad \text{for} \quad y = 0, 1, 2, ...$$

The Poisson distribution has an important role in modeling the behavior of count data. It provides a good representation for experimental data whose variance is equal to the mean, a phenomenon known as equidispersion. In entomology, however, it is very rare to have a good fit of the Poisson model to data from counts of insects, eggs, among others. The canonical link function for the Poisson model is the logarithmic function, i.e. $\eta = \log(\mu)$.

2.3.1.2 Overdispersion

Overdispersion is a phenomenon in which the variability is greater than expected by the Poisson and binomial models. In the Poisson model, for example, it is assumed that the variance is equal to the mean, a characteristic that rarely occurs in entomology. In addition to inherent characteristics of the distributions, experimental conditions can generate overdispersion, among them we can mention: natural variability of the experimental material, the correlation between individual responses, hierarchical or aggregated data structure and omission of covariates in the linear predictor (Hinde and Demétrio, 1998).

Assuming a model with correct linear predictor and link function fitted to a data set whose variability is greater than expected, consistent, converging in probability to the true value of the parameters, the standard errors are underestimated. Therefore, the selection of models and hypothesis tests in relation to the parameters can be compromised, leading to the selection of more complex models and, consequently, incorrect conclusions about the scientific hypotheses.

A simple check of the occurrence of overdispersion in a set of counts or proportions data can be done by comparing the residual deviance with the number of residual degrees of freedom. Since, asymptotically, the residual deviance has a chi-square distribution with degrees of freedom equal to the number of residual degrees of freedom, it is expected that these quantities will be approximately equal for a Poisson or Binomial model, if they are a good fit. When the residual deviance is much greater, there is evidence of overdispersion. In cases like this, finding a distribution that can accommodate the variability is a better choice. The negative binomial distribution (NB) is an alternative.

2.3.1.3 Negative Binomial model

The probability function of Negative Binomial type-I distribution $Y \sim \text{NB}(\mu, \sigma)$ can be written as

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

for $y = 0, 1, 2, ...$, $\mu > 0$ and $\sigma > 0$ (Anscombe, 1949).

The Negative Binomial type-II distribution can be defined as it follows,

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

for $y = 0, 1, 2, ...$, $\mu > 0$ and $\sigma > 0$ (Evans, 1953; Johnson, N. L.; Kotz, S.; Kemp, 1993).

2.3.1.4 Zero Inflation

In some cases the overdispersion can be caused by the occurence of excess zeros. If there is a much higher number of zeros than expected for the Poisson or Negative Binomial distributions, it is said that there is zero inflation (Zuur *et al.*, 2009). Not considering zero inflation can either lead to biased estimates of parameters and standard errors.

In view of the issues associated with zero inflation, the subject of interest concerns, initially, the process of generating zeros in sample data sets. Several authors have described zero-generating processes and proposed ways to characterize them such as in bird abundance contexts (Kuhnert *et al.*, 2005; Martin *et al.*, 2005), population dynamics of marine fauna (Hemmingsen *et al.*, 2005) and also in areas outside of ecology.

Zeros-inflated data can be analyzed through the use of mixed models (Zero-inflated Poisson or "ZIP" and Zero-inflated negative binomail or "ZINB").

2.3.1.5 Zero Inflated Poisson model (ZIP)

The zero inflated Poisson (ZIP), with $Y \sim \text{ZIP}(\mu, \sigma)$, is a discrete mixture of a component of value 0 with probability σ and a Poisson distribution with mean μ and probability $1 - \sigma$, where σ is the associated parameted of the zero inflation overdispersion.

The probability density function is given as

$$\begin{aligned} P(Y = 0|\mu, \sigma) &= \sigma + (1 - \sigma)e^{-\mu}, \text{ and} \\ P(Y = y|\mu, \sigma) &= (1 - \sigma)\frac{\mu^y}{y!}e^{-\mu}, \text{ if } \mathbf{y} > 0, \end{aligned}$$

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

The mean and the variance are, respectively $E[Y] = (1 - \sigma)\mu$ e $V[Y] = (1 - \sigma)\mu + \sigma(1 - \sigma)\mu^2$.

2.3.1.6 Zero Inflated Negative Binomial model (ZINB)

The zero-inflated negative binomial distribution (ZINB) with parameters μ, σ, ν, corresponds to the mixture of 0 with probability ν , where ν is the non-struturaded zero probability (Zuur *et al.*, 2009), and the negative binomial distribution $\text{NB}(\mu, \sigma)$ with probability $(1 - \nu)$. The probability density function is given as

$$\begin{aligned} P(Y = 0|\mu, \sigma, \nu) &= \nu + (1 - \nu)P(Y_1 = 0|\mu, \sigma) \\ P(Y = y|\mu, \sigma, \nu) &= (1 - \nu)P(Y_1 = y|\mu, \sigma), \text{ if } y > 0, \end{aligned}$$

for $Y_1 \sim \text{NB}(\mu, \sigma)$, so,

$$P(Y_1 = 0|\mu, \sigma) = (1 + \sigma\mu)^{-\frac{1}{\sigma}}$$

and

$$P(Y_1 = y|\mu, \sigma) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left(\frac{\sigma\mu}{1 + \sigma\mu}\right)^y \left(\frac{1}{1 + \sigma\mu}\right)^{1/\sigma}$$

where $y_1 = 0, 1, 2, 3, ...,\ \mu, \sigma > 0$ and $0 < \nu < 1$. The mean and the variance are, respectively, $E[Y] = (1 - \nu)\mu$ and $V[Y] = (1 - \nu)\mu[1 + (\sigma + \nu)\mu]$.

2.3.2 Introduction to generalized additive mixed models (GAMs)

In cases where there is some structure of correlation between observations, such as the case of longitudinal data, additional functions can be included to improve the fit, as is the case of Generalized Additive Models.

Generalized additive models defined by (Hastie and Tibshirani, 1986), can be compared with the generalized linear models with the difference related to the linear predictors that involve a sum of smoothing functions of explanatory variables. The generalized additive model has the linear predictor described as

$$\eta_i = \mathbf{x}_i^T\theta + f_1(x_{1i}) + f_2(x_{2i}) + \ldots,$$

where $\mathbf{x}_i^T$ is the $i - th$ design matrix row $(i = 1, 2, \ldots, n)$, $\theta = \theta_1, \theta_2, \ldots$ corresponds to the associated parameter vector to the $\mathbf{x}_i^T$ matrix and $f_j(\cdot)$ are the j smoothing functions of the covariates x_i that can assume different bases (i.e. cubic splines, regression splines, B-splines). Such models can be specified in terms of smoothing functions making them flexible, and theoretically more complex, since there is a need for the representation of the smoothing function and choice of degree of smoothing (Wood, 2006).

Once they are semi-parametric regression models, GAMs add a greater flexibility in modeling once they don't require assuming much about the structure and behavior of the data and are also used to model erratic behavior observed in data.

2.3.2.1 Smoothing methods using splines

The smoothing function $f_j(x)$ can be a spline of a particular type, such as smoothing splines and natural splines.

For smoothing splines , we introduce a penalty term when estimating $f_j(x)$, in order to balance over-fitting and smoothness. For a simple model with normal errors, $f_j(x)$ can be estimated by minimising the objective function $S(f)$, subject to the penalty $\lambda \geq 0$

$$S(f) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx, \tag{2.2}$$

where $f(x)$ is generated from a spline basis with λ being the smoothing parameter.

2.3.2.2 Cubic splines

Let $x_1, \ldots, x_n$ belong to the interval $[a, b]$ where $a < x_1 < \ldots < x_n < b$. The function $f(x)$ defined in the $[a, b]$ interval is a cubic spline (Green and Silverman, 1994) only if :

- in each interval $(a, x_1), \ldots, (x_n, b)$, $f(x)$ is a third-degree polynomial,
- the polynomial segments to each x_i have continuous first and second derivatives.

So, a cubic spline, under these two principles, can be described as

$$f(x) = d_i(x - x_i)^3 + c_i(x - x_i)^2 + b_i(x - x_i) + a_i \text{ for } x_i \leq x \leq x_{i+1} \tag{2.3}$$

with a_i, b_i, c_i, d_i, $i = 0, \ldots, n$ coeficients of the third-degree polynomial function and we define $i_0 = a$ and $i_{n+1} = b$.

The goodness of fit can be evaluated using the generalized Akaike information criterion (GAIC) (Stasinopoulos and Rigby, 2007) defined as follows:

$$\text{GAIC(k)} \ = \hat{D} + k.df \tag{2.4}$$

where $\hat{D} = -2\hat{l}_d$, is the global fitted deviance $\hat{l}_d$ is the fitted log-likelihood, $k = 2(1+1/df)[1-(df+2)/n]^{-1}$ is the model penalty of the corrected AIC_c model (Clifford M . Hurvich and Tsai, 1998) with df (effective) degrees of freedom of the likelihood function for a fitted model. The selected model is the one wich presents the minimum AIC value.

It is also possible to use the worm plot as a general diagnostic tool for analysis of residuals, visualizing differences between two conditional distributions (Van Buuren and Fredriks, 2001). The worm plot, GAIC and also fitting a GAM, can be done using the package `gamlss()` (Stasinopoulos and Rigby, 2007) present in R software.

2.4 Case study analysis

2.4.1 Fitting Poisson models with polynomial functions

To begin the process of selecting models it is better to choose simpler models and later more complex models. Since we are working with count data, a first obvious choice is to fit Poisson models. When selecting the linear predictor that will be used, we first include all the main factors, the main parameters of experiment (Block, Spacing, Section and Week). As an initial alternative, we chose polynomial functions in an attempt to improve the fit. Later we use smoothing functions. Assuming $Y_{ijkt} \sim \text{PO}(\mu_{ijkt})$, $\log \mu_{ijkt} = \eta^{(1)}_{ijkt}$;, the first fitted models for wingless aphids and winged aphids can be described with the linear predictor as it follows

$$\eta^{(1)}_{ijkt} = \beta_0 + \text{Block}_i + \text{Spacing}_j + \text{Section}_k + \sum_{l=1}^{z} \beta_l (Week_t)^l, \tag{2.5}$$

where β_0 is the intercept, $i = 1, 2, 3, 4; j = 1.2, 3; k = 1, 2, 3; t = 1, \ldots, 16$ and the polynom of z-th degree with $z = 5$ for the counts of aphids.

According to the likelihood ratio test for nested models (Tables 2.1 and 2.2), the degree of polynomial selected for the models in question is the polynomial of the fifth degree, and with this we can follow the process with the evaluation of the next step, testing the interaction.

The graphical verification of the fit can also be performed with the help of the half-normal plot as shown in Figure 2.7, and it's possible to observe that increasing the polynomial degree does not lead to a better fit, neither for wingless nor for winged aphids.

Table 2.1. Likelihood ratio test for nested models to choose the polynomial degree of the Poisson model for winged aphids.

Polynomial degree	Likelihood ratio (LR)	d.f.	*p-value*
3rd Vs 4th	91.40	1	< 0.0001
4th Vs 5th	95.12	1	< 0.0001

Table 2.2. Likelihood ratio test for nested models to choose the polynomial degree of the model for wingless aphids.

Polynomial degree	Likelihood ratio (LR)	d.f.	*p-value*
3rd Vs 4th	6192.24	1	< 0.0001
4th Vs 5th	5157.87	1	< 0.0001

After selecting the degree of the polynomial function that most assists with the accommodation of temporal variation, it is interesting to verify the existence of an interaction between the fixed effects on our models. To do so, the next fitted models seek to evaluate the inclusion of the interaction between the Spacing and Section effects of the plants, and the linear predictor can now be described as

$$\eta^{(1)}_{ijkt} = \beta_0 + \text{Block}_i + \text{Spacing}_j + \text{Section}_k + (\text{Spacing} \times \text{Section})_{jk} + \sum_{l=1}^{z} \beta_l (Week_t)^l. \tag{2.6}$$

For both winged aphids and wingless aphids the interaction was not significant, therefore, we can continue to find a model that better fits the data once there is overdispersion generated by the excess of zeros.

The next step is to establish a model that considers the zero inflation, and since we are following the order starting from a Poisson distribution, the only change in the model will be the definition of a ZIP (Zero inflated Poisson), $Y_{ijkt} \sim \text{ZIP}(\mu_{ijkt}, \sigma)$, $\log \mu_{ijkt} = \eta^{(1)}_{ijkt}$;, model with constant inflation for the random part of the model as follows:

$$\eta^{(1)}_{ijkt} = \beta_0 + \text{Block}_i + \text{Spacing}_j + \text{Section}_k + \sum_{l=1}^{z} \beta_l (Week_t)^l, \tag{2.7}$$

where μ is the location parameter or mean and σ is the additional parameter of scale or variability.

The addition of the variability parameter to accommodate the overdispersion generated by the zero inflation is significant for winged aphids $LR = 258.98$ and wingless aphids $LR = 23191.49$, indicated by the improvement in the fit of the model.

Finally, for $Y_{ijkt} \sim \text{ZIP}(\mu_{ijkt}, \sigma_t)$, $\log \mu_{ijkt} = \eta^{(2)}_{ijkt}, \text{logit}(\sigma_t) = \gamma_0 + f(\text{Week}_t)$;, we need to evaluated if the fit will improve with a regression for the weekly time effect by including it in the part of the zero-inflation model as described by linear predictor

$$\eta^{(1)}_{ijkt} = \beta_0 + \text{Block}_i + \text{Spacing}_j + \text{Section}_k + \sum_{l=1}^{z} \beta_l (Week_t)^l, \tag{2.8}$$

This is the chosen model (winged aphids $LR = 17.01$ and wingless aphids $LR = 432.20$) The regression model presents better fit and the comparison of the fit between the different models can be observed with the half normal plots Figure 2.8 for winged and 2.9 for wingless.

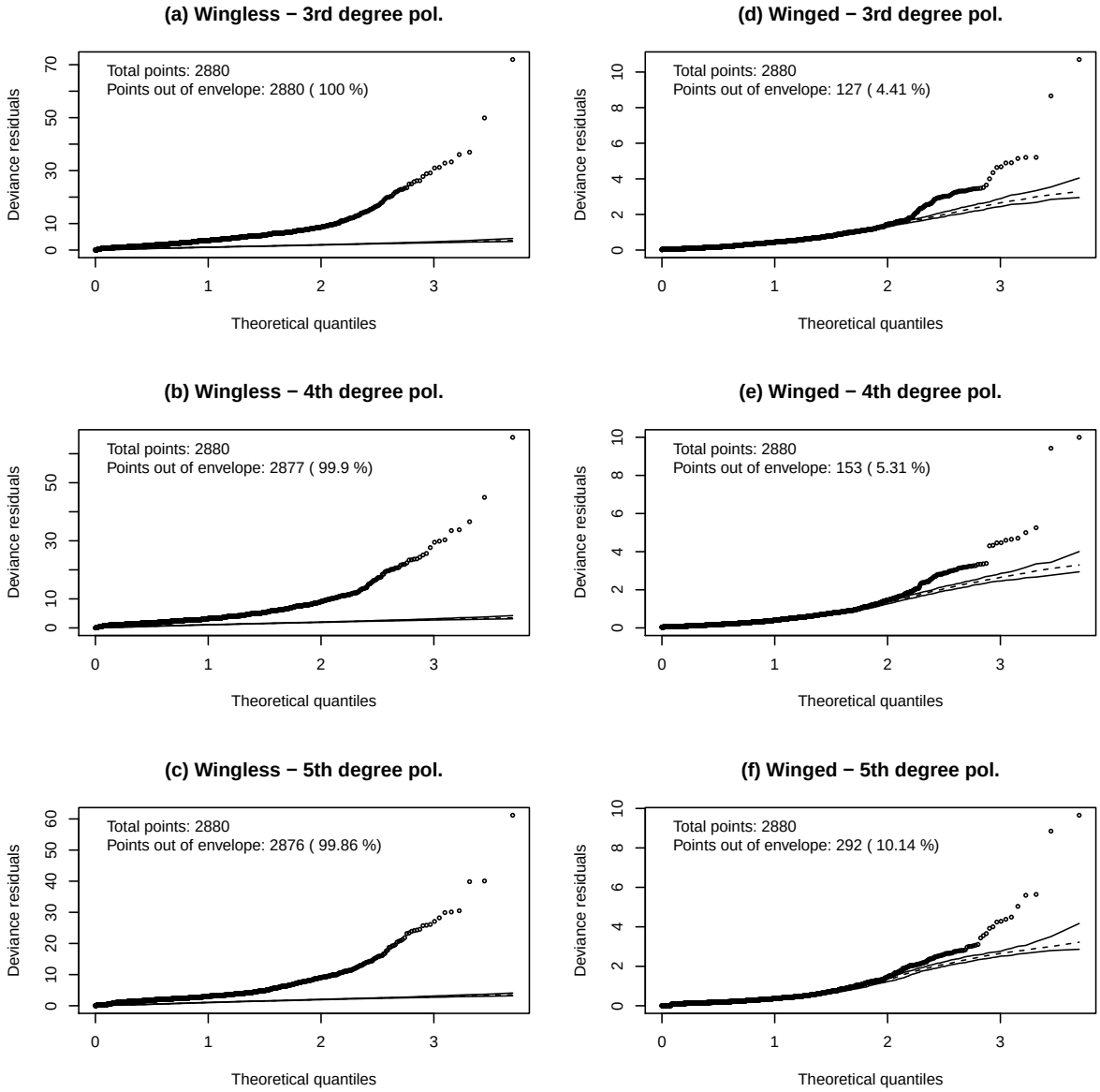

Figure 2.7. Half-normal plot for the different degrees of polynomial functions for winged and wingless aphids.

With the selected model, finally we are able to include the covariates. One by one, we test inclusion of a natural enemy by comparing the goodness of fit with the model 2.8.

For all significant values of the inclusion of the parameters of natural enemies we have to, in practical terms, they influence the population dynamics of aphids. If a natural enemy is significant for improving the fit of the model it means that, in nature, it is responsible for controlling the decrease in the number of pest population, in this case, responsible for controlling the aphid population.

Now, with the selected model for a Poisson distribution, the same process is made for the Negative binomial distribution.

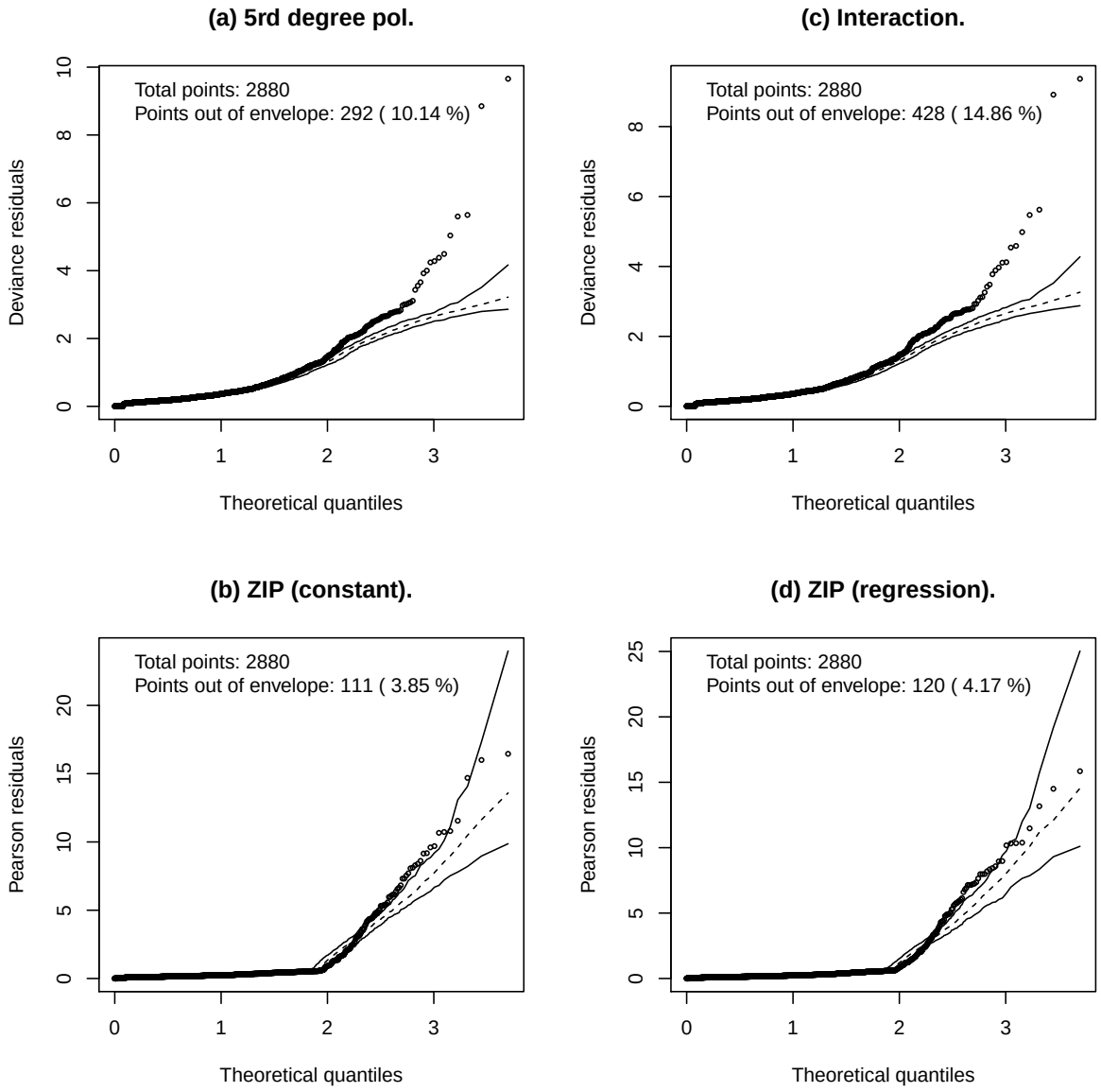

Figure 2.8. Half-normal plot to compare the different models for winged aphids and 5-th degree of the polynom.

Table 2.3. Likelihood ratio test for nested models to evaluate the inclusion of natural enemies for wingless aphids with the ZIP 5-th degree polynomial model.

Natural enemie	**Likelihood ratio (LR)**	d.f.	***p-value***
Lysipheblus	4.67	1	0.0307
Chrysopidae	1.04	1	0.3077
Scymnus	4.25	1	0.0394
Cycloneda	1.32	1	0.2505
Syrphidae	1.19	1	0.2746
Spiders	35.92	1	< 0.0001

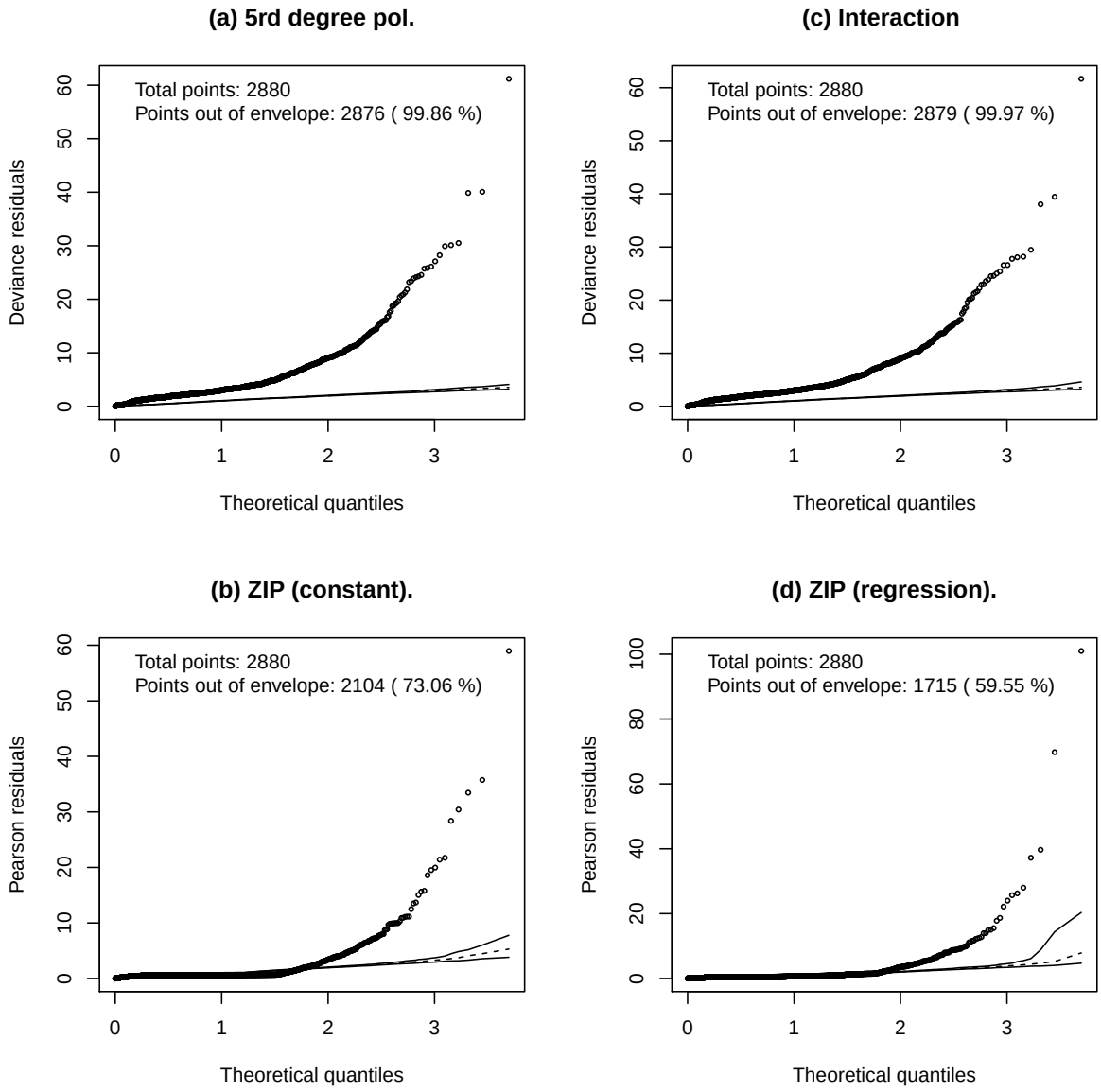

Figure 2.9. Half-normal plot to compare the different models for wingless aphids (a) 5-th degree of the polynom, (b) interaction, (c) ZIP (constant) and ZIP (regression).

Table 2.4. Likelihood ratio test for nested models to evaluate the inclusion of natural enemies for wingled aphidswith the ZIP 5-th degree polynomial model.

Natural enemie	**Likelihood ratio (LR)**	d.f.	***p-value***
Lysipheblus	5.53	1	0.0187
Chrysopidae	0.40	1	0.5291
Scymnus	6.23	1	0.0126
Cycloneda	6.90	1	0.0086
Syrphidae	19.58	1	< 0.0001
Spiders	4.59	1	0.0321

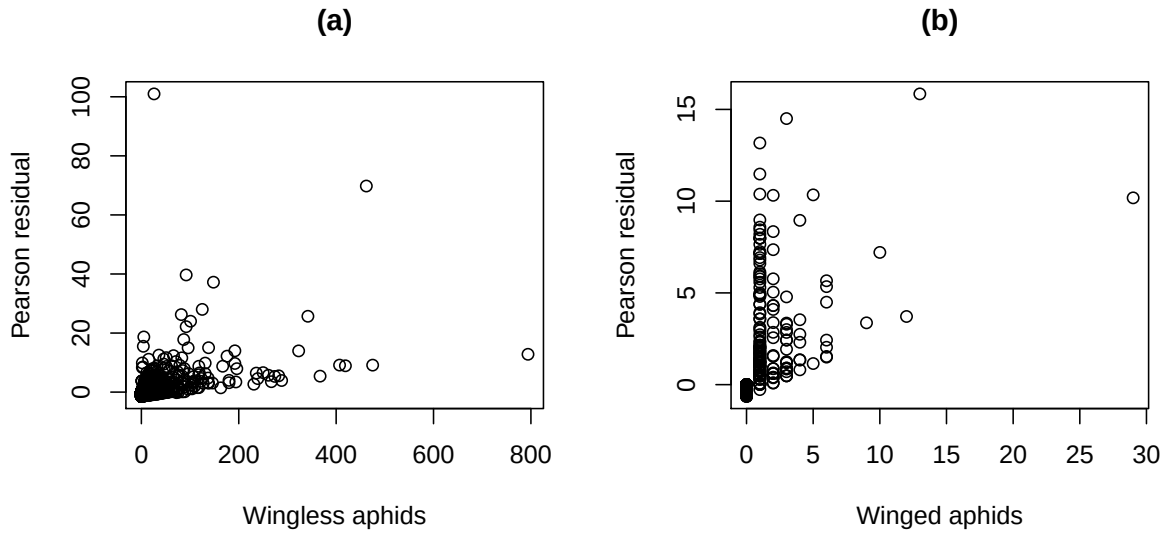

Figure 2.10. Residuals versus dependent variables for wingless aphids (a) and winged aphids (b).

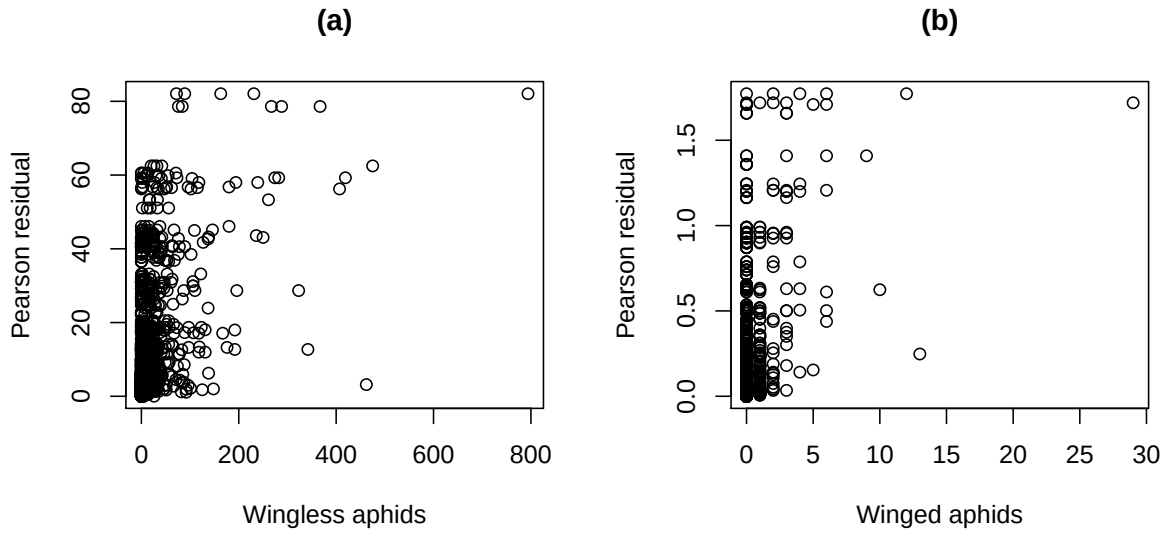

Figure 2.11. Residuals versus predicted variables for wingless aphids (a) and winged aphids (b).

2.4.2 Fitting negative binomial models

Assuming $Y_{ijkt} \sim \text{NB}(\mu_{ijkt}, \sigma)$, $\log \mu_{ijkt} = \eta^{(1)}_{ijkt}$; the sequence of fitted models for wingless aphids and winged aphids can be first described by the linear pretictors described bellow

$$\eta^{(1)}_{ijkt} = \beta_0 + \text{Block}_i + \text{Spacing}_j + \text{Section}_k + \sum_{l=1}^{z} \beta_l (Week_t)^l, \tag{2.9}$$

where β_0 is the intercept, $i = 1, 2, 3, 4; j = 1.2, 3; k = 1, 2, 3; t = 1, \ldots, 16$ and the choosen

polynom of z-th degree with $z = 5$ for the counts of aphids, μ is the location parameter or mean and σ is the additional parameter of scale or variability for the zero inflated negative binomial distribution.

Then, we test $Y_{ijkt} \sim \text{NB}(\mu_{ijkt}, \sigma)$, $\log \mu_{ijkt} = \eta^{(2)}_{ijkt}$;, assuming the second linear predictor

$$\eta^{(2)}_{ijkt} = \beta_0 + \text{Block}_i + \text{Spacing}_j + \text{Section}_k + (\text{Spacing} \times \text{Section})_{jk} + \sum_{l=1}^{z} \beta_l (Week_t)^l. \tag{2.10}$$

The third model $Y_{ijkt} \sim \text{ZINB}(\mu_{ijkt}, \sigma_t, \nu)$, $\log \mu_{ijkt} = \eta^{(2)}_{ijkt}, \log \sigma_t = \delta_0 + f(\text{Week}_t)$; is fit for a constant variation over time. The last model here includes a regression for the time effect described as $Y_{ijkt} \sim \text{ZINB}(\mu_{ijkt}, \sigma_t, \nu_t)$, $\log \mu_{ijkt} = \eta^{(2)}_{ijkt}, \log \sigma_t = \delta_0 + f(\text{Week}_t)$,
$\text{logit}(\nu_t) = \gamma_0 + f(\text{Week}_t)$.

It is possible to see the difference between all the models by using the half-normal plot (Figure 2.12 and 2.13).

Using the likehood ratio we can choose the ZINB with 5-th degree polynomial and, after the residual analysis, test the goodness of fit including the covariates Tables 2.5 and 2.6.

Table 2.5. Likelihood ratio test for nested models to evaluate the inclusion of natural enemies for winged aphids with the ZINB 5-th degree polynomial model.

Natural enemies	**Likelihood ratio (LR)**	d.f.	***p-value***
Lysipheblus	14.33	3	0.0025
Chrysopidae	12.46	3	0.0059
Scymnus	15.15	3	0.0017
Cycloneda	9.91	3	0.0192
Syrphidae	26.81	3	< 0.0001
Spiders	10.22	3	0.0167

Table 2.6. Likelihood ratio test for nested models to evaluate the inclusion of natural enemies for wingless aphids with the ZINB 5-th degree polynomial model.

Natural enemie	**Likelihood ratio (LR)**	d.f.	***p-value***
Lysipheblus	67.62	3	< 0.0001
Chrysopidae	67.64	3	< 0.0001
Scymnus	70.38	3	< 0.0001
Cycloneda	68.44	3	< 0.0001
Syrphidae	67.40	3	< 0.0001
Spiders	67.41	3	< 0.0001

The same can be evaluated for the negative binomial model. Here, more swings of natural enemies are significant, leading to the consideration of better fit over the Poisson model.

2.4.3 Including splines in the linear predictor

As the main objective is to compare the fit of models that use polynomial function with that acommodate temporal variability by means of smoothing functions, the next step is repeat the whole previous model selection process replacing the polynomial function by splines. The first thing is to choose the effective degrees of freedom, nodes of the function, without the independent variables by using AIC, so we can define

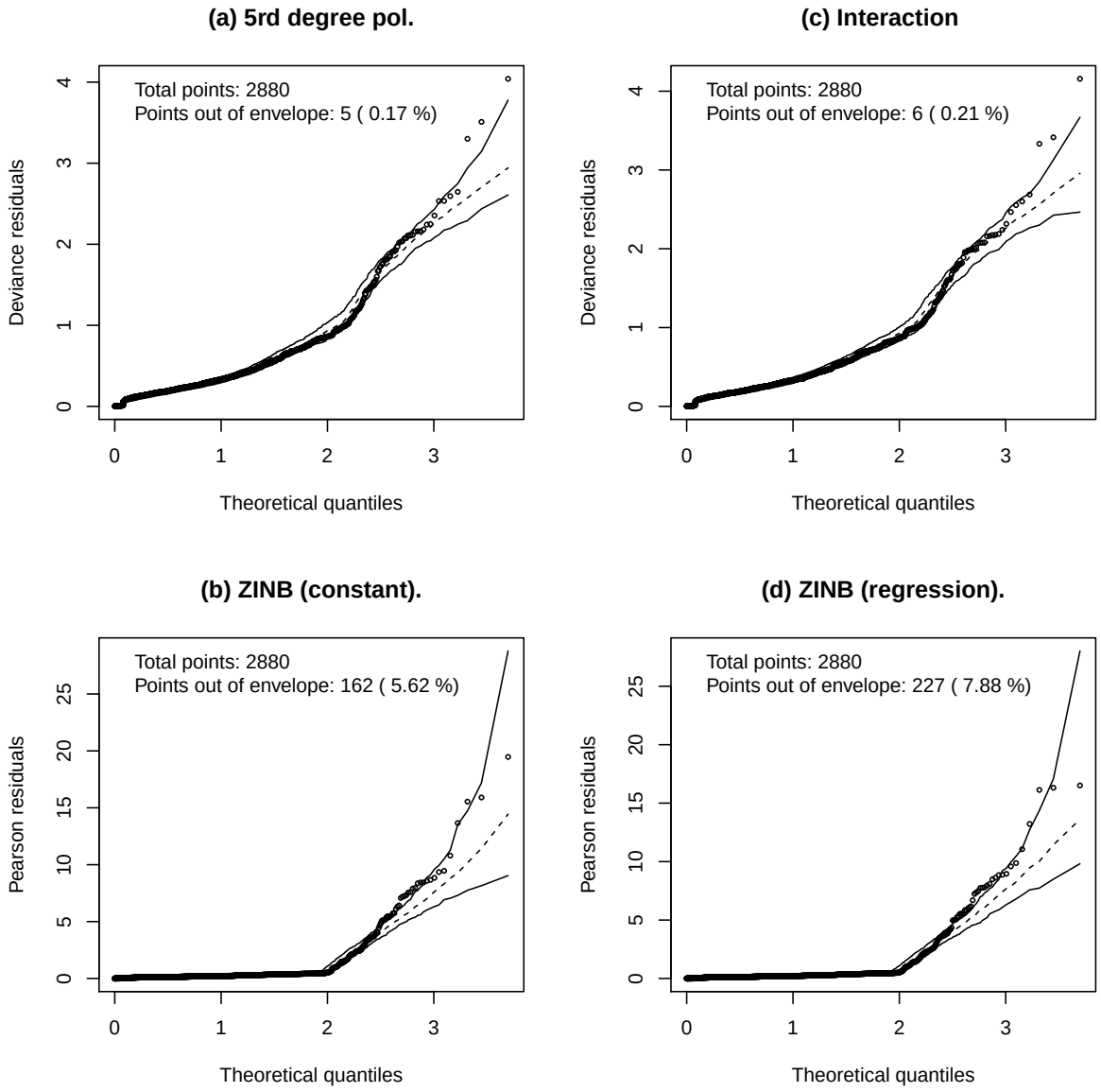

Figure 2.12. Half-normal plot to compare the different models for winged aphids and 5-th degree of the polynom.

$$\eta^{(3)}_{ijkt} = \beta_0 + \text{Block}_i + \text{Spacing}_j + \text{Section}_k + f(Week_t), \tag{2.11}$$

where β_0 is the intercept, $i = 1, 2, 3, 4; j = 1.2, 3; k = 1, 2, 3; t = 1, \ldots, 16$ and $f(.)$ is a smoothing function based on cubic splines with eight nodes over time. The models are described below according to the previous linear predictor:

P1 $Y_{ijkt} \sim \text{P}(\mu_{ijkt})$, $\log \mu_{ijkt} = \eta^{(3)}_{ijkt}$;

P2 $Y_{ijkt} \sim \text{P}(\mu_{ijkt})$, $\log \mu_{ijkt} = \eta^{(4)}_{ijkt}$;

P3 $Y_{ijkt} \sim \text{ZIP}(\mu_{ijkt}, \sigma)$, $\log \mu_{ijkt} = \eta^{(3)}_{ijkt}$;

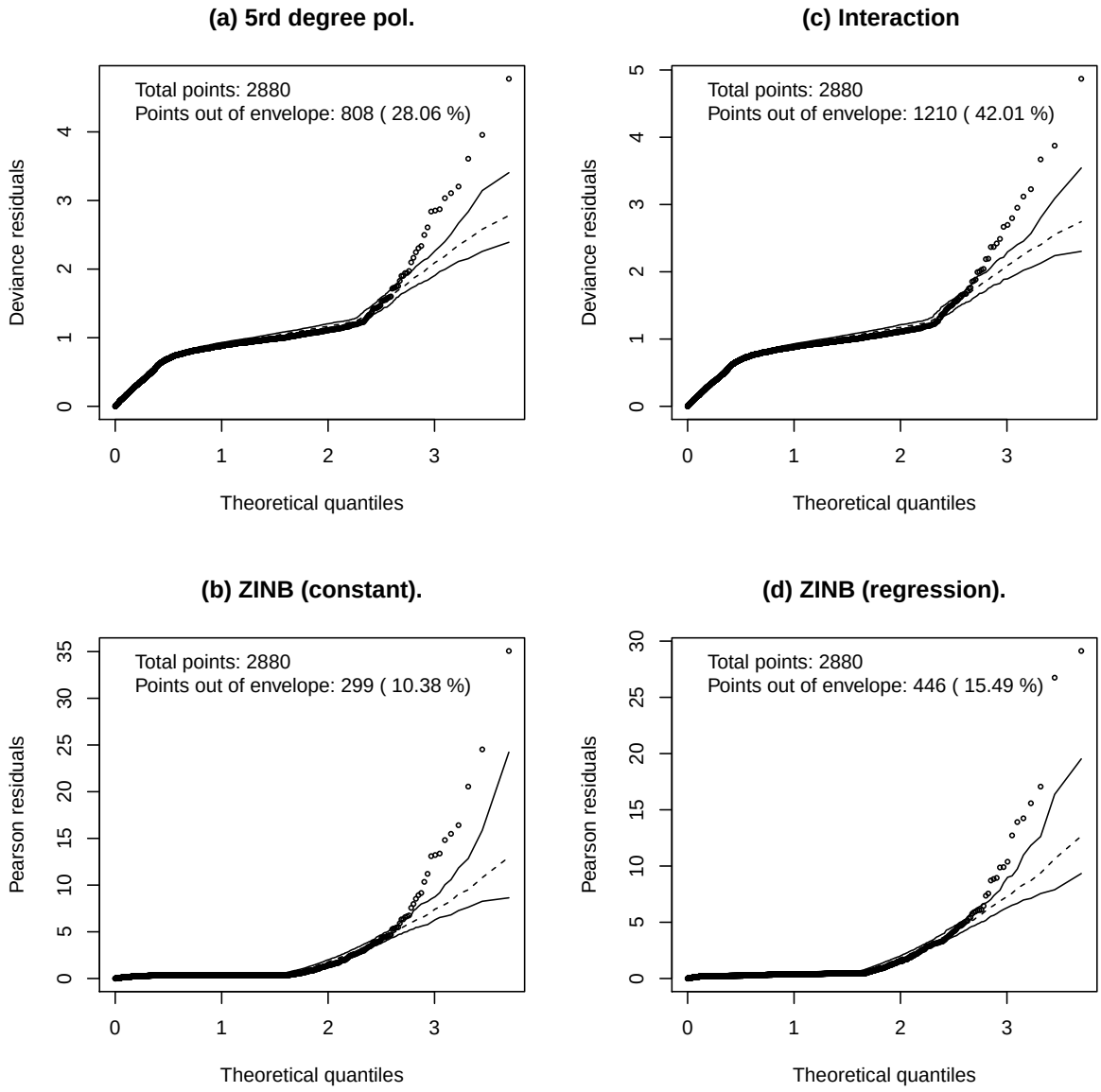

Figure 2.13. Half-normal plot to compare the different models for wingless aphids and 5-th degree of the polynom.

P4 $Y_{ijkt} \sim \text{ZIP}(\mu_{ijkt}, \sigma_t)$, $\log \mu_{ijkt} = \eta^{(3)}_{ijkt}, \text{logit}(\sigma_t) = \gamma_0 + f(\text{Week}_t)$;

NB1 $Y_{ijkt} \sim \text{NB}(\mu_{ijkt}, \sigma)$, $\log \mu_{ijkt} = \eta^{(3)}_{ijkt}$;

NB2 $Y_{ijkt} \sim \text{NB}(\mu_{ijkt}, \sigma)$, $\log \mu_{ijkt} = \eta^{(4)}_{ijkt}$;

NB3 $Y_{ijkt} \sim \text{ZINB}(\mu_{ijkt}, \sigma_t, \nu)$, $\log \mu_{ijkt} = \eta^{(3)}_{ijkt}, \log \sigma_t = \delta_0 + f(\text{Week}_t)$;

NB4 $Y_{ijkt} \sim \text{ZINB}(\mu_{ijkt}, \sigma_t, \nu_t)$, $\log \mu_{ijkt} = \eta^{(3)}_{ijkt}, \log \sigma_t = \delta_0 + f(\text{Week}_t)$,
$\text{logit}(\nu_t) = \gamma_0 + f(\text{Week}_t)$.

Testing the interaction Spacing/Section using AIC we select for wingless aphids the model including interaction and winthout interaction for winged aphids (equation 2.12).

$$\eta^{(4)}_{ijkt} = \beta_0 + \text{Block}_i + \text{Spacing}_j + \text{Section}_k + (\text{Spacing} \times \text{Section})_{jk} + f(Week_t), \tag{2.12}$$

According to the values of AIC for the following table, we can choose the best model.

Table 2.7. Values of AIC for model selection using splines.

Wingless				
Model	d.f	AIC (k=2)	AIC (k=3.84)	AIC (k=7.97)
NB1	16.99863	62968.84	63000.11	63070.32
NB2	20.99863	62539.35	62577.99	62664.71
NB3	29.99909	42116.97	42172.16	42296.06
NB4	39.00027	**41578.97**	**41650.73**	**41811.80**
Winged				
Model	d.f	AIC (k=2)	AIC (k=3.84)	AIC (k=7.97)
NB1	17.00136	1877.488	1908.771	1978.986
NB2	21.00136	1881.630	1920.272	2007.008
NB3	29.99923	**1609.483**	**1664.682**	**1788.578**
NB4	38.99849	1614.651	1686.409	1847.472

The goodness-of-fit of the model can be observed with the worm plot Figure 2.14.

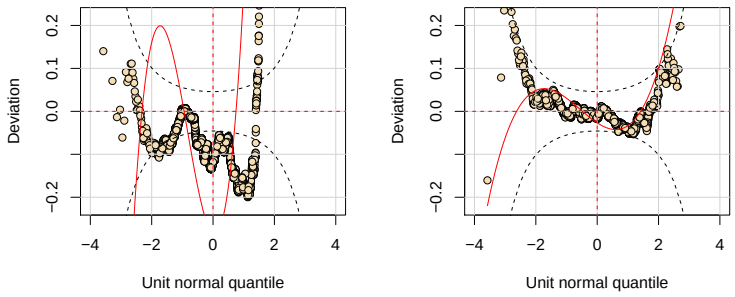

Figure 2.14. Wormplot of the selected models for wingless aphids and winged aphids respectively.

So, it is time to include the covariates, one by one, and compare with selected model in the previous step.

According to this, the model is better including natural enemies except Lysipheblus and Chrysopidae, for wingless aphids, and only Chrysopidae is not significative for winged aphids as seen in Table 2.8.

Now, the last distribution is the Negative binomial. We start fitting a simple model (NB1) with the linear predictor **??**, than a model with interaction (NB2), a zero inflated model with constant zero inflation (NB3) and a zero inflated model with smooothing funtion aplied to the zero inflation parameter (NB4). After comparing using AIC we can finaly choose the NB1 model with the linear predictor described in **??**.

The fitness of the model can be observed with the quantile plot Figure 2.15.

After comparing using AIC we can finaly choose the NB1 model as it shows Table 2.9 bellow.

Table 2.8. Values of AIC for model selection using splines including each natural enemy.

Wingless				
Model	d.f	AIC (k=2)	AIC (k=3.84)	AIC (k=7.97)
P4	39.00027	41578.97	41650.73	41811.80
P4 + Lysipheblus	40.00030	41584.43	41658.03	41823.23
P4 + Chrysopidae	40.00028	41580.59	41654.19	41819.39
P4 + Scymnus	40.00028	41573.86	41647.46	41812.66
P4 + Cycloneda	40.00209	41573.69	41647.30	41812.50
P4 + Syrphidae	40.00018	41571.44	41645.04	41810.24
P4 + Spiders	40.00028	41549.87	41623.47	41788.67
Winged				
Model	d.f	AIC (k=2)	AIC (k=3.84)	AIC (k=7.97)
P3	29.99923	1609.483	1664.682	1788.578
P3 + Lysipheblus	30.99924	1604.320	1661.358	1789.385
P3 + Chrysopidae	30.99923	1610.357	1667.396	1795.423
P3 + Scymnus	30.99922	1604.028	1661.066	1789.093
P3 + Cycloneda	30.99922	1605.511	1662.549	1790.576
P3 + Syrphidae	30.99917	1591.518	1648.557	1776.583
P3 + Spiders	30.99924	1607.030	1664.068	1792.095

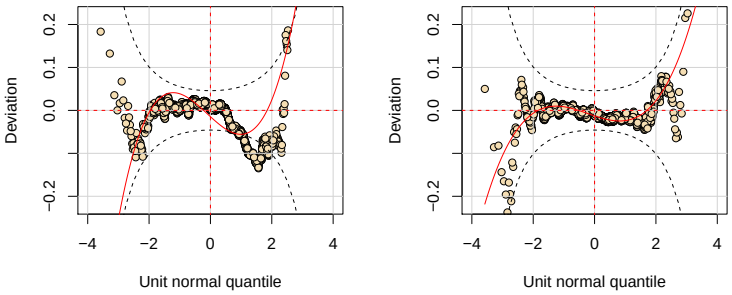

Figure 2.15. Wormplot of the selected models for wingless aphids and winged aphids respectively.

Table 2.9. Values of AIC for model selection using splines .

Wingless				
Model	d.f	AIC (k=2)	AIC (k=3.84)	AIC (k=7.97)
NB1	18.00116	10094.05	10127.17	10201.51
NB2	22.00116	10098.18	10138.67	10229.53
NB3	19.00081	33865.12	33900.08	33978.55
NB4	28.00196	33883.12	33934.65	34050.29
Winged				
Model	d.f	AIC (k=2)	AIC (k=3.84)	AIC (k=7.97)
NB1	17.99881	1511.401	1544.519	1618.854
NB2	21.99882	1517.397	1557.875	1648.730
NB3	18.99899	1512.030	1546.988	1625.454
NB4	27.99829	1518.511	1570.028	1685.661

And the same way as for the Poisson, we now add the covariates and compare each one with

the model selected before.

Table 2.10. Values of AIC for model selection using splines including each natural enemy.

Wingless				
Model	d.f	AIC (k=2)	AIC (k=3.84)	AIC (k=7.97)
NB1	18.00116	10094.05	10127.17	10201.51
NB1 + Lysipheblus	27.99711	9795.604	9847.118	9962.746
NB1 + Chrysopidae	27.99711	9795.752	9847.266	9962.895
NB1 + Scymnus	27.99711	9792.820	9844.335	9959.963
NB1 + Cycloneda	27.99711	9796.094	9847.609	9963.237
NB1 + Syrphidae	27.99711	9795.875	9847.390	9963.018
NB1 + Spiders	27.99711	9796.263	9847.778	9963.406
Winged				
Model	d.f	AIC (k=2)	AIC (k=3.84)	AIC (k=7.97)
NB1	17.99881	1511.401	1544.519	1618.854
NB1 + Lysipheblus	18.99855	1504.550	1539.507	1617.971
NB1 + Chrysopidae	18.99881	1511.825	1546.783	1625.248
NB1 + Scymnus	18.99884	1506.397	1541.355	1619.820
NB1 + Cycloneda	18.99882	1511.732	1546.690	1625.155
NB1 + Syrphidae	18.99887	1492.828	1527.786	1606.251
NB1 + Spiders	18.99882	1510.848	1545.806	1624.271

Here, Table 2.10, we can see that all the covariates are significant for wingless aphids but only Lysipheblus and Syrphidae are significant for winged aphids.

2.5 Discussion

The aim of this paper was to propose a tutorial, on how to fit and to assess goodness-of-fit for a range of different models for overdispersed and zero inflated count data. We described and showed, step by step, how to fit and interpret the simplest, Poisson and Negative Binomial models, as well as the more complex model. More than this, it was important to see the diference between using different function to accomodate variabilty along the time. Considering the selection between the two types of models proposed so far (polynomial and splines), one can conclude that the model that includes the smoothing function in the explanatory variable time, is the one that best describes the behavior of the response variable. By incorporating variability associated with the excess of zeros and, also, the great variation between observations related to covariables, or the negative inflated Binomial model of zeros with cubic spline fits better. Thus, such models, are the most indicated, since it captures more sutile biological interactions.

3 SIMULATION STUDY FOR COUNT DATA USING GAMS

Abstract

Count data are commonly observed in several areas of knowledge, whether by numbers of individuals or proportions. Two common features of counting studies are overdispersion and zero-inflation, characteristics treated using zero-inflated counting models. However, modeling the data variability and zero-inflation alone may not be enough for a good fit when dealing with longitudinal counting data. In such cases, the use of models including smoothing functions are appropriate. In this simulation study, we compared the efficiency of zero-inflated models with the simulation of three different scenarios to test the inclusion of smoothing functions using cubic splines. For these simulated data with temporal effect, the inclusion of splines did not improve the googdness fit.

Keywords: Count data; simulation study; generalized additive models; zero-inflated models; smothing functions.

3.1 Introduction

It is common in several fields of science to observe counting data, which generally have great variability and large occurrence of zeros. Whether in evolutionary biology or ecology (Harrison, 2014), in genetics and medicine (Plagnol *et al.*, 2012) or psychology (Loeys *et al.*, 2012), count data often presents overdispersion or zero-inflation that cannot be optimally modelled with normal distribution. Normally in these cases, the relationship between observations and explanatory variables are performed using models involving Poisson or negative binomial distributions. Once the Poisson model assumes equality of the mean and variance, negative binomial models have greater flexibility modeling the relationships between the mean and the variance (Colin and Pravin, 2013). Although they are present in many packages of statistical software, negative binomial models are limited to modeling over-dispersed data, and unable to deal with under-dispersed data (McCullagh and Nelder, 1989).

Zero inflated models are alternatives when there is great variability related with the zero occurences. Zero inflated Poisson models (ZIP) (Lambert, 1992) and Zero inflated negative binomial models (ZINB) (Greene, 1994) are used to deal with these characteristics. Although used in many areas such as psychology (Atkins and Gallop, 2007), computer science (Souza *et al.*, 2016) and several other applied fields (Farhadi Hassankiadeh *et al.*, 2018; Njambi Wanjau, 2019), if the non-zero part of data is over-dispersed, the parameter estimates of a ZIP can be biased just as standard errors may be underestimated. Zero-inflated models are also used in the analisys of longitudinal data (Neelon *et al.*, 2010; Rose *et al.*, 2006). Eventually, models for count data are not as flexible because they assume the linearity of covariates on the log-transformed expectation. To deal with this problem, spline based approaches became quite common. Introduced by Hastie and Robert (1990), the Generalized Additive Models (GAM) tend to be more flexible, using smoothing functions to accommodate the extra variability. Smoothing splines are one of the most popular methods for prediction of the nonparametric regression, estimating a nonparametric function that minimizes the penalized least squares criterion (Aydin *et al.*, 2013).

The aim of this work is to evaluate the fit improvement of zero-inflated models with the inclusion of smoothing functions under different variability and proportion of zeros for simulated data. Changing dispersion parameters and proportion of zeros, we compare Poisson, Negative binomial, ZIP and ZINB models, with and without a nonparametric regression for simulated data.

3.2 Count data models

Count data models are included in the class of Generalized linear models (GLM) (Nelder and Wedderburn, 1972) that deals with a larger distributions family, the exponecial family, making it possible to analyze datasets with different characteristics (continuous and discrete). The GLMs are defined by three distinct components:

- A random component, corresponding to the random variables $Y_1, \ldots, Y_n$ that belong to the exponential family of distributions each in terms of a distinct parameter θ_i.
- The linear predictor related, to the explanatory variables defined as $\eta = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}$, where $\boldsymbol{\beta}$ is a vector of p unknown parameters and $\mathbf{x} = [x_1, \ldots, x_n]'$ is the i-th column of the $n \times p$ design matrix.
- A link function $g(\mu_i) = \eta_i$, relating the systematic and the random component (Hinde and Demétrio, 1998)

All counting models used are members of the exponential distributions family and will be described in the following sections.

3.2.1 Poisson and Negative binomial models

The basic model used for count data is the Poisson model. The Poisson probability density function (pdf) is

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \quad \text{for} \quad y = 0, 1, 2, ...$$

where the expected value of Y is μ.i.e., $E(y) = \mu$ and $Var(y) = \mu$, and link function defined as $g(\mu) = \log(X\beta)$ for β vector of the estimated model coefficients. Because count data are often overdispersed, the Poisson model is generally considered inappropriate (Colin and Pravin, 2013).

Since inherent characteristics to the sampling processes can generate overdispersed data, the Poisson distribuiton is not a good option anymore, so the negative binomial distribution is an alternative (Hinde and Demétrio, 1998). The probability function of Negative Binomial type-I distribution $Y \sim \text{NB}(\mu, \phi)$ can be written as

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

for $y = 0, 1, 2, ...$, $\mu > 0$ and $\sigma > 0$ (Anscombe, 1949).

The Negative Binomial type-II distribution can be defined as it follows

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

for $y = 0, 1, 2, ...$, $\mu > 0$ and $\sigma > 0$ (Evans, 1953; Johnson, N. L.; Kotz, S.; Kemp, 1993).

3.2.2 Zero-inflated models: ZIP and ZINB

Since the number of zeros can be impressive in model fit, the need for zero-inflated models is even greater to accommodate the high variability not associated with the experimental structure.

The zero-inflated Poisson (Lambert, 1992) and zero-inflated negative binomial (Stasinopoulos and Rigby, 2007) are alternatives to increase the goodness of fit.

The probability density function of ZIP model is given as

$$\begin{aligned} P(Y=0|\mu,\sigma) &= \sigma + (1-\sigma)e^{-\mu}, \text{ and} \\ P(Y=y|\mu,\sigma) &= (1-\sigma)\frac{\mu^y}{y!}e^{-\mu}, \text{ if } \mathbf{y} > 0, \end{aligned}$$

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

The mean and the variance are, respectively $E[Y] = (1-\sigma)\mu$ e $V[Y] = (1-\sigma)\mu + \sigma(1-\sigma)\mu^2$.

The probability density function of a ZINB model is given as

$$\begin{aligned} P(Y=0|\mu,\sigma,\nu) &= \nu + (1-\nu)P(Y_1=0|\mu,\sigma) \\ P(Y=y|\mu,\sigma,\nu) &= (1-\nu)P(Y_1=y|\mu,\sigma), \text{ if } y > 0, \end{aligned}$$

for $Y_1 \sim \text{NB}(\mu,\sigma)$, so,

$$P(Y_1=0|\mu,\sigma) = (1+\sigma\mu)^{-\frac{1}{\sigma}}$$

and

$$P(Y_1=y|\mu,\sigma) = \frac{\Gamma(y+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}$$

where $y_1 = 0,1,2,3,...,\ \mu,\sigma > 0$ and $0 < \nu < 1$. The mean and the variance are, respectively, $E[Y] = (1-\nu)\mu$ and $V[Y] = (1-\nu)\mu[1+(\sigma+\nu)\mu]$.

All these distributions can be applied through the theory of generalized additive models (GAM) (Hastie and Tibshirani, 1986). Such models can be specified in terms of smoothing functions making them flexible, and theoretically more complex, since there is a need for the representation of the smoothing function and choice of degree of smoothing (Wood, 2006).

It is possible to use a semi-parametric function to acommodate variability and here we will use cubic splines (Green and Silverman, 1994), implemented in the package `gamlss()` (Stasinopoulos and Rigby, 2007) present in R software. Let $x_1,\ldots,x_n$ belong to the interval $[a,b]$ where $a < x_1 < \ldots < x_n < b$. The function $f(x)$ defined in the $[a,b]$ interval is a cubic spline only if :

- in each interval $(a,x_1),\ldots,(x_n,b)$, $f(x)$ is a third-degree polynomial,
- the polynomial segments to each x_i have continuous first and second derivatives.

So, a cubic spline, under these two principles, can be described as

$$f(x) = d_i(x-x_i)^3 + c_i(x-x_i)^2 + b_i(x-x_i) + a_i \text{ for } x_i \le x \le x_{i+1} \tag{3.1}$$

with a_i, b_i, c_i, d_i, $i = 0,\ldots,n$ coeficients of the third-degree polynomial function and we define $i_0 = a$ and $i_{n+1} = b$.

3.3 Simulation design

To test the effects of variability and proportions of zeros in the Poisson, Negative binomial (NB), Zero inflated Poisson (ZIP) and Zero inflated negative binomial (ZINB) models, different datasets were simulated. In an attempt to evaluate counting data models in different ways, and be able to compare with the spline application, three conditions were adopted to elaborate three scenarios of simulations with different parameter definitions. The first simulation was performed using a ZINB model with a defined mean $\mu = 1.0$, and the other parameters varied in two levels each, $\sigma = 0.5, 1$ and $\nu = 0.3, 0.7$, generating four combinations in which all models were fitted, with a total of 16 models tested, being M1 to M4 replicated in each of the four combinations of parameters. The second scenario was simulated using a ZIP model with a defined $\mu = 1.0$ and two different σ levels, $\sigma = 0.1, 0.3$ and then, all models were fitted, with a total of 8 models, M1 to M4 replicated in the two combinations.

The third one considered a nonlinear zero-inflated simulation. We initially defined two factors, time and individuals, whose main objective is to reproduce the relation of individual observations, factor $I = 1, \ldots, 20$, sampled over Time, factor $T = 1, \ldots, 50$. The simulated data considered $\mu = exp(\eta_a)$ for $\eta_a = \beta_0 + \beta_1 * T$, and $\sigma = P[\eta_b]$ where P is the probability of a logistic distribution and $\eta_b = \gamma_0 + \gamma_1 * \sin(2 * \pi / Period * Time) + \gamma_2 * \cos(2 * \pi / Period * Time)$, with $Period = 20$, $\beta_0 = -1$, $\beta_1 = 0.1$, $\gamma_0 = 0.2$, $\gamma_1 = 0.5$ and $\gamma_2 = -0.5$. Then seven models were fitted with the link-function $\eta_{ij}^{(1)} = Time$, four models without a function in the parameter σ and four with the function, for comparison in the sequence N1 to N7. The 31 models are described bellow considering α_μ the model intercept, α_σ and α_ν the constant σ, ν respectively:

- M1 $Y_i \sim \text{ZINB}(\mu, \sigma, \nu)$, $\log \mu = \alpha_\mu, \log \sigma = \alpha_\sigma, \text{logit}(\nu) = \alpha_\nu$
- M2 $Y_i \sim \text{ZIP}(\mu, \sigma)$, $\log \mu = \alpha_\mu, \text{logit}(\sigma) = \alpha_\sigma$;
- M3 $Y_i \sim \text{NB}(\mu, \sigma)$, $\log \mu = \alpha_\mu$;
- M4 $Y_i \sim \text{P}(\mu)$, $\log \mu = \alpha_\mu$;
- N1 $Y_{ij} \sim \text{ZINB}(\mu_{ij}, \sigma, \nu)$, $\log \mu_{ij} = \eta_{ij}^{(1)}$,$\log \sigma_t = \alpha_\sigma, \text{logit}(\nu) = \alpha_\nu$;
- N2 $Y_{ij} \sim \text{ZINB}(\mu_{ij}, \sigma_t, \nu)$, $\log \mu_{ij} = \eta_{ij}^{(1)}$, $\log \sigma_t = f(\text{Time}_t), \text{logit}(\nu) = \alpha_\nu$;
- N3 $Y_{ij} \sim \text{ZIP}(\mu_{ij}, \sigma)$, $\log \mu = \eta_{ij}^{(1)}, \text{logit}(\sigma_t) = \alpha_\sigma$;
- N4 $Y_{ij} \sim \text{ZIP}(\mu_{ij}, \sigma_t)$, $\log \mu = \eta_{ij}^{(1)}, \text{logit}(\sigma_t) = f(\text{Time}_t)$;
- N5 $Y_{ij} \sim \text{NB}(\mu_{ij}, \sigma)$, $\log \mu_{ij} = \eta_{ij}^{(1)}, \log \sigma = \alpha_\sigma$;
- N6 $Y_{ij} \sim \text{NB}(\mu_{ij}, \sigma_t)$, $\log \mu_{ij} = \eta_{ij}^{(1)}, \log \sigma_t = f(\text{Time}_t)$;
- N7 $Y_{ij} \sim \text{P}(\mu_{ij})$, $\log \mu_{ij} = \eta_{ij}^{(1)}$;

The three simulation scenarios were elaborated considering the same execution pattern that follows the steps described below:

- sample size defined as $n = 1000$ observation,
- number of simulated data sets $m = 1000$,

- definition of the distribution parameters as shown in the Table 3.1, where we see four diferent combinations of parameters for the ZINB scenario (4×1000 datasets), two combinations for the Poisson scenario (2×1000 datasets) and one combination for the non-linear scenario (1000 datasets),
- fitting all models for each scenario (ZINB, ZIP, Negative binomial and Poisson),
- calculation the mean estimates of the parameters,
- calculation of the selection frequency of a specific model.

In practical terms, the model selection frequency calculation was done manually for each of the m = 1000 dataset with n = 1000 observations. For each combination of a specific simulation scenario, the Poisson, NB, ZIP, ZINB models were fitted and the value of AIC was extracted from each one. A table was created where in each column there were the AIC values of a specific model (eg column01 = Poisson AIC, column02 = NB AIC, ...) and, consequently, each row would contain the AIC values of all models fitted for the same dataset. Finally, for each line, the best model among the four proposed would be chosen, thus proposing the calculation of the selection frequencies for each of the parameter combinations within a given scenario.

Table 3.1. Parameters of the respective distributions used in the simulation scenarios.

Distribution	μ	σ	ν
ZINB (1)	1	0.5	0.3
ZINB (2)	1	0.5	0.7
ZINB (3)	1	1.0	0.3
ZINB (4)	1	1.0	0.7
ZIP (1)	1	0.1	-
ZIP (2)	1	0.3	-
Non-linear ZINB	$exp(\eta_a)$	$P[\eta_b]$	0.3

The selection frequency was made evaluating the goodness of fit of the four models, for each one of the $m = 1000$ data sets of a same simulation scenario, using the Akaike's information criterion (AIC), defined as follows:

$$\text{AIC} = -2\text{logL} + 2p \tag{3.2}$$

where logL is the maximum likelihood function for a fitted model and p is the number of parameters in this model. According to Burnham and Anderson (2004), the selected model is the one wich presents the minimum AIC value. All model and simulations were made using the package `gamlss()` (Stasinopoulos and Rigby, 2007) present in R software.

3.4 Results

The results of the proportion of selected models for each scenario are presented in Tables 3.2, 3.3, 3.4.

Table 3.2. Proportion of selected models for ZINB simulation with respective values of the parameters.

Model	$\sigma = 0.5, \nu = 0.3$	$\sigma = 0.5, \nu = 0.7$	$\sigma = 1.0, \nu = 0.3$	$\sigma = 1.0, \nu = 0.7$
ZINB	69.9%	44.8%	56.4%	48.0%
ZIP	2.5%	38.2%	-	10.6%
N.B.	27.6%	17.0%	43.6%	41.4%
PO	-	-	-	-

In Table 3.2, the model ZINB was the best for the four different simulations of the first scenario. Considering the four simulations, the only one which the ZIP model was selected more often than the negative binomial model was for $\sigma = 0.5, \nu = 0.7$ (38.2% of the datasets). Poisson models were not selected in any case, because of the absence of the parameters.

Almost the totality of the selected models were the zero-inflated negative binomial models, for the three different scenarios, ZINB(μ, σ, ν), ZIP(μ, σ), non-linear ZINB(μ, σ, ν), as defined in Table 3.1. For ZIP simulations with $\sigma = 0.3$ (Table 3.3), the best models were the zero-inflated Poisson (selected in 943 simulated datasets, 94.3% of the cases). The Poisson models were selected for 90 datasets in this scenario.

Special attention should be given to the non-linear simulation scenario (Table 3.4), where for all simulated datasets, the ZIP, N.B. and Poisson models were not selected in any case. Here is possible to observe that the zero-inflated negative binomial with the spline was selected for 429 datasets, agains 571 datasets where the inclusion of the cubic spline was not better.

Table 3.3. Proportion of selected models for ZIP simulation with respective values of the parameters.

Model	$\sigma = 0.1$	$\sigma = 0.3$
ZINB	50.3%	4.6%
ZIP	33.5%	94.3%
N.B.	8.2%	2.1%
PO	9.0%	-

Table 3.4. Proportion of selected models for ZINB simulation with respective values of the parameters.

Model	Proportion
ZINB	57.1%
ZINB(f)	42.9%
ZIP	-
ZIP(f)	-
N.B.	-
N.B.(f)	-
PO	-

Table 3.5. Mean estimates of the parameters for the ZINB simulation and respective values of the parameters.

ZINB $\sigma = 0.5, \nu = 0.3$			
Model	μ	σ	ν
ZINB	1.009	0.504	0.298
ZIP	1.319	0.886	-
N.B.	0.700	1.284	-
PO	0.700	-	-
ZINB $\sigma = 0.5, \nu = 0.7$			
Model	μ	σ	ν
ZINB	0.703	1.351	0.565
ZIP	1.319	3.408	-
N.B.	0.300	4.967	-
PO	0.300	-	-
ZINB $\sigma = 1.0, \nu = 0.3$			
Model	μ	σ	ν
ZINB	1.026	0.965	0.309
ZIP	1.590	1.280	-
N.B.	0.698	2.040	-
PO	0.698	-	-
ZINB $\sigma = 1.0, \nu = 0.7$			
Model	μ	σ	ν
ZINB	0.649	2.372	0.532
ZIP	1.584	4.310	-
N.B.	0.299	6.779	-
PO	0.299	-	-

In Table 3.5 it is possible to see the influence of the parameter, used to simulate the data, in the model estimation.

For $\nu = 0.3$, the ZINB models presents mean estimates similar to the parameter values used to simulate the data. Considering the first case with $\mu = 1.0, \sigma = 0.5,$ and $\nu = 0.3$, we have approximate mean estimates for the ZINB model with values of $\mu = 1.0091696, \sigma = 0.5046620, \nu = 0.2983536$.

For a large proportion of zeros $\nu = 0.7$, the mean estimates for the ZINB models underestimate μ and overestimate σ, while the ZIP models overestimate both. Considering for example the simulated parameter for the second case, $\mu = 1.0, \sigma = 0.5,$ and $\nu = 0.7$, the ZINB mean estimates are $\mu = 0.7038998, \sigma = 1.3510002, \nu = 0.56513276$ and the ZIP mean estimates $\mu = 1.319720, \sigma = 3.408092$.

Table 3.6. Mean estimates of the parameters for the ZIP simulation and respective values of the parameters.

ZIP $\sigma = 0.1$			
Model	μ	σ	ν
ZINB	1.007	2.288e+163	0.094
ZIP	1.003	0.113	-
N.B.	0.901	0.118	-
PO	0.901	-	-
ZIP $\sigma = 0.3$			
Model	μ	σ	ν
ZINB	0.966	6.513e+296	0.259
ZIP	0.989	0.416	-
N.B.	0.699	0.484	-
PO	0.699	-	-

Table 3.7. Mean estimates of the parameters for the non-linear ZINB simulation and respective values of the parameters.

Model	Intercept	μ	Intercept	σ	ν
ZINB	0.368	1.105	-	0.575	0.298
ZINB(f)	0.367	1.105	0.562	1.003	0.298
ZIP	0.448	1.100	-	0.549	-
ZIP(f)	0.472	1.099	0.353	1.010	-
N.B.	0.257	1.105	-	2.140	-
N.B.(f)	0.257	1.105	1.096	1.020	-
PO	0.266	1.104	-	-	-

The Table 3.6 when the greater the σ the smaller the μ estimates. The mean estimates of σ for the ZINB models presented extreme values due to the difference of parameterization between the ZIP distribution that considers a different variance function from that used by the ZINB model.

In Table 3.7 there is the comparison between models with and without the use of cubic splines for the ZINB, ZIP and N.B. models. The mean estimates of μ are all very similar considering a same distribution. According to the selected models using AIC, the inclusion of a cubic spline, which leads to the inclusion of the σ intercept, does not influence in the model fit.

3.5 Discussion

In this work it was considered to establish different scenarios of variability and zero-inflation to test the fit of parametric and semiparametric models. As expected, in the simulations where the associated variability was due to the variation of the parameters of the distributions used, the ZINB models presented not only higher frequency of selection but also better estimates of the parameters.

According to Feng and Zhu (2011), a combination of a model with a non-parametric part can improve model selection and change the non-linear effect on the covariates. However, although the nonparametric simulation scenario created variables with correlated observations, the inclusion of smoothing functions did not improve the fit.

It is clear that for a large proportion of zeros, models like ZIP end up associating the great variability by inflating σ estimates, since all inflation of zero is being estimated as variability generated by

the structural condition of the experiment (Hinde and Demétrio, 1998), as seen in the first simulation scenario.

When we have simulated data from a zero-inflated Poisson distribution, described in the second scenario, the proportion of zeros and the variability associated with the data structure are not composed of distinct parameters, so the difference of parameterization between ZINB and ZIP generate completely different estimates for the fitted models.

The use of the smoothing function for the third scenario of non-linear simulation, did not lead to a better fit for the used models in any of the cases. So, for the simulation created here, there was no difference using splines to fit the choosen models because there was no correlation according with the non-linear function used for the μ and σ.

REFERENCES

Anscombe, F. J., 1949 Wiley International Biometric Society. International Biometric Society **5**: 165–173.

Atkins, D. C., and R. J. Gallop, 2007 Rethinking How Family Researchers Model Infrequent Outcomes: A Tutorial on Count Regression and Zero-Inflated Models. Journal of Family Psychology **21**: 726–735.

Atkinson, A. C., 1985 *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford Statistical Science Series. Clarendon Press.

Aydin, D., M. Memmedli, and R. E. Omay, 2013 Smoothing Parameter Selection for Nonparametric Regression Using Smoothing Spline. European Journal of Pure and Applied Mathematics **6**: 222–238.

Battel, A. P. M., R. A. Moral, and W. A. Godoy, 2012 Modelos matemáticos predador-presa e aplicações ao manejo integrado de pragas. Oecologia Australis **16**: 43–62.

Burnham, K. P., and D. R. Anderson, 2004 Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods and Research **33**: 261–304.

Clifford M . Hurvich, J. S. . S., and C.-L. Tsai, 1998 Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. Wiley for the Royal Statistical Society **60**: 271–293.

Colin, C. A., and T. Pravin, 2013 *Regression analysis of count data, Second edition*. Cambridge University Press, Cambridge, 2nd ed. edition.

Demétrio, C. G. B., Hinde, J., & Moral, R. A., 2014 Models for overdispersed data. In W. G. E. C. Ferreira, editor, *Ecological modelling applied to entomology. Entomology in focus*. Springer International Publishing, New Delhi, pp. 219–259.

Evans, D. A., 1953 Experimental Evidence Concerning Contagious Distributions in Ecology. Biometrika **40**: 186–211.

Farhadi Hassankiadeh, R., A. Kazemnejad, M. Gholami Fesharaki, and S. Kargar Jahromi, 2018 Efficiency of Zero-Inflated Generalized Poisson Regression Model on Hospital Length of Stay Using Real Data and Simulation Study. Caspian Journal of Health Research **3**: 5–9.

Feng, J., and Z. Zhu, 2011 Semiparametric analysis of longitudinal zero-inflated count data. Journal of Multivariate Analysis **102**: 61–72.

Green, P. J., and B. W. Silverman, 1994 *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall.

Greene, W. H., 1994 Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Biology & Philosophy **9**: 265–265.

Harrison, X. A., 2014 Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ **2**: e616.

Hastie, T., and T. Robert, 1990 Generalized additive models 1 Introduction 2 Smoothing methods and generalized addi- tive models. Preventive Medicine .

Hastie, T., and R. Tibshirani, 1986 [Generalized Additive Models]: Rejoinder. Statist. Sci. **1**: 314–318.

Hemmingsen, W., P. A. Jansen, and K. MacKenzie, 2005 Crabs, leeches and trypanosomes: an unholy trinity? Marine Pollution Bulletin **50**: 336–339.

Hinde, J., and C. G. Demétrio, 1998 Overdispersion: Models and estimation.

Johnson, N. L.; Kotz, S.; Kemp, A. W., 1993 Univariate discrete distributions. Second Edition. Biometrical Journal : 565.

Kidd, D., and P. Amarasekare, 2012 The role of transient dynamics in biological pest control: Insights from a host-parasitoid community. Journal of Animal Ecology **81**: 47–57.

Kot, M., 2001 *Elements of Mathematical Ecology.* Elements of Mathematical Ecology. Cambridge University Press.

Kuhnert, P. M., T. G. Martin, K. Mengersen, and H. P. Possingham, 2005 Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. Environmetrics **16**: 717–747.

Lambert, D., 1992 Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics **34**: 1–14.

Liang, K. Y., and S. Zeger, 1986 Longitudinal data analysis using generalized linear models. Biometrika **73**: 13–22.

Loeys, T., B. Moerkerke, O. de Smet, and A. Buysse, 2012 The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. British Journal of Mathematical and Statistical Psychology **65**: 163–180.

Martin, T. G., B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, *et al.*, 2005 Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. Ecology Letters **8**: 1235–1246.

McCullagh, P., and J. A. Nelder, 1989 *Generalized Linear Models, Second Edition.* Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Moral, R. A., J. Hinde, and C. G. B. Demétrio, 2017 Half-Normal Plots and Overdispersed Models in R : The hnp Package . Journal of Statistical Software **81**.

Neelon, B. H., A. J. O'Malley, and S. L. T. Normand, 2010 A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. Statistical Modelling **10**: 421–439.

Nelder, J. A., and R. W. M. Wedderburn, 1972 Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General) **135**: 370–384.

Njambi Wanjau, A., 2019 Assessment and Selection of Competing Models for Count Data: An Application to Early Childhood Caries. International Journal of Data Science and Analysis **4**: 24.

Plagnol, V., J. Curtis, M. Epstein, K. Y. Mok, E. Stebbings, *et al.*, 2012 A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics **28**: 2747–2754.

Rose, C. E., S. W. Martin, K. A. Wannemuehler, and B. D. Plikaytis, 2006 On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. Journal of Biopharmaceutical Statistics **16**: 463–481.

Souza, R. M., F. J. Cysneiros, and R. A. Fagundes, 2016 Zero-inflated prediction model in software-fault data. IET Software **10**: 1–9.

Stasinopoulos, D. M., and R. A. Rigby, 2007 Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. Journal of Statistical Software **23**.

Van Buuren, S., and M. Fredriks, 2001 Worm plot: A simple diagnostic device for modelling growth reference curves. Statistics in Medicine **20**: 1259–1277.

Wood, S. N., 2006 ON CONFIDENCE INTERVALS FOR GENERALIZED ADDITIVE MODELS BASED ON PENALIZED REGRESSION SPLINES. Australian & New Zealand Journal of Statistics **48**: 445–464.

Zuur, A., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith, 2009 *Mixed Effects Models and Extensions in Ecology with R*. Statistics for Biology and Health. Springer New York.

APPENDIX

Appendix A

Computation script for chapter 2

Requiring the data

```
dados <- read.csv("contotal2013.csv", h=TRUE)
dados$Bloco <- as.factor(dados$Bloco)
dados$Espacamento <- as.factor(dados$Espacamento)
dados$Planta <- as.factor(dados$Planta)
dados$Seccao <- as.factor(dados$Seccao)
```

Choosing the polynomial degree - Poisson model - ZIP - (Analog for Negative binomial - ZINB)

```
modap01 <- glm(Apteros ~ poly(Semana, 3) + Bloco + Espacamento +
Seccao, family=poisson, data=dados)
modap02 <- glm(Apteros ~ poly(Semana, 4) + Bloco + Espacamento +
Seccao, family=poisson, data=dados)
modap03 <- glm(Apteros ~ poly(Semana, 5) + Bloco + Espacamento +
Seccao, family=poisson, data=dados)

modal01 <- glm(Alados ~ poly(Semana, 3) + Bloco + Espacamento +
Seccao, family=poisson, data=dados)
modal02 <- glm(Alados ~ poly(Semana, 4) + Bloco + Espacamento +
Seccao, family=poisson, data=dados)
modal03 <- glm(Alados ~ poly(Semana, 5) + Bloco + Espacamento +
Seccao, family=poisson, data=dados)
```

Likehood ratio test

```
lrtest <- function(x1, x2) {
l1 <- logLik(x1)
l2 <- logLik(x2)
d1 <- df.residual(x1)
d2 <- df.residual(x2)
lr <- 2*abs(l1 - l2)
nu <- abs(d1 - d2)
p <- 1 - pchisq(lr, nu)
return(data.frame(lr, nu, p))}

lrtest(modap01, modap02)
lrtest(modap02, modap03) # 5 degree choosen
lrtest(modal01, modal02)
lrtest(modal02, modal03) # 5 degree choosen
```

Interaction model

```
modap05 <- glm(Apteros ~ poly(Semana, 5) + Espacamento*Seccao
+ Bloco, family=poisson, data=dados)
modal05 <- glm(Alados ~ poly(Semana, 5) + Espacamento*Seccao
+ Bloco, family=poisson, data=dados)
```

Zero inflated Poisson - Constant zero inflation

```
modap07 <- zeroinfl(Apteros ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco | 1, dist="poisson", data=dados)
modal07 <- zeroinfl(Alados ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco | 1, dist="poisson", data=dados)
```

Zero inflated Poisson - Regression for zero inflation

```
modap09 <- zeroinfl(Apteros ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco | poly(Semana, 5), dist="poisson", data=dados)
modal09 <- zeroinfl(Alados ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco | poly(Semana, 5), dist="poisson", data=dados)
```

Inclusion of the covariates - Natural enemies

```
# wingless aphids model
modap11a <- zeroinfl(Apteros ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Lysipheblus | poly(Semana, 5),
dist="poisson", data=dados)
modap11b <- zeroinfl(Apteros ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Lixeiro | poly(Semana, 5),
dist="poisson", data=dados)
modap11c <- zeroinfl(Apteros ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Scymmus | poly(Semana, 5),
dist="poisson", data=dados)
modap11d <- zeroinfl(Apteros ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Cycloneda | poly(Semana, 5),
dist="poisson", data=dados)
modap11e <- zeroinfl(Apteros ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Sirfideo | poly(Semana, 5),
dist="poisson", data=dados)
modap11f <- zeroinfl(Apteros ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Aranha | poly(Semana, 5),
dist="poisson", data=dados)

# winged aphids model
modal11a <- zeroinfl(Alados ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Lysipheblus | poly(Semana, 5),
dist="poisson", data=dados)
modal11b <- zeroinfl(Alados ~ poly(Semana, 5) + Espacamento
```

```
+ Seccao + Bloco + Lixeiro | poly(Semana, 5),
dist="poisson", data=dados)
modal11c <- zeroinfl(Alados ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Scymmus | poly(Semana, 5),
dist="poisson", data=dados)
modal11d <- zeroinfl(Alados ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Cycloneda | poly(Semana, 5),
dist="poisson", data=dados)
modal11e <- zeroinfl(Alados ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Sirfideo | poly(Semana, 5),
dist="poisson", data=dados)
modal11f <- zeroinfl(Alados ~ poly(Semana, 5) + Espacamento
+ Seccao + Bloco + Aranha | poly(Semana, 5),
dist="poisson", data=dados)

lrtest(modap09, modap11a) # 4.671542  1 0.03066638
lrtest(modap09, modap11b) # 1.040407  1 0.3077268 - NAO
lrtest(modap09, modap11c) # 4.245019  1 0.03936562
lrtest(modap09, modap11d) # 1.320536  1 0.2504959 - NAO
lrtest(modap09, modap11e) # 1.193657  1 0.2745931 - NAO
lrtest(modap09, modap11f) # 35.92167  1 2.054116e-09

lrtest(modal09, modal11a) # 5.533813  1 0.01865242
lrtest(modal09, modal11b) # 0.396090  1 0.5291156 - NAO
lrtest(modal09, modal11c) # 6.228525  1 0.01257084 - NAO
lrtest(modal09, modal11d) # 6.901628  1 0.008611728
lrtest(modal09, modal11e) # 19.57989  1 9.647959e-06
lrtest(modal09, modal11f) # 4.589454  1 0.03216926
```

Example of hnp

```
ph9ap <- hnp(modap09, verb=T, print.on = T)
ph9al <- hnp(modal09, verb=T, print.on = T)
```

Fitting models with smoothing function (GAMLSS *splines*) - Poisson - ZIP - Negative binomial - ZINB)

Choosing the right degrees of freedom

```
gampoap01a <- gamlss(formula = Apteros ~ cs(Semana, df = 3),
family = PO, data = dados)
gampoap01b <- gamlss(formula = Apteros ~ cs(Semana, df = 4),
family = PO, data = dados)
gampoap01c <- gamlss(formula = Apteros ~ cs(Semana, df = 5),
family = PO, data = dados)
gampoap01d <- gamlss(formula = Apteros ~ cs(Semana, df = 6),
family = PO, data = dados)
gampoap01e <- gamlss(formula = Apteros ~ cs(Semana, df = 7),
```

```
family = PO, data = dados)
gampoap01f <- gamlss(formula = Apteros ~ cs(Semana, df = 8),
family = PO, data = dados)
gampoap01g <- gamlss(formula = Apteros ~ cs(Semana, df = 9),
family = PO, data = dados)
gampoap01h <- gamlss(formula = Apteros ~ cs(Semana, df = 10),
family = PO, data = dados)
gampoap01i <- gamlss(formula = Apteros ~ cs(Semana, df = 11),
family = PO, data = dados)
gampoap01j <- gamlss(formula = Apteros ~ cs(Semana, df = 12),
family = PO, data = dados)
gampoap01k <- gamlss(formula = Apteros ~ cs(Semana, df = 13),
family = PO, data = dados)

GAIC.table(gampoap01a,gampoap01b,gampoap01c,gampoap01d,
gampoap01e,gampoap01f,gampoap01g,gampoap01h,gampoap01i,
gampoap01j,gampoap01k)


gampoap01f <- gamlss(formula = Apteros ~ cs(Semana, df = 8)
+ Bloco + Espacamento + Seccao, family = PO, data = dados)



gampoal01a <- gamlss(formula = Alados ~ cs(Semana, df = 3),
family = PO, data = dados)
gampoal01b <- gamlss(formula = Alados ~ cs(Semana, df = 4),
family = PO, data = dados)
gampoal01c <- gamlss(formula = Alados ~ cs(Semana, df = 5),
family = PO, data = dados)
gampoal01d <- gamlss(formula = Alados ~ cs(Semana, df = 6),
family = PO, data = dados)
gampoal01e <- gamlss(formula = Alados ~ cs(Semana, df = 7),
family = PO, data = dados)
gampoal01f <- gamlss(formula = Alados ~ cs(Semana, df = 8),
family = PO, data = dados)
gampoal01g <- gamlss(formula = Alados ~ cs(Semana, df = 9),
family = PO, data = dados)
gampoal01h <- gamlss(formula = Alados ~ cs(Semana, df = 10),
family = PO, data = dados)
gampoal01i <- gamlss(formula = Alados ~ cs(Semana, df = 11),
family = PO, data = dados)
gampoal01j <- gamlss(formula = Alados ~ cs(Semana, df = 12),
family = PO, data = dados)
gampoal01k <- gamlss(formula = Alados ~ cs(Semana, df = 13),
```

```
family = PO, data = dados)

GAIC.table(gampoal01a,gampoal01b,gampoal01c,gampoal01d,
gampoal01e,gampoal01f,gampoal01g,gampoal01h,gampoal01i,
gampoal01j,gampoal01k)

gampoal01f <- gamlss(formula = Alados ~ cs(Semana, df = 8)
+ Bloco + Espacamento + Seccao, family = PO, data = dados)
```

Interaction model

```
gampoap02f <- gamlss(formula = Apteros ~ cs(Semana, df = 8)
+ Bloco + Espacamento*Seccao, family = PO, data = dados)
gampoal02f <- gamlss(formula = Alados ~ cs(Semana, df = 8)
+ Bloco + Espacamento*Seccao, family = PO, data = dados)
```

f(Week, by=Spacing); f(Week, by=Section); f(Week, by=Spacing*Section)

```
gampoap03f <- gamlss(formula = Apteros ~ cs(Semana, df = 8)*Espacamento
+ Bloco + Espacamento*Seccao, family = PO, data = dados)
gampoal03f <- gamlss(formula = Alados ~ cs(Semana, df = 8)*Espacamento
+ Bloco + Espacamento + Seccao, family = PO, data = dados)

gampoap04f <- gamlss(formula = Apteros ~ cs(Semana, df = 8)*Seccao
+ Bloco + Espacamento*Seccao, family = PO, data = dados)
gampoal04f <- gamlss(formula = Alados ~ cs(Semana, df = 8)*Seccao
+ Bloco + Espacamento + Seccao, family = PO, data = dados)

gampoap05f <- gamlss(formula = Apteros ~
cs(Semana, df = 8)*Espacamento*Seccao
+ Bloco + Espacamento*Seccao, family = PO, data = dados)
gampoal05f <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao
+ Bloco + Espacamento + Seccao, family = PO, data = dados)
```

Zero inflated Poisson - Constant zero inflation

```
gampoap06f <- gamlss(formula = Apteros ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento*Seccao, sigma.formula = ~1, family = ZIP, data = dados)

gampoal06f <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento + Seccao, sigma.formula = ~1, family = ZIP, data = dados)
```

Zero inflated Poisson - Regression for zero inflation

```
gampoap07f <- gamlss(formula = Apteros ~
```

```
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento*Seccao, sigma.formula = ~ cs(Semana, df=8),
family = ZIP, data = dados)
gampoal07f <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento + Seccao, sigma.formula = ~ cs(Semana, df=8),
family = ZIP, data = dados)
```

Inclusion of the covariates - Natural enemies

```
gampoap07f1 <- gamlss(formula = Apteros ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento*Seccao+ Lysipheblus, sigma.formula = ~ cs(Semana, df=8),
family = ZIP, data = dados)
gampoap07f2 <- gamlss(formula = Apteros ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento*Seccao+ Lixeiro , sigma.formula = ~ cs(Semana, df=8),
family = ZIP, data = dados)
gampoap07f3 <- gamlss(formula = Apteros ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento*Seccao+ Scymmus, sigma.formula = ~ cs(Semana, df=8),
family = ZIP, data = dados)
gampoap07f4 <- gamlss(formula = Apteros ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento*Seccao + Cycloneda, sigma.formula = ~ cs(Semana, df=8),
family = ZIP, data = dados)
gampoap07f5 <- gamlss(formula = Apteros ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento*Seccao+ Sirfideo, sigma.formula = ~ cs(Semana, df=8),
family = ZIP, data = dados)
gampoap07f6 <- gamlss(formula = Apteros ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento*Seccao + Aranha, sigma.formula = ~ cs(Semana, df=8),
family = ZIP, data = dados)

gampoal06f1 <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento + Seccao+ Lysipheblus, sigma.formula = ~1,
family = ZIP, data = dados)
gampoal06f2 <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento + Seccao+ Lixeiro, sigma.formula = ~1,
family = ZIP, data = dados)
gampoal06f3 <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento + Seccao+ Scymmus, sigma.formula = ~1,
```

```
family = ZIP, data = dados)
gampoal06f4 <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento + Seccao+ Cycloneda, sigma.formula = ~1,
family = ZIP, data = dados)
gampoal06f5 <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento + Seccao+ Sirfideo, sigma.formula = ~1,
family = ZIP, data = dados)
gampoal06f6 <- gamlss(formula = Alados ~
cs(Semana, df = 8)*Espacamento*Seccao + Bloco +
Espacamento + Seccao+ Aranha, sigma.formula = ~1,
family = ZIP, data = dados)
```

Example AIC with GAIC.table

```
GAIC.table(gampoap07f, gampoap07f1, gampoap07f2, gampoap07f3,
gampoap07f4, gampoap07f5, gampoap07f6)
GAIC.table(gampoal06f, gampoal06f1, gampoal06f2, gampoal06f3,
gampoal06f4, gampoal06f5, gampoal06f6)
```

Appendix B

Computation script for chapter 3

Simulating the data

```
require(gamlss)

n <- 1000
m <- 1000
y0 <- matrix(0,nrow=n,ncol=m)
for(i in 1:m){
y0[,i] <- rZINBI(n, mu = 1, sigma = 0.5, nu = 0.3)
}
```

Fitting the m models for the simulated data

```
estimates01 <- matrix(0, nrow = 6, ncol = m)
for(i in 1:m) {
fit <- gamlss(y0[,i] ~ 1,
sigma.formula = ~ 1,
nu.formula = ~ 1,
family = ZINBI)
coefs <- c(fit$mu.coefficients, fit$sigma.coefficients,
fit$nu.coefficients,lpred(fit,se.fit=TRUE)$se.fit[1],logLik(fit),
AIC(fit))
estimates01[,i] <- coefs
}
```

Extracting estimates of the parametes and bias

```
rownames(estimates01) <- c("mu","sigma","nu","St.Error","LogLike","AIC")
estimates01[1:2,] <- exp(estimates01[1:2,])
estimates01[3,] <- plogis(estimates01[3,])
estimates01 <- as.data.frame(t(estimates01))

estimates01a <- estimates01[,1:3]
estimates01b <- estimates01[,4:6]

require(reshape)
estimates01a <- melt(estimates01a)
names(estimates01a) <- c("parameter", "estimate")
estimates01a$bias <- estimates01a$estimate - c(rep(1, m), rep(0.5, m), rep(.3, m))

with(estimates01a, tapply(estimate, parameter, mean)) # mean estimate
with(estimates01a, tapply(bias, parameter, mean)) # mean bias
with(estimates01a, tapply(bias^2, parameter, mean)) # mean squared error
```

Simulating the nonlinear scenario

```
require(gamlss)

n_ind <- 20
Time <- 1:50

beta0 <- -1
beta1 <- .1

gamma0 <- .2
gamma1 <- .5
gamma2 <- -.5
period <- 20

log.mu <- beta0 + beta1 * Time
mu <- exp(log.mu)

logit.omega <- gamma0 + gamma1*sin(2*pi/period*Time) +
                        gamma2*cos(2*pi/period*Time)
omega <- plogis(logit.omega)

set.seed(32871)

m <- 1000
y0 <- matrix(0,nrow=n_ind*length(Time),ncol=m)
for(i in 1:m){
y0[,i] <- rZINBI(n_ind * length(Time), mu = rep(mu, n_ind),
                 sigma = rep(omega, n_ind))
}
y1 <- data.frame(ID = gl(n_ind, length(Time)), Time = Time, y0)
```

Fitting the m models for the nonlinear scenraio

```
estimates01 <- matrix(0, nrow = 7, ncol = m)
for(i in 3:m+2) {
try(fit <- gamlss(y1[,i] ~ Time,
sigma.formula = ~ 1, #sigma.formula = cs(Time) to include cubic spline
nu.formula = ~ 1,
family = ZINBI, data = y1))
coefs <- c(fit$mu.coefficients, fit$sigma.coefficients,
fit$nu.coefficients,lpred(fit,se.fit=TRUE)$se.fit[1],logLik(fit),
AIC(fit))
estimates01[,i] <- coefs
}
```

Function to extract AIC and select the best model

```
AICS01 <- data.frame(estimates01b[,3],estimates02b[,3],
                     estimates03b[,3],estimates04b[,3])
colnames(AICS01) <- c("ZINB","ZIP","N.B.","POIS")

sentido <- c("linha","coluna")

minimoAIC <- function(x,sentido,função){
#sentido: "linha" ou "coluna"
resultado <- c()
fc<-função
if(sentido=="linha"){
for (i in 1:nrow(x)){
resultado[i]<- fc(x[i,])
}
}
else{
for (i in 1:ncol(x)){
resultado[i] <- fc(x[,i])
}
}
return(resultado)
}
minimoAIC(AICS01,"linha",min)

funcanzinha <- function(x,y){
resultado <- c()
for (i in 1:1000){
resultado[i]<- identical(x[i],y[i])
}
return(resultado)}

AICS01 <- data.frame(estimates01b[,3],estimates02b[,3],estimates03b[,3],
                     estimates04b[,3])
colnames(AICS01) <- c("ZINB","ZIP","N.B.","POIS")

AICS01$Minimum <- minimoAIC(AICS01,"linha",min)

sum(funcanzinha(AICS01[,1],AICS01[,5]),na.rm = TRUE)
```