University of São Paulo
Luiz de Queiroz College of Agriculture

Flexible models for hierarchical and overdispersed data in agriculture

Ricardo Klein Sercundes

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

Piracicaba
2018

Ricardo Klein Sercundes
Agronomist

Flexible models for hierarchical and overdispersed data in agriculture

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **CLARICE GARCIA BORGES DEMÉTRIO**

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

Piracicaba
2018

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

Sercundes, Ricardo Klein

Flexible models for hierarchical and overdispersed data in agriculture / Ricardo Klein Sercundes. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2018 .

64 p.

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

1. Modelo combinado 2. Modelo linear generalizado misto 3. B-spline 4. Distribuição multinomial 5. Distribuição beta 5. Verossimilhança 4. . I. Título.

DEDICATION

To my lovely Maria Cristina Martins

ACKNOWLEDGMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis, especially my parents Ricardo Bezerra Sercundes and Marie Luise Klein Sercundes and my sister Michelle Klein Sercundes Sinti, for their love and supporting me throughout my life.

To my girlfriend Maria Cristina Martins, for her huge patience throughout the whole process of my Masters/Doctorate life. Thank you so much for understanding me.

To my adviser, Prof. Dr. Clarice Garcia Borges Demétrio, for the continuous support of my Doctorate, for her patience, motivation, enthusiasm and immense knowledge.

To Prof. Dr. Geert Molenberghs and Prof. Dr. Geert Verbeke from UHasselt and KU Leuven, for receiving me on my internship period, for the scientific contribution, intellectual input and for all the support during my stay in Belgium.

To all my friends from ESALQ/USP, KU Leuven and UHasselt.

To the staff of the Department of Exact Science at ESALQ/USP, the secretaries Luciane Brajão, Rosni Pinto and Solange Sabadin, in special to the computer technicians Eduardo Bonilha and Jorge Wiendl.

This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil.

CONTENTS

Resumo . . . 6
Abstract . . . 7
1 Introduction . . . 9
References . . . 11
2 Random smoothing splines to model fluctuations in insect populations . . . 15
2.1 Introduction . . . 15
2.2 Biological control dataset . . . 16
2.3 Building blocks . . . 17
2.3.1 Generalized linear models and overdispersion . . . 17
2.3.2 Generalized linear mixed models . . . 19
2.3.3 Random smoothing splines as GLMM . . . 21
2.4 Analysis of the biological control dataset . . . 22
2.5 Concluding remarks . . . 26
References . . . 26
3 A combined overdispersed longitudinal model for nominal data . . . 31
3.1 Introduction . . . 31
3.2 Grazing management dataset . . . 33
3.3 Building blocks . . . 34
3.3.1 Generalized linear models and overdispersion . . . 34
3.3.2 Generalized linear mixed models and combined models . . . 36
3.4 Combined model for nominal outcomes . . . 38
3.5 Parameter estimation . . . 39
3.6 Simulation . . . 40
3.7 Analysis of the grazing management data . . . 44
3.8 Concluding remarks . . . 46
References . . . 47
4 Final considerations . . . 51
Appendices . . . 53

RESUMO

Modelos flexíveis para dados hierárquicos e superdispersos na agricultura

Nesse trabalho, exploramos e propusemos modelos flexíveis para a análise de dados hierárquicos e superdispersos na agricultura. Um modelo linear generalizado semi-paramétrico misto foi aplicado e comparado com os principais modelos para a análise de dados de contagem e, um modelo combinado que leva em consideração a superdispersão e a hierarquia dos dados por meio de dois efeitos aleatórios distintos foi proposto para a análise de dados nominais. Todos os códigos computacionais foram implementados no *software* SAS 9.4 sendo disponibilizados no apêndice.

Palavras-chave: Modelo combinado, Modelo linear generalizado misto, B-spline, Distribuição multinomial, Distribuição beta, Verossimilhança.

ABSTRACT

Flexible models for hierarchical and overdispersed data in agriculture

In this work we explored and proposed flexible models to analyze hierarchical and overdispersed data in agriculture. A semi-parametric generalized linear mixed model was applied and compared with the main standard models to assess count data and, a combined model that take into account overdispersion and clustering through two separate sets of random effects was proposed to model nominal outcomes. For all models, the computational codes were implemented using the SAS 9.4 software and are available in the appendix.

Keywords: Combined model, Generalized linear mixed model, B-spline, Multinomial distribution, Beta distribution, Likelihood.

1 INTRODUCTION

The growing demand for theory and data analysis tools has made the agricultural sciences a great niche of research in statistics. Data with complex structures, such as breeding trials across regions (Oliveira *et al.*, 2005; Gonçalves *et al.*, 2014), longitudinal studies in animals (Andersen *et al.*, 2007; Machado *et al.*, 2012), entomological zero-inflated data (Demétrio *et al.*, 2014), and heritability studies that consider many animals and family relations (Vazquez *et al.*, 2013) are just some of the challenges provided in this area.

When non-Gaussian data are studied, many approaches have been proposed, which often can be placed within the generalized linear modeling (GLM) framework (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Agresti, 2010), i.e., a unifying framework based on the so-called exponential family distributions, although taking into account the nature of the outcomes using several density or probability functions, e.g., Bernoulli/Binomial, gamma, Poisson and multinomial, the GLMs may sometimes be very restrictive because of the so-called mean-variance relationship, i.e., the variance is expressed as a deterministic function of the mean. In many practical situations in agricultural field experiments, mainly when hierarchical structures or highly variable data arise, this restriction is not in line with a particular set of data, and may cause serious flaws in point and precision estimation and inference on important parameters (Plackett and Wedderburn, 1978; Hinde and Demétrio, 1998; Cox, 1983). This may lead to incorrect conclusions; for instance, a treatment which does not have a significant effect could be assessed as if it does. Two phenomena can occur: overdispersion and underdispersion. The former one arises when the observed variance from the data is greater than the theoretical variance (restricted by the model's mean-variance relationship) from the model, while the latter one is obtained when the observed variance is smaller than the theoretical variance. In this research, emphasis is placed on overdispersion.

Several routes can be taken to model properly the mean-variance relationship, being one of the most popular frameworks developed by Breslow and Clayton (1993) called generalized linear mixed models (GLMM), where the GLM and the random-effects ideas are combined. Although flexible, to build a GLMM is not a trivial task because many aspects have to be considered, such as the correct specification of the response variable distribution (binomial, Poisson, Gaussian, gamma, etc.), the definition of a linear predictor, a data-coherent link function, and additionally the random effects structure. However, in some cases, the simple inclusion of a random effect is not sufficient to model the data properly, necessitating the inclusion of more elements, such as semi-parametric

approaches (Durbán *et al.*, 2005; Faes *et al.*, 2006; Ruppert *et al.*, 2003) or models that include more than one random effect (Molenberghs *et al.*, 2007, 2010, 2012; Ivanova *et al.*, 2014; Molenberghs *et al.*, 2017).

In this sense, the aims of this work are to explore and develop some flexible models in order to solve problems related with agricultural experiments. For this, two motivating case studies were considered, the first one related with biological control in citrus orchards and the second one with pasture production.

The citrus production is one of the most important sectors of modern agribusiness. Around the world, the annual production currently stands to 100 million tonnes, covering an area of approximately 7.5 million hectares in more than 100 countries (FAO, 2012). Brazil is one of the largest orange producers currently being responsible for over 50% of the world's orange juice production. According to Neves *et al.* (2011) the citriculture is currently present in over three thousand Brazilian municipalities, with almost four hundred of them located in São Paulo State, generating more than two hundred thousand direct and indirect jobs and US$ 1.5–2.5 billion every year.

In 2009, the citrus sector was the second most intensive crop in Brazil to use pesticides (cotton was the first), totaling 725,577 tonnes of commercial products and corresponding to a total of US$ 288.2 million (Neves *et al.*, 2011). The indiscriminate usage of pesticides can lead to many problems such as pest resistance, the reduction of natural enemies and the emergence of secondary pests (Cutler, 2013; Hardin *et al.*, 1995). To promote a more competitive and green productive system, few biological products have been developed to control the pests in the orange orchards. However, fungi-based biopesticides have increased in popularity because they have the capacity to infect a large number of pests and to remain in the environment (Alves, 1998).

To assess the impact of these new products, many field experiments with longitudinal studies have been carried out. In these studies, counts and binary data usually arise to quantify the abundance, diversity and treatment effects. Due to climate changes, the insect life cycle and migration, field experiments usually show high variability and a nonlinear association between the outcome and covariates. In this way, classical, fully parametrically models cannot explain the biological phenomena properly, requiring more flexible versions, such as the semi-parametric generalized linear mixed model studied in Chapter 2. We applied this framework and compared with the main standard models for count data in order to model the correlation between repeated measures and the overdispersion.

The beef and milk chain are other very important Brazilian economic sectors, representing around 7% of Brazil's Gross Domestic Product (GDP) (MAPA, 2014). The production in these sectors is performed, mainly on pasture based systems (Stock *et al.*, 2011; Millen *et al.*, 2009). These systems have several benefits, e.g. lower costs, better animals welfare and nutrient cycling in the enviroment, but it is only sustainable and competitive if performed with an efficient management. Grazing management has been the focus of the research with forage plants in Brazil for many years. Howbeit, it was only during the last decade that significant changes and advances occurred regarding the understanding of important factors and processes that determine adequate use of tropical forage plants in pastures (Silva and Nascimento Júnior, 2007).

According to Pereira *et al.* (2014) tall-tufted tussock-forming species represent the main growth form among the tropical grasses with higher potential for herbage production utilized in South America. However, knowledge on how environmental factors and management affect the horizontal structure and lateral expansion of tussocks or how grazing affects the soil occupation capacity of those plants is scarce. In grazing management studies, it is common to observe several types of outcomes in the same plot or paddock over a period of time. When the outcome is nominal (more than two categories without order between them), few techniques are available in the literature to analyse the relationship between the longitudinal outcome and extra-variability sources. Thus, in Chapter 3, a combined model that take into account overdispersion and clustering through two separate sets of random effects was proposed to model nominal outcomes. A simulation study and also the analysis of a dataset involving a longitudinal experiment with grass pasture and dairy cows were performed. For all Chapters, in the appendices, we show in details how to implement these models in the statistical software package SAS 9.4.

References

Agresti, A., 2010 *Categorical data analysis*. Wiley, New York.

Alves, S. B., 1998 *Controle microbiano de insetos*. Piracicaba, second edition.

Andersen, H. M., E. Jørgensen, L. Dybkjær, and B. Jørgensen, 2007 The ear skin temperature as an indicator of the thermal comfort of pigs. Applied Animal Behaviour Science **113**: 43–56.

Breslow, N. E. and D. G. Clayton, 1993 Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association **88**: 9–25.

Cox, D. R., 1983 Some remarks on overdispersion. Biometrika **70**: 269–274.

Cutler, G. C., 2013 Insects, insecticides and hormesis: Evidence and considerations for study. Dose-Response **11**: 154–177.

Demétrio, C. G. B., J. Hinde, and R. A. Moral, 2014 Models for Overdispersed Data in Entomology. In *Ecological modeling applied to entomology*, Springer.

Durbán, M., J. Harezlak, M. P. Wand, and R. J. Carroll, 2005 Simple fitting of subject-specific curves for longitudinal data. Statistics in Medicine **24**: 1153–1167.

Faes, C., M. Aerts, H. Geys, L. Bijnens, L. Ver Donck, and W. J. Lammers, 2006 GLMM approach to study the spatial and temporal evolution of spikes in the small intestine. Statistical Modelling **6**: 300–320.

FAO, 2012 Food and Agriculture Organization of the United Nations.

Gonçalves, G. M., A. P. Viana, M. Deon, and V. D. Resende, 2014 Breeding new sugarcane clones by mixed models under genotype by environmental interaction. Scientia Agricola pp. 66–71.

Hardin, M. R., B. Benrey, M. Coll, W. O. Lamp, G. K. Roderick, and P. Barbosa, 1995 Arthropod pest resurgence: an overview of potential mechanisms. Crop Protection **14**: 3–18.

Hinde, J. and C. G. B. Demétrio, 1998 Overdispersion: Models and estimation. Computational Statistics and Data Analysis **27**: 151–170.

Ivanova, A., G. Molenberghs, and G. Verbeke, 2014 A model for overdispersed hierarchical ordinal data. Statistical Modelling **14**: 399–415.

Machado, N. S., C. Akemi, and W. Marques, 2012 Resfriamento da cobertura de aviários e seus efeitos na mortalidade e nos índices de conforto térmico. Nucleus **9**: 59–74.

MAPA, 2014 Ministério da Agricultura, Pecuária e Abastecimento.

McCullagh, P. and J. Nelder, 1989 *Generalized linear models*. Boca Raton.

Millen, D. D., R. D. Pacheco, M. D. Arrigoni, M. L. Galyean, and J. T. Vasconcelos, 2009 A snapshot of management practices and nutritional recommendations used by feedlot nutritionists in Brazil. Journal of Animal Science **87**: 3427–3439.

Molenberghs, G., G. Verbeke, and C. G. B. Demétrio, 2007 An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Analysis pp. 513–531.

Molenberghs, G., G. Verbeke, and C. G. B. Demétrio, 2017 Hierarchical models with normal and conjugate random effects: a review. SORT-Statistics and Operations Research Transactions **41**: 191–254.

Molenberghs, G., G. Verbeke, C. G. B. Demétrio, and A. M. C. Vieira, 2010 A Family of Generalized Linear Models for Repeated Measures with Normal and Conjugate Random Effects. Statistical Science **25**: 325–347.

Molenberghs, G., G. Verbeke, S. Iddi, and C. G. B. Demétrio, 2012 A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. Journal of Multivariate Analysis **111**: 94–109.

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

Neves, M. F., V. G. Trombim, and F. F. Lopes, 2011 *The Orange Juice Business: a Brazilian Perspective*.

Oliveira, R. A. D., M. Deon, V. D. Resende, E. Daros, C. Zambon, O. T. Ido, H. Weber, and H. S. Koehler, 2005 Genotypic evaluation and selection of sugarcane clones in three environments in the state of Paraná. Crop breeding and applied technology pp. 426–434.

Pereira, L. E., A. J. Paiva, E. V. Geremia, and S. C. da Silva, 2014 Grazing management and tussock distribution in elephant grass. Grass and Forage Science **70**: 406–417.

Plackett, P. S. R. and R. W. M. Wedderburn, 1978 Inference sensitivity for Poisson mixtures. Biometrika **65**: 591–602.

Ruppert, D., M. Wand, and R. Carroll, 2003 *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics.

Silva, S. C. D. and D. D. Nascimento Júnior, 2007 Avanços na pesquisa com plantas forrageiras tropicais em pastagens: características morfofisiológicas e manejo do pastejo. Revista Brasileira de Zootecnia **36**: 121–138.

Stock, L. A., R. Zoccal, G. R. de Carvalho, and N. B. Siqueira, 2011 Competitividade do agronegócio do leite brasileiro. Embrapa p. 326.

Vazquez, A. I., D. M. Bates, G. J. M. Rosa, D. Gianola, and K. A. Weigel, 2013 Technical note: An R package for fitting generalized linear mixed models in animal breeding. Journal of animal science **88**: 497–504.

2 RANDOM SMOOTHING SPLINES TO MODEL FLUCTUATIONS IN INSECT POPULATIONS

Abstract: When entomological experiments are performed, non-Gaussian data following longitudinal design are usually observed. To analyze such data, generalized linear mixed models (GLMMs) are frequently used because they can handle different distributions as well as correlations between repeated measures. Although less familiar, this framework can also be used for smoothing purposes, allowing to fit highly variable data. We explore this methodology in a biological control dataset where the entomopathogenic fungus *Isaria fumosorosea* ESALQ-1296 was sprayed onto an organic citrus orchard and the fluctuations in population levels of *Diaphorina citri*, a pest, were compared to plots without application. For this, standard models for count data, the negative binomial and Poisson-normal model, were fitted and compared with versions that incorporate random smoothing splines. The results allow to conclude that the fungus has potential to reduce the number of *D. citri* in the field. The random smoothing splines framework also proved to be a good way to model properly the overdispersion generated by the nonlinearity between the longitudinal covariate and the response variable. We also show how to implement these models in the statistical software package SAS 9.4.
Keywords: Generalized linear mixed model; Semi-parametric model; B-spline; Citrus; Integrated pest management; Biological control.

2.1 Introduction

The field behavior of pests and natural enemies is often one of the main objectives of many entomological experiments (Yamamoto *et al.*, 2001; Beloti *et al.*, 2013; Rodrigues *et al.*, 2014). In these studies, non-Gaussian data with high variability and longitudinal structures may arise, making the data analysis surprisingly challenging. To handle this, classical generalized linear mixed models are commonly used, but this metodology, in many cases, is not able to explain the biological phenomena properly, requiring more flexible versions, such as semi-parametric or non-parametric models.

There are many references on non-parametric regression with independent data using kernel and spline methods (Härdle *et al.*, 1999; Green and Silverman, 1994). Generalized additive models (Hastie and Tibshirani, 1990) are also widely used and well understood. However, only very limited work has been done on non-parametric regression when the data have a hierarchical structure. Speed (1991) was the first to make the connection between non-parametric regression and mixed models. Since then, the use

of smoothing splines within the mixed-model framework is becoming more and more appreciated and the number of applications in the literature is growing, see, for example, the work of Verbyla *et al.* (1999); Ruppert *et al.* (2003); Wand (2003); Ngo and Wand (2004); Durbán *et al.* (2005); Molenberghs and Verbeke (2005).

The beauty of random smoothing splines lies in the fact that these models are very flexible and allow to capture a sample unit structure. Moreover, they can be expressed using the standard mixed-effects theory and it is possible to select the degree of smoothing in the model automatically (Molenberghs and Verbeke, 2005; Faes *et al.*, 2006). The objective of this Chapter is to describe some aspects of the random smoothing splines methodology and to present an application in agriculture, where, in general, this methodology is not employed.

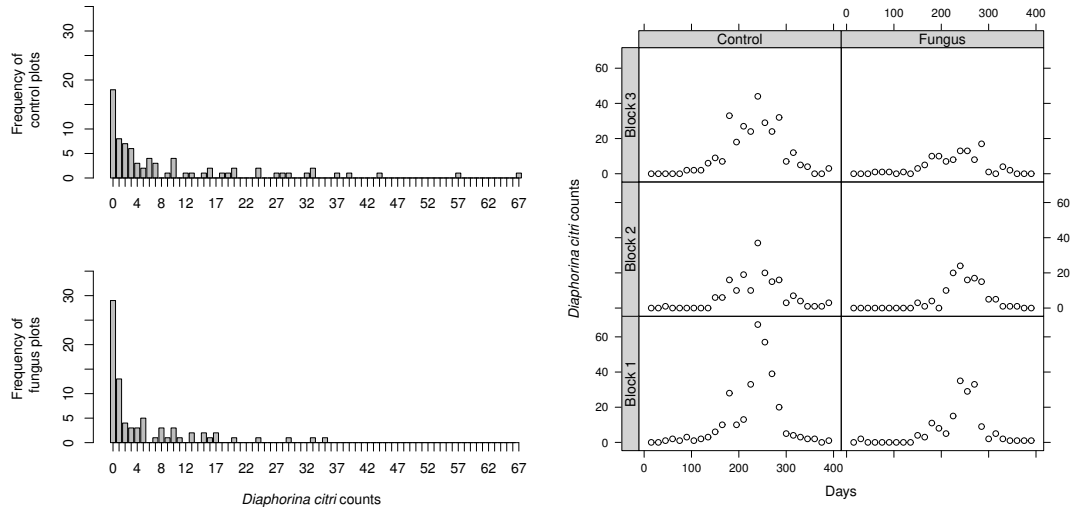
The Chapter is organized as follows. In Section 2.2 a motivating dataset is introduced. Basic ingredients for the modeling framework, generalized linear models, overdispersion, generalized linear mixed models, random smoothing splines and parameter estimation are the subject of Section 2.3. The dataset is analyzed is Section 2.4. Concluding remarks are offered in Section 2.5. Finally, we show in detail how to implement these models in SAS 9.4 in the Appendix A.

2.2 Biological control dataset

To assess the effect of the entomopathogenic fungus *I. fumosorosea* ESALQ-1296 in the population levels of *D. citri*, Conceschi (2017) conducted an experiment at a commercial organic citrus orchard in the city of Itirapina, São Paulo, Brazil. The experimental area was divided into three blocks, each one with two plots of 30,000 square meters. Six traps were set at the center line of each plot spaced 24 meters from each other and the number of *D. citri* were counted every 15 days for 26 time points.

The treatments were control (without fungus spraying for *D. citri* management), and spraying a fungi suspension containing 5×10^6 conidia.ml^{-1} in a total volume of 60 ml.m^{-3} for the canopy. The applications were carried out monthly using a motorized air blast sprayer (Jacto ARBUS 2000), requiring an increase in the frequency based on pest monitoring. Control of all other pests and diseases were the same for all plots. The experiment began in April 2014 and was concluded in June 2015. Due to the high number of zeros, we worked with the sums of counts per trap, resulting in 2 treatments$\times$3 blocks$\times$ 26 days = 156 observations. As a summary of the data, Fig. (2.1.a) shows frequency plots where a very skewed distribution for both treatments is observed. The longitudinal behavior for each plot is shown in Fig. (2.1.b). There is an increase of counts between 135

and 330 days (November to April), coinciding with the intense vegetative growth of the orchard. According to Patt and Sétamou (2010), and Kobori *et al.* (2011), adults of *D. citri* are attracted by flushing shoots to oviposit, resulting in an increase of insects in the field. Similar patterns of fluctuations were also observed by Yamamoto *et al.* (2001) and Beloti *et al.* (2013) in other orchards of São Paulo State.



(a) (b)

Figure 2.1: Biological control of *Diaphorina citri* with *Isaria fumosorosea* ESALQ-1296 data: (a) Frequency plot for each treatment group; (b) Number of sampled *Diaphorina citri* for each treatment and block over 26 days.

2.3 Building blocks

In this section, we briefly present the main concepts of the random smoothing splines framework. In Section 2.3.1 we introduce the exponential family, the generalized linear models and overdispersion. In Section 2.3.2 we present some properties of generalized linear mixed models and in Section 2.3.3 we explain how smoothing methods can be incorporated within the mixed modeling framework.

2.3.1 Generalized linear models and overdispersion

The class of generalized linear models (GLM) was introduced by Nelder and Wedderburn (1972) as a framework for handling a range of common statistical models for Gaussian and non-Gaussian data. A GLM is defined basically in terms of three components. The first is a set of independent random variables, $Y_1, \ldots, Y_N$, with probability

or density function that belongs to the exponential family and can be written as

$$f(y_i|\eta_i,\phi) \;=\; \exp\left\{\phi^{-1}[y_i\eta_i - \psi(\eta_i)] + c(y_i,\phi)\right\},$$

where $\psi(\cdot)$ and $c(\cdot)$ are known functions and ϕ and η_i are called the dispersion and natural or canonical parameter, respectively. The exponential family embraces several distributions, e.g. normal, Bernoulli, Binomial, Poisson, gamma, multinomial and, consequently, models belonging to this family share some basic features, such as the first two moments derived from the function $\psi(\cdot)$ given by

$$\mathrm{E}(Y_i) = \mu_i = \psi'(\eta_i), \qquad \mathrm{Var}(Y_i) = \sigma^2 = \phi\psi''(\eta_i).$$

With exception of the normal distribution, the mean and variance of the exponential family distributions are related through $\sigma^2 = \phi\psi''[\psi'^{-1}(\mu)] = \phi v(\mu)$, with $v(\cdot)$ the so-called variance function. The second component, called linear predictor or natural parameter, η_i, is the quantity which incorporates the information about the independent variables into the model. The third component, called link function, $h(\cdot)$, provides the relationship between the linear predictor and the mean of the distribution as $\mu_i = h(\eta_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$, where $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are vectors of covariates and fixed unknown regression coefficients, respectively.

For count data, the Poisson distribution is usually a natural starting point for data analysis because it can handle the non-negative support function and the discrete behavior of the outcomes. Thus, if we assume that Y_i is Poisson distributed. The probability function in this case is given by

$$f(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0,1,2\ldots, \quad \lambda_i > 0, \tag{2.1}$$

with natural parameter $\eta_i = \ln(\lambda_i)$, mean $\mu_i = \lambda_i$, dispersion parameter $\phi = 1$ and variance function $\nu(\mu_i) = \mu_i = \lambda_i$. The logarithm is the natural link function, leading to the classical Poisson regression model with $\ln(\lambda_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

It is well known that (2.1) implies the variance and the mean to be equal. However, comparing the sample average with the sample variance might already reveal that frequently, this assumption is not met in real data. According to Grunwald *et al.* (2011) and Demétrio *et al.* (2014), cases where the variance is greater than the mean are largely reported in the literature as overdispersion, which may occur due to the absence of relevant covariates, heterogeneity of sampling units, hierarchical structures and excess of zeros. It should be noted that underdispersion can occur as well. Thus, it is important to adapt models to take into account this feature in order to avoid incorrect and misleading inferences (Hinde and Demétrio, 1998).

An elegant way to extend the Poisson model to account for overdispersion is through a two-step approach allowing the Poisson mean, λ_i, to vary according to some distribution (Hinde and Demétrio, 1998). If we treat λ_i as continuous, since it must also be positive, a natural choice is to assume that λ_i follows a gamma distribution, giving rise to the negative binomial model with probability function

$$f(y_i|\lambda_i,\phi) = \frac{\Gamma(y_i+\frac{1}{\phi})}{\Gamma\left(\frac{1}{\phi}\right)\Gamma(y_i+1)}\left(\frac{\phi\lambda_i}{1+\phi\lambda_i}\right)^{y_i}\left(\frac{1}{1+\phi\lambda_i}\right)^{\frac{1}{\phi}},\ y_i=0,1,2,\ldots,\quad \phi>0,\ \lambda_i>0,$$

where $\mathrm{E}(Y_i)=\lambda_i$ and $\mathrm{Var}(Y_i)=\lambda_i+\lambda_i^2\phi$. The negative binomial distribution is similar to the Poisson distribution, but incorporates a variance that is larger than its mean. As a result, it is more flexible and can accommodate more distributional shapes than the Poisson distribution (Gbur, 2012; Demétrio *et al.*, 2014). Another way to model overdispersion for count data consists in adding an observation-level random effect to the linear predictor, as described in Section 2.3.2.

2.3.2 Generalized linear mixed models

In many entomological field experiments, the studies are carried out in such a way that several measurements are taken from the same cluster, subject or sample unit over time, characterizing a longitudinal study. To analyze these data, appropriate statistical approaches are needed, such as generalized linear mixed models (GLMMs) where the generalized linear modeling and the linear mixed modeling frameworks are combined to accommodate both correlation and overdispersion. In full generality, one assumes that, conditionally on q-dimensional random effects $\boldsymbol{b}_i$, assumed to be drawn independently from a Gaussian distribution, $N_q(\boldsymbol{0},\boldsymbol{D})$, the outcomes Y_{ij} measured on the i-th sample unit at j-th time point ($i=1,\ldots,N; j=1,\ldots,n_i$) are independent with densities of the form

$$f_i(y_{ij}|\boldsymbol{b}_i,\boldsymbol{\beta},\phi) = \exp\left\{\phi^{-1}[y_{ij}\eta_{ij}-\psi(\eta_{ij})]+c(y_{ij},\phi)\right\},$$

with $\eta(\mu_{ij})=\eta[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i)]=\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i$, with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, respectively. Finally, let $f(\boldsymbol{b}_i|\boldsymbol{D})$ be the density of the $N_q(\boldsymbol{0},\boldsymbol{D})$ distribution for the random effects $\boldsymbol{b}_i$. For count data, if the standard Poisson model is used, this gives rise to a Poisson-normal model where

$$\begin{aligned} Y_{ij}|\boldsymbol{b_i} &\sim \text{Poisson}(\lambda_{ij}),\ \lambda_{ij}>0 \\ \ln(\lambda_{ij}) &= \boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \\ \boldsymbol{b_i} &\sim N_q(\boldsymbol{0},\boldsymbol{D}). \end{aligned}$$

In the special case of univariate data with a single normal random effect $b_i \sim \mathrm{N}(0, d)$, the mean and the variance of the Poisson-normal model are given by

$$\mathrm{E}(Y_i) = \lambda_i = \exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \frac{1}{2}d\right) \quad \text{and} \quad \mathrm{Var}(Y_i) = \lambda_i + \lambda_i^2(e^d - 1),$$

the latter being a quadratic form, very close to the variance function of the negative binomial model. Thus, the Poisson-normal model can also accommodate a certain amount of overdispersion besides correlation. Estimates of $\boldsymbol{\beta}$, $\boldsymbol{D}$ and ϕ for GLMMs are obtained from maximizing the marginal likelihood, integrating out the random effects and commonly written as:

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{D}, \phi) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{b}_i|\boldsymbol{D}) d\boldsymbol{b}_i. \tag{2.2}$$

The key problem in maximizing (2.2) is the presence of N integrals over the q-dimensional random effects $\boldsymbol{b_i}$. In some special cases (e.g. the linear mixed models for continuous outcomes or the probit-normal model) these integrals can be solved analytically but in general, numerical approximations are needed such as Bayesian methods using Gibbs sampling, the EM algorithm (Demidenko, 2004), penalized quasi-likelihood (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993), the double penalized quasi-likelihood (Lin and Zhang, 1999) and adaptive and nonadaptive Gaussian quadrature (Molenberghs and Verbeke, 2005). It is important to highlight that default methods in software packages are not always the best option to be used. The choice should be made according to the characteristics of data and design.

In this Chapter, we focus on the Laplace approximation to solve the integrals over the random effects. This method is identical to the adaptative Gaussian quadrature with one quadrature point and essentially approximates the integrands by a tractable one, making the numerical maximization feasible. When the complexity of the model and the number of random effects increases, the number of quadrature points makes the adaptative Gaussian quadrature method computationally intensive and eventually prohibitive (Gbur, 2012). However, the Laplace method works well for a considerable number of mixed models and is implemented in a wide range of software packages as glimmix and nlmixed in SAS and glmer and lme4 in R. It also yields an exact approximation for Gaussian kernels and becomes more precise when the number of repeated measures per sample unit increases (Molenberghs and Verbeke, 2005; Joe, 2008; Capanu *et al.*, 2013). Its behavior in count data, unless when counts are very small, is usually acceptable.

2.3.3 Random smoothing splines as GLMM

There are many causes of variability in entomological field experiments such as environmental changes, the insect life cycle, migration, competition, predation and parasitism. Although very flexible, the parametric GLMM framework is not aways appropriate to describe how the response variable is affected by time, other covariates and the features of the individual profiles. Hence, in this section, we explore how smoothing methods that use base function with penalization can be formulated as estimators in a generalized mixed model framework. In analogy with VERBYLA *et al.* (1999), DURBÁN *et al.* (2005) and FAES *et al.* (2006), we can write the unknown smooth function f in the linear predictor as

$$\eta_{ij} = f(t_{ij}) = \beta_0 + \beta_1 t_{ij} + \ldots + \beta_d t_{ij}^d + \sum_{k=1}^{K} \gamma_{ik} B_{k,d}(t_{ij}), \quad \gamma_{ik} \sim N(0, \tau^2),$$

where $\beta_0, \ldots, \beta_d$ are regression coefficients, t_{ij} is the time covariate at sample unit i and time point j, γ_{ik} is the random spline coefficient at knot k, $k = 1, \ldots, K$, and $B_{k,d}(t_{ij})$ is a base function (e.g. truncated power base, radial base or B-spline base) with degree d. According to RUPPERT *et al.* (2003) a change of base does not change the fit but, for B-splines, it can induce numerical stability and consequently greater accuracy to the estimates. In this sense, we placed emphasis on B-spline functions that are defined recursively with base of degree zero ($d = 0$) for the k-th interval knots κ_k and κ_{k+1} ($\kappa_k \leq \kappa_{k+1}$) given by

$$B_{k,0}(t_{ij}) = \begin{cases} 1 & \text{if } \kappa_k \leq t_{ij} \leq \kappa_{k+1}, \\ 0 & \text{otherwise.} \end{cases}$$

Higher-degree base are then determined from the values of the lower degree base function, already evaluated, and the width of the adjoining intervals between knots:

$$B_{k,d}(t_{ij}) = \frac{t_{ij} - \kappa_k}{\kappa_{k+d} - \kappa_k} B_{k,d-1}(t_{ij}) + \frac{\kappa_{k+d+1} - t_{ij}}{\kappa_{k+d+1} - \kappa_{k+1}} B_{k+1,d-1}(t_{ij}), \quad d = 1, 2, \ldots$$

The flexibility of a spline curve is given by the degree of the base function, the number of the knots and their location. Often, the researcher needs to perform a sensitivity analysis in order to find the best option for each problem. However, some general recommendations can be used such as base of degree 2 or 3, knots located in the quantiles of the longitudinal covariate and the number of knots derived by the rule of thumb $K = \min\{(\text{number of unique } t_{ij}/4), 35\}$.

In matrix notation, with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^T$, $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{KN})^T$, $\boldsymbol{X}$ the matrix with the i-th row $\boldsymbol{X_i} = [1\ t_i \ldots t_i^d]$ and $\boldsymbol{Z}$ the matrix with the i-th row $\boldsymbol{Z_i} = [B_{1,d}(t_i) \ldots B_{K,d}(t_i)]$, the natural parameter can be written as

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \qquad \boldsymbol{\gamma} \sim N_{KN}(\mathbf{0}, \tau^2 \boldsymbol{I}_{KN\times KN}),$$

and the penalized spline function as

$$f(\boldsymbol{y}|\phi, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \phi^{-1}\Big[\boldsymbol{y}(\boldsymbol{X\beta} + \boldsymbol{Z\gamma}) - \mathbf{1}^T\psi(\boldsymbol{X\beta} + \boldsymbol{Z\gamma})\Big] - \frac{1}{2}\Omega^2 \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}^T \boldsymbol{D} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}, \quad (2.3)$$

with $\mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}) = \psi'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma})$, $\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\gamma}) = \mathrm{diag}\{\psi''(\boldsymbol{X\beta} + \boldsymbol{Z\gamma})\}$ and $\boldsymbol{D} = \mathrm{diag}(\mathbf{0}_{d+1}, \mathbf{I}_{KN\times KN})$ a known positive semi-definite penalty matrix (Wahba, 1978; Green and Silverman, 1994). The first term of (2.3) measures the goodness-of-fit while the second is the roughness penalty with smoothing parameter defined as $\Omega^2 = \phi/\tau^2$, selected automatically via the fitting algorithm. Large values of Ω produce smoother curves while smaller values produce wiggly curves. Estimates of (2.3) are obtained from maximizing the marginal likelihood similarly to GLMMs. Thus, the Laplace method is an interesting option to work with random smoothing splines because the flexibility of the fitted curve is given by several random effects that follows the same distribution for each adjoining intervals of the longitudinal covariate.

2.4 Analysis of the biological control dataset

For the data analysis we assumed that the outcome, Y_{ij}, is the sum of counts of *D. citri* measured in the i-th plot ($i = 1, 2, \ldots, 6$) on the j-th day ($j = 15, 30, \ldots, 390$). We started assessing the mean-variance relationship of the data. The majority of points in the sample means and variances dispersion diagram (Fig.2.2.a) are above the identity line, clearly suggesting the presence of overdispersion.

To assess the effects of the experimental factors, four models for count data that take into account overdispersion were proposed: the negative binomial (NB), Poisson-normal (PN), Poisson model with random B-splines (Prs) and negative binomial model with random B-splines (NBrs). To avoid numerical instability, the day covariate was standardized as $t_{ij} = \{[\mathrm{day}_{ij} - \mathrm{mean}(\mathrm{day}_{ij})]/\mathrm{standard\ deviation}(\mathrm{day}_{ij})\}$. For all cases, the canonical link function, i.e., the exponential link, was used and the linear predictors were given by:

$$\eta_{ij} = \beta_0 + \beta_1 \mathrm{B1}_i + \beta_2 \mathrm{B2}_i + \beta_3 \mathrm{TF}_i + \beta_4 \mathrm{t}_{ij} + \beta_5 \mathrm{TF}_i \mathrm{t}_{ij} + \beta_6 \mathrm{t}_{ij}^2 + \beta_7 \mathrm{TF}_i \mathrm{t}_{ij}^2,$$

with

$$\begin{array}{ll} \mathrm{Y}_{ij} \sim \text{NB:} & \ln(\lambda_{ij}) = \eta_{ij}, \\ \mathrm{Y}_{ij}|b_i \sim \text{PN:} & \ln(\lambda_{ij}) = \eta_{ij} + b_i, \\ \mathrm{Y}_{ij}|\gamma_{ik} \sim \text{Prs and NBrs:} & \ln(\lambda_{ij}) = \eta_{ij} + \sum_{k=1}^{6} \gamma_{ik} B_{k,2}(t_{ij}), \end{array}$$

where $\mathrm{B1}_i$, $\mathrm{B2}_i$ and TF_i are dummy variables for blocks 1, 2 and fungus treatment respectively, t_{ij} the standardized time that the outcome was measured, $\mathrm{TF}_i \mathrm{t}_{ij}$ the interaction between treatment and time, $\mathrm{TF}_i \mathrm{t}_{ij}^2$ the interaction between treatment and time-squared, $b_i \sim N(0, \sigma^2)$ the random effect associated with the i-th plot, $\gamma_{ik} \sim N(0, \tau^2)$ the random spline effect associated with the i-th plot and k-th knot ($K = \min\{26/4, 35\} \approx 6$, it means, $k = 1, \ldots, 6$) and $B_{k,2}(t_{ij})$ a 2 degree B-spline base function with knots located in the quantiles of the unique values of time as suggested in Section 2.3.3. We also assessed other bases and polynomial degrees but no improvements were observed. Results of fitting these models, using the SAS procedure GLIMMIX, are presented in Table 2.1.

The dispersion parameter for the NB model suggests overdispersion ($\phi = 0.2418$), but the variance function looks overestimated, i.e., this function describes a pattern above the majority of the observed points (Fig. 2.2.a). Hence, the PN model might represent a good alternative given that the variance function is also quadratic. However, this model presented a small estimate for the random effect variance ($\sigma^2 = 0.0107$), underestimating the variance function (Fig. 2.2.a). This probably happened because in this study, we worked with the sum of counts per trap, making it difficult to capture the correlation between measurements taken on the same trap.

The Prs model shows a clear improvement in terms of likelihood and also smaller AIC and BIC when compared with the PN model, which is usually the standard model for longitudinal data (Table 2.1). This suggests that not all variability in this study follows from the correlation between the repeated measures, but also from the nonlinearity between the outcome and the time covariate. The Prs model also shows a more realistic variance function when compared with the NB and the PN models. Although the Prs model looks reasonable, the fitted curve is too smooth when compared with the NBrs model. It happens because the smoothing parameter, Ω, in the first case, is the unique piece that is modeling the overdispersion, being inflated when compared with the second case that has both smoothing and dispersion parameters. Similar results are also found in Namata *et al.* (2007), where an increase in the number of knots provided a reduction

in the random spline variance component, generating larger values for the smoothing parameter.

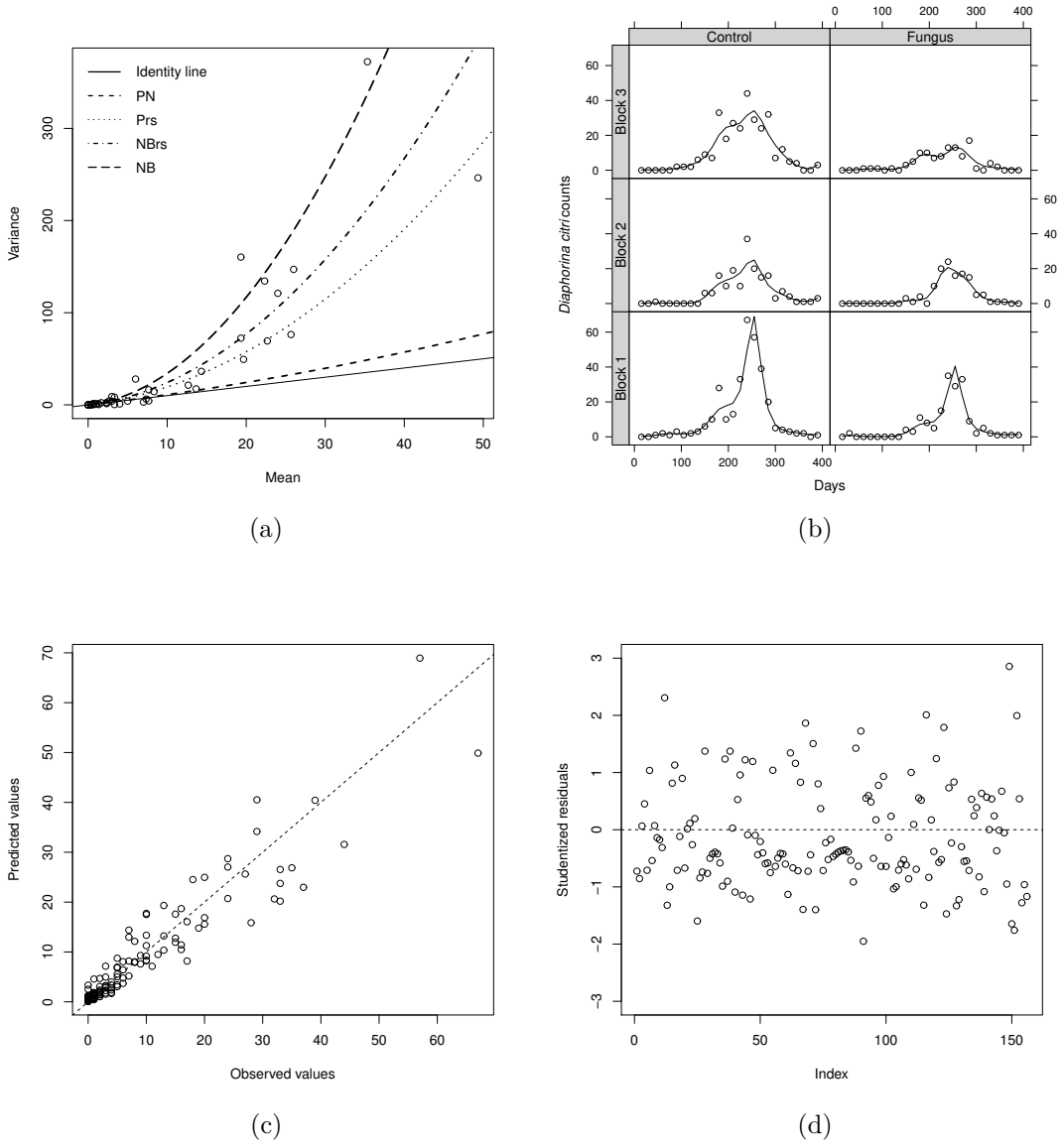
Thus, the NBrs model seems to explain better the insect fluctuations showing an improvement in terms of likelihood and the smallest AIC and BIC, modeling properly the variability of the data (Fig.2.2.b). Although the NBrs model has more parameters, the associated standard errors are bigger than the ones in the NB and PN models. Actually, this is not a problem since the omission of these features might reduce standard errors, but produce biased results.

Table 2.1: Parameter estimates (standard errors) for the coefficients in the negative binomial model (NB), the Poisson-normal model (PN), the Poisson model with random B-splines (Prs) and the negative binomial model with random B-splines (NBrs). Estimation was done by maximum likelihood using the Laplace approximation method over the normal random effect, if present.

Effect	Par.	NB	PN	Prs	NBrs
Intercept	β_0	1.0834(0.1602)	0.7329(0.1530)	1.1442(0.2947)	1.1681(0.2511)
Block 1	β_1	0.1527(0.1438)	0.2401(0.1268)	0.1779(0.3507)	0.1434(0.2926)
Block 2	β_2	-0.2952(0.1490)	-0.2120(0.1326)	-0.4023(0.3578)	-0.3860(0.2991)
Treat fungus	β_3	-1.0605(0.2583)	-1.1531(0.2533)	-0.8621(0.3414)	-0.8997(0.3118)
Time	β_4	8.5209(0.6912)	10.2425(0.5359)	7.5595(1.0164)	7.7075(0.9254)
Treat $\times$ time	β_5	1.7950(0.6912)	2.0767(1.0463)	0.4888(1.6494)	0.7715(1.5472)
Time^2	β_6	-7.5764(0.6151)	-9.0557(0.4700)	-6.7265(0.9049)	-6.8672(0.8231)
Treat $\times$ time^2	β_7	-1.3892(1.0628)	-1.5340(0.8996)	-0.4622(1.4537)	-0.6532(1.3523)
Disp. parameter	ϕ	0.2418(0.0584)	1	1	0.1421(0.0469)
Var. rand. intercept	σ^2	-	0.0107(0.0100)	-	-
Var. rand. spline	τ^2	-	-	0.7016(0.2711)	0.4075(0.2188)
Smooth paramater	ϕ/τ^2	-	-	1.4253	0.3487
-2loglik		689.40	805.53	716.94	**680.69**
AIC		707.40	823.53	734.94	**700.69**
BIC		734.85	821.65	733.06	**698.60**
No. of parameters		9	9	9	10

To evaluate the significance of the spline effect of NBrs model, we tested the hypothesis $H_0 : \tau^2 = 0$. Based on the work by Stram and Lee (1994); Self and Liang (1987); Molenberghs and Verbeke (2007), the likelihood ratio, score, and Wald tests do not follow the conventional asymptotic chi-squared null distribution in this case, but rather a mixture of chi-squared distributions. For a single variance parameter, this is a 50:50 mixture of a χ^2_0 (the degenerate chi-squared distribution in 0) and χ^2_1, often denoted as $\chi^2_{0:1}$. Thus, we obtain $p = P(\chi^2_{0:1} \geq 8.71) = 0.5P(\chi^2_0 \geq 8.71) + 0.5P(\chi^2_1 \geq 8.71) =$

0.0016, showing that the inclusion of the random spline function was important to model the data.



(a) (b)

(c) (d)

Figure 2.2: (a) Variance function for the NB, PN, Prs and NBrs models; (b) Observed data and fitted curves using the model NBrs; (c) Observed versus predicted values for NBrs model; (d) Studentized residuals for NBrs model.

We also assessed the need for a random spline function for each treatment in the NBrs model, that is equivalent to testing the hypotesis $H_0 : \tau_c^2 = \tau_f^2 = \tau^2$, where τ_c^2 and τ_f^2 are different random smoothing splines for the control and the treatment group, respectively. Using the standard likelihood ratio test we obtain $p = P(\chi_1^2 \geq 0.37) = 0.5430$ suggesting that both treatments can be modeled with the same smoothing function. It

suggests that the nonlinearity of the outcomes is not directly affected by the treatments but probably by environmental factors that did not differ too much between the plots. To assess the significance of treatment effect we tested the hypothesis $H_0 : \beta_3 = \beta_5 = \beta_7 = 0$. Using the standard likelihood ratio test we obtain $p = P(\chi^2_3 \geq 10.88) = 0.0123$, giving evidence of significant treatment effect. Therefore, the treatment effect changes the behavior of the fluctuations in population levels of *D. citri* over days.

Using the NBrs model, plots of observed versus predicted values and the studentized residuals versus indices were produced and are displayed, respectively, in Fig. 2.2.c and Fig. 2.2.d. Neither of these show anything particularly unusual, indicating the adequacy of the NBrs model. Thus, the negative binomial model with random B-splines provides a good fit to the data, giving evidence that the fungus *I. fumosorosea* ESALQ-1296 has the potential to reduce the fluctuations in population levels of *D. citri* in citrus orchards.

2.5 Concluding remarks

In this Chapter, generalized linear mixed models with random smoothing splines were explored and compared with some of the main standard models to analyze overdispersed longitudinal count data. We discussed the importance of modeling the variance function and also the nonlinearity in order to describe the biological variation properly.

Although the Poisson-normal model is a standard choice in longitudinal data analysis, it did not perform so well when compared to models with random smoothing splines. The mixed model representation of penalized splines allows one to take advantage of the existing methodology and the use of software packages. The greatest advantage of this technique is its flexibility, allowing to model behavior that fully parametric techniques would not be able to capture.

Our analyses indicated that the biological control using fungal pathogen *I. fumosorosea* ESALQ-1296 reduces the population levels of *D. citri*, suggesting that the fungus can be used for the management of this pest contributing to the sustainability of citriculture.

References

Beloti, V. H., G. R. Rugno, M. R. Felippe, A. Do Carmo-Uehara, L. F. Garbim, W. A. C. Godoy, and P. T. Yamamoto, 2013 Population Dynamics of Diaphorina citri Kuwayama (Hemiptera: Liviidae) in Orchards of 'Valencia' Orange, 'Ponkan' Mandarin and 'Murcott' Tangor Trees. Florida Entomologist **96**: 173–179.

Breslow, N. E. and D. G. Clayton, 1993 Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association **88**: 9–25.

Capanu, M., M. Gönen, and C. B. Begg, 2013 An assessment of estimation methods for generalized linear mixed models with binary outcomes. Statistics in Medicine **32**: 4550–4566.

Conceschi, M. R., 2017 *Parâmetros a serem considerados nas pulverizações do fungo Isaria fumosorosea para o manejo de Diaphorina citri*. Ph.d. diss., University of São Paulo.

Demétrio, C. G. B., J. Hinde, and R. A. Moral, 2014 Models for Overdispersed Data in Entomology. In *Ecological modeling applied to entomology*, Springer.

Demidenko, E., 2004 *Mixed models Theory and Applications*. Willey & Sons, Inc.

Durbán, M., J. Harezlak, M. P. Wand, and R. J. Carroll, 2005 Simple fitting of subject-specific curves for longitudinal data. Statistics in Medicine **24**: 1153–1167.

Faes, C., M. Aerts, H. Geys, L. Bijnens, L. Ver Donck, and W. J. Lammers, 2006 GLMM approach to study the spatial and temporal evolution of spikes in the small intestine. Statistical Modelling **6**: 300–320.

Gbur, E., 2012 *Analysis of generalized linear mixed models in agricultural and natural resources sciences*. Crop Science Society of America.

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

Grunwald, G. K., S. L. Bruce, L. Jiang, M. Strand, and N. Rabinovitch, 2011 A statistical model for under- or overdispersed clustered and longitudinal count data. Biometrical Journal **53**: 578–594.

Härdle, W., H. Liang, and J. Gao, 1999 *Partially Linear Models*. Springer-Verlang, New York.

Hastie, T. J. and R. J. Tibshirani, 1990 *Generalized Additive Models*. Chapman & Hall, London.

Hinde, J. and C. G. B. Demétrio, 1998 Overdispersion: Models and estimation. Computational Statistics and Data Analysis **27**: 151–170.

Joe, H., 2008 Accuracy of Laplace approximation for discrete response mixed models. Computational Statistics and Data Analysis **52**: 5066–5074.

Kobori, Y., T. Nakata, Y. Ohto, and F. Takasu, 2011 Dispersal of adult Asian citrus psyllid, Diaphorina citri Kuwayama (Homoptera: Psyllidae), the vector of citrus greening disease, in artificial release experiments. Applied Entomology and Zoology **46**: 27–30.

Lin, X. and D. Zhang, 1999 Inference in generalized additive mixed models by using smoothing splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**: 381–400.

Molenberghs, G. and G. Verbeke, 2005 *Models for discrete longitudinal data*. Springer-Verlang, New York.

Molenberghs, G. and G. Verbeke, 2007 Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space. The American Statistician **61**: 22–27.

Namata, H., Z. Shkedy, C. Faes, M. Aerts, G. Molenberghs, H. Theeten, P. Van Damme, and P. Beutels, 2007 Estimation of the force of infection from current status data using generalized linear mixed models. Journal of Applied Statistics **34**: 923–939.

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

Ngo, L. and M. P. Wand, 2004 Smoothing with Mixed Model Software. Journal of statistical software **9**: 1–54.

Patt, J. M. and M. Sétamou, 2010 Responses of the Asian Citrus Psyllid to Volatiles Emitted by the Flushing Shoots of Its Rutaceous Host Plants. Environmental Entomology **39**: 618–624.

Rodrigues, C. A., A. P. M. B. Battel, N. M. Martinelli, R. D. A. Moral, R. K. Sercundes, and W. A. C. Godoy, 2014 Dynamics and Predation Efficiency of Chrysoperla externa (Neuroptera: Chrysopidae) on Enneothrips flavens (Thysanoptera: Thripidae). Florida Entomologist **92**: 653–658.

Ruppert, D., M. Wand, and R. Carroll, 2003 *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics.

Self, S. G. and K. Y. Liang, 1987 Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. Journal of the American Statistical Association **82**: 605–610.

Speed, T., 1991 Comment on paper by Robinson. Statistical Science **6**: 42–44.

Stram, D. O. and J. W. Lee, 1994 Variance Components Testing in the Longitudinal Mixed Effects Model. Biometrics pp. 1171–1177.

Verbyla, A. P., B. R. Cullis, M. G. Kenward, and S. J. Welham, 1999 The analysis of designed experiments and longitudinal data using smoothing splines. Journal of the Royal Statistical Society Series C (Applied Statistics) **48**: 269–311.

Wahba, G., 1978 Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression. Journal of the Royal Statistical Society Series B **40**: 364–372.

Wand, M. P., 2003 Smoothing and mixed models. Computational Statistics **18**: 223–249.

Wolfinger, R. and M. O'Connell, 1993 Generalized linear mixed models a pseudo-likelihood approach. Journal of Statistical Computation and Simulation **48**: 233–243.

Yamamoto, P. T., P. E. Paiva, and S. Gravena, 2001 Flutuação Populacional de Diaphorina citri Kuwayama (Hemiptera: Psyllidae) em Pomares de Citros na Região Norte do Estado de São Paulo. Neotropical Entomology **30**: 165–170.

3 A COMBINED OVERDISPERSED LONGITUDINAL MODEL FOR NOMINAL DATA

Abstract: Longitudinal studies where the outcomes are nominal occur in several scientific areas. They are usually modeled with the generalized linear mixed model (GLMM). Widely used, this approach allows for the modeling of the hierarchy of the data and a certain amount of overdispersion. In this Chapter, a combined model (CM) that takes into account overdispersion and clustering through two separate sets of random effects is fomulated. The maximum likelihood method with analytic-numerical integration is used to estimate the model parameters. To examine the performance of the CM and the GLMM, simulation studies were conducted, exploring different scenarios of sample size, random effects, and overdispersion. Both models were also applied to an agricultural dataset and compared. Moreover, we show how to implement these models in the statistical software package SAS 9.4.
Keywords: Multinomial distribution; Beta distribution; Likelihood; Hierarchical data.

3.1 Introduction

There are many research fields such as medicine, marketing, education, and agriculture where nominal data are collected. A nominal outcome has its measurement scale formed by a set of categories that have no intrinsic order, being classified as binary, if only 2 categories are observed (e.g., dead or alive), or polytomous, if 3 or more categories are observed (e.g., political party affiliation: democrat, republican, or independent). Although polytomous response data are qualitative, all nominal outcomes may be written as a set of binary variables (Agresti, 2010; Hartzel *et al.*, 2001; Clayton, 1992). For cross-sectional studies, a whole collection of modeling approaches can be used, such as the generalized linear modeling (GLM) framework based on the exponential family of distributions (Nelder and Wedderburn, 1972).

One of the key features of the GLM framework, with exception of the Gaussian distribution, is the so-called mean-variance relationship, where the variance is a deterministic function of the mean. For example, for Bernoulli outcomes with success probability $\mu = \pi$, the variance is $\nu(\mu) = \pi(1 - \pi)$, which may be overly restrictive for real datasets with hierarchies. Scenarios where the mean is larger or smaller than the variance are reported in the literature as over- or underdispersion, respectively (Grunwald *et al.*, 2011; Demétrio *et al.*, 2014). For purely binary data, hierarchies need to be present in order to violate the mean-variance relationship (Molenberghs *et al.*, 2010, 2012, 2017). Thus,

studies where several measurements are taken from the same cluster, subject, or sample unit over time, characterizing a longitudinal study may violate this assumption.

Some of the main approaches used to analyze longitudinal data with nominal outcomes are generalized estimating equations (GEE) (Liang and Zeger, 1986; Lipsitz *et al.*, 1994; Touloumis *et al.*, 2013), transition models (TM) (Diggle *et al.*, 2002; Molenberghs and Verbeke, 2005; Lara *et al.*, 2017) and generalized linear mixed models (GLMM) (Hartzel *et al.*, 2001; Diggle *et al.*, 2002; Hedeker, 2003; Molenberghs and Verbeke, 2005). The first one is based on estimating the average response over a set of association parameters in a 'working' correlation matrix, the second one are models where any response within a sequence of repeated measures is modeled conditionally upon the previous outcomes and the third one is based on the assumption that, for every subject, the response can be expressed by a generalized linear model, but with subject-specific regression coefficients. Widely used, these approaches allow for the modeling of the hierarchy of the data and a certain amount of overdispersion. In this sense, Molenberghs and Verbeke (2007), Molenberghs *et al.* (2010), Molenberghs *et al.* (2012), Ivanova *et al.* (2014), and Molenberghs *et al.* (2017) showed that accommodating either overdispersion or hierarchically-induced association may fall short of properly modeling the data. Therefore, they proposed a so-called combined modeling framework encompassing both. Molenberghs and Verbeke (2007) focused on counts, Molenberghs *et al.* (2010) laid out a general framework, Molenberghs *et al.* (2012) worked with binary and binomial outcomes, Ivanova *et al.* (2014) tackles ordinal outcomes whereas Molenberghs *et al.* (2017) contributed with a review of all proposed combined models. The topic of the current research is the modeling of repeated, overdispersed nominal data. For this, we propose a combined model (CM) for nominal outcomes to model the hierarchy and the overdispersion using two different sets of random effects.

The Chapter is organized as follows.In Section 3.2, a motivating case study, stemming from an agricultural experiment on an elephant grass pasture and dairy cows is introduced. Basic ingredients for our modeling framework, standard generalized linear models, extensions for overdispersion, the generalized linear mixed model, and the combined model framework are the subject of Section 3.3. The proposed combined model is described in Section 3.4. Parameter estimation is the focus of Section 3.5. A simulation study comparing the proposed model and the GLMM is described and results presented in Section 3.6. The case study is analyzed in Section 3.7. Concluding remarks are offered in Section 3.8. Finally, we show in detail the algebraic development in Appendix B and how to implement these models in SAS 9.4 in Appendix C.

3.2 Grazing management dataset

The data used in this work results from an experiment on an elephant grass pasture (*Pennisetum purpureum* Schum. cv. Napier) grazed by dairy cows (Pereira *et al.*, 2015a,b). It was set up in a complete randomized block design with the treatments allocated according to a 2×2 factorial arrangement, where treatments are the combinations of two pre-grazing conditions (95% and maximum 98% canopy light interception during regrowth) and two post-grazing heights (35 and 45 cm). The experiment was carried out from January 2011 until April 2012, the period classified into six seasons due to climate conditions: summer1 (Jan-Mar), autumn (Apr-June), winter (July-Sept), early spring (Oct - mid-Nov), late spring (mid-Nov - Dec) and summer2 (Jan-Apr).

The response analyzed in the study is the type of vegetation observed in the field, which can be weeds, bare ground or tussocks. Forty (40) points were observed in each one of the four paddocks in each block. Since there are always 40 points observed in each paddock, we can analyze the proportions of each type of vegetation under the total, characterizing a multinomial outcome with three levels. There are $40 \times 16 = 640$ points per season, but in the early spring, one of the paddocks was affected by frost damage and thus the total number of observations was $N = 640 \times 6 - 40 = 3800$. A sample of the dataset and a sketch of the experiment are show in Table 3.1 and Figure 3.1, respectively.

Table 3.1: Sample of the grazing management dataset.

Seasons	Blocks	Pre-grazing	Post-grazing	ID point	Outcome*
Summer1	1	max	35	1	3
Autumn	1	max	35	1	1
Winter	1	max	35	1	2
Early spring	1	max	35	1	3
Late spring	1	max	35	1	1
Summer2	1	max	35	1	2
Summer1	1	max	35	2	3
Autumn	1	max	35	2	3
⋮	⋮	⋮	⋮	⋮	⋮
Summer1	4	95	45	640	2
Autumn	4	95	45	640	1
Winter	4	95	45	640	1
Early spring	4	95	45	640	3
Late spring	4	95	45	640	3
Summer2	4	95	45	640	3

* Where weed=1, bare ground=2 and tussock=3

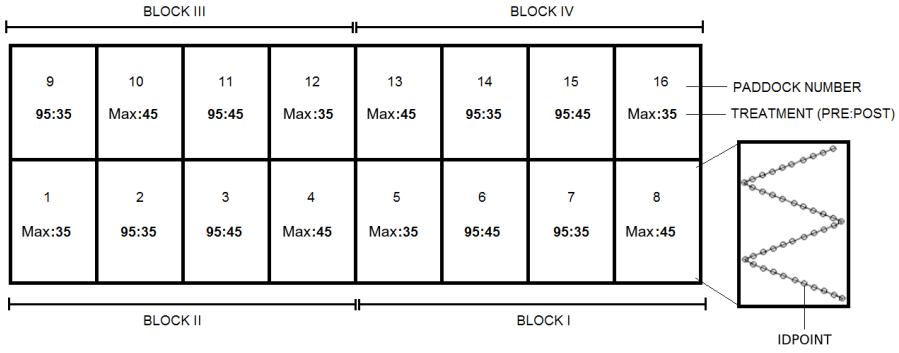


Figure 3.1: Sketch of the experiment carried out.

3.3 Building blocks

In this section, we briefly present the main concepts to formulate the combined model for nominal outcomes. In Section 3.3.1, we introduce the exponential family, generalized linear models and overdispersion. In Section 3.3.2, we present some properties of generalized linear mixed models and the general framework of combined models.

3.3.1 Generalized linear models and overdispersion

The class of generalized linear models (GLM) was introduced by Nelder and Wedderburn (1972) as a framework for handling a range of common statistical models for Gaussian and non-Gaussian data. A GLM is defined basically in terms of three components. The first is a set of independent random variables, $Y_1, \ldots, Y_N$, with probability or density function that belongs to the exponential family and that can be written as

$$f(y_i|\eta_i, \phi) \;=\; \exp\left\{\phi^{-1}[y_i\eta_i - \psi(\eta_i)] + c(y_i, \phi)\right\}, \tag{3.1}$$

where $\psi(\cdot)$ and $c(\cdot)$ are known functions and ϕ and η_i are called dispersion and natural or canonical parameter, respectively. The exponential family embraces several distributions, e.g., Gaussian, Bernoulli, binomial, Poisson, gamma and multinomial. The models belonging to the exponential family, (3.1), share some basic features, such as the first two moments being given by

$$\mathrm{E}(Y_i) = \mu_i = \psi'(\eta_i) \quad \text{and} \quad \mathrm{Var}(Y_i) = \sigma^2 = \phi\psi''(\eta_i).$$

Thus, the mean and variance of these distributions are related through $\sigma^2 = \phi\psi''[\psi'^{-1}(\mu)] = \phi v(\mu)$, with $v(\cdot)$ called variance function. For the Gaussian distribution, this relation is constant, that is $v(\mu) = 1$, and therefore, there is no relation between the mean and the variance. The second component, called linear predictor or natural

parameter, η_i, is the quantity that incorporates the information about the independent variables into the model and the third component, called link function, $h(\cdot)$, provides the relationship between the linear predictor and the mean of the distribution as $\mu_i = h(\eta_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$, where $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are vectors of covariates and fixed unknown regression coefficients, respectively.

For polytomous outcomes, the multinomial distribution is usually a natural starting point for data analysis. This distribution arises as a natural extension of the binomial distribution when each independent trial has more than two possible mutually exclusive outcomes. Consider a series of m independent trials of an experiment, each resulting in one of the R mutually exclusive events $\mathrm{E}_1, \ldots, \mathrm{E}_R$. In each replicate within the experiment, the probability of the occurrence of event E_r is equal to π_r with $\sum_{r=1}^{R} \pi_r = 1$. Let $\boldsymbol{Y}^* = (Y_1, \ldots, Y_R)^T$ denote the random vector of the number of occurrences of events $\mathrm{E}_1, \ldots, \mathrm{E}_R$ out of m trials, with $\sum_{r=1}^{R} Y_r = m$. Let $\boldsymbol{y}^* = (y_1, \ldots, y_R)^T$ represent a realization of $\boldsymbol{Y}^*$, $\sum_{r=1}^{R} y_r = m$. Then, the random vector $\boldsymbol{Y}^*$ is said to have a multinomial distribution with parameters m, $\boldsymbol{\pi}^* = (\pi_1, \ldots, \pi_R)^T$, and joint probability function given by

$$P(\boldsymbol{Y}^* = \boldsymbol{y}^*) = \frac{m!}{y_1!y_2!\ldots y_R!}\pi_1^{y_1}\pi_2^{y_2}\ldots\pi_R^{y_R} \quad \pi_i \in [0,1], \quad i = 1, \ldots, R.. \tag{3.2}$$

We can safely reduce the dimensionality of $\boldsymbol{Y}^*$ and $\boldsymbol{\pi}^*$ by removing their respective last elements. Then, define $\boldsymbol{Y} = (Y_1, \ldots, Y_{R-1})^T$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{R-1})^T$ and the realization of $\boldsymbol{Y}$ as $\boldsymbol{y} = (y_1, \ldots, y_{R-1})$. Thus, without loss of generality, we say that $\boldsymbol{Y}$ has multinomial distribution with parameters m and $\boldsymbol{\pi}$ with joint probability function as in equation (3.2) with $y_R = m - \sum_{h=1}^{R-1} y_h$ and $\pi_R = 1 - \sum_{h=1}^{R-1} \pi_h$. Hence, the mean and the variance of $\boldsymbol{Y}$ are, respectively

$$\mathrm{E}(\boldsymbol{Y}) = m\boldsymbol{\pi} \quad \text{and} \quad \mathrm{Var}(\boldsymbol{Y}) = m\Delta(\boldsymbol{\pi}), \tag{3.3}$$

where $\Delta(\boldsymbol{\pi}) = \mathrm{diag}\,(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$. Note that $\Delta(\boldsymbol{\pi})$ is an $R \times R$ variance-covariance matrix of full rank where the diagonal elements are $\pi_r(1 - \pi_r)$ and off-diagonal elements $-\pi_r\pi_{r'}$ for $r \neq r'$. It is well known that (3.3) implies a restrictive variance function. According to Grunwald *et al.* (2011) and Demétrio *et al.* (2014), cases where the variance is greater than the mean are largely reported in the literature as overdispersion, which may occur due to the absence of relevant covariates, heterogeneity of sampling units, hierarchical structures and excess of zeros. It should be noted that underdispersion can occur as well. Thus, it is important to adapt models to take into account this feature in order to avoid incorrect inferences (Hinde and Demétrio, 1998).

An elegant way to extend the multinomial model to handle overdispersion is to multiply the multinomial covariance matrix by a constant scalar parameter. A quasi-likelihood approach using a scale adjustment is presented by P and NELDER (1983), being later extended by MOREL and KOEHLER (1995) in order to allow for different levels of overdispersion for each category using a diagonal matrix of overdispersion parameters and a Cholesky decomposition of the multinomial variance-covariance matrix. Mixture of distributions can also be used to allow for overdispersion parameters such as the random-clumped multinomial distribution proposed by MOREL and NAGARAJ (1993) and NEERCHAL and MOREL (1998). This model is a finite mixture of multinomial distributions that captures the extra variation caused by clumped sampling. Another convenient route to tackle overdispersion is through the so-called two-stage approach, which considers a distribution for the model parameter. Thus, some models for the multinomial distribution were proposed such as the Dirichlet-Multinomial model, where the parameter $\boldsymbol{\pi}$ follows a Dirichlet distribution (MOSIMANN, 1962). Although the estimation of a dispersion parameter might provide some flexibility to the standard GLMs, this is not always sufficient, especially when hierarchical structures or highly variable data arise.

3.3.2 Generalized linear mixed models and combined models

When non-Gaussian data are hierarchically organized (repeated measures or clustering, for example), the theory of the so-called generalized linear mixed model (GLMM) is well-known in the literature (MOLENBERGHS and VERBEKE, 2005; DIGGLE *et al.*, 2002). In full generality, one assumes that, conditionally on q-dimensional random effects $\boldsymbol{b}_i$, assumed to be drawn independently from a normal distribution, $N_q(\boldsymbol{0}, \boldsymbol{D})$, the outcomes Y_{ij} measured on the i-th subject or sample unit at j-th time point ($i = 1, \ldots, N; j = 1, \ldots, n_i$) are independent with densities of the form

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\eta_{ij} - \psi(\eta_{ij})] + c(y_{ij}, \phi)\right\}, \tag{3.4}$$

where $\eta(\mu_{ij}) = \eta[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i)] = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$, with $\boldsymbol{x}_{ij}$ ($\boldsymbol{z}_{ij}$) the design vector for the fixed (random) effects. Finally, let $f(\boldsymbol{b}_i|\boldsymbol{D})$ be the density of the Gaussian distribution, $N(\boldsymbol{0}, \boldsymbol{D})$, for the random effects $\boldsymbol{b}_i$.

For nominal data, it is assumed that the outcome Y_{ij} can take values $r = 1, \ldots, R$. Without loss of generality, we can replace it by a set of R dummy variables where $W_{r,ij}$ is equal to 1 if $Y_{ij} = r$ and 0 otherwise. Evidently, there are redundant dummies, but any subset of $R-1$ components is not, as described in Section 3.3.1. Thus, $\boldsymbol{W}_{ij} \sim$ multinomial($\boldsymbol{\pi}_{ij}$) with probabilities $\boldsymbol{\pi}_{ij} = (\pi_{1,ij}, \ldots, \pi_{r,ij}, \ldots, \pi_{R,ij})^T$. Assuming that category R is the ref-

erence category, a baseline-category logit model (Agresti, 2010; Hartzel *et al.*, 2001) can be written as

$$\begin{aligned} \ln\left(\frac{\pi_{r,ij}}{\pi_{R,ij}}\right) &= \eta_{r,ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_r + \boldsymbol{z}_{ij}^T\boldsymbol{b}_{r,i}, \quad r = 1,\ldots,R-1, \\ \boldsymbol{b}_{r,i} &\sim N(\boldsymbol{0}, \boldsymbol{D}), \end{aligned} \tag{3.5}$$

where $\boldsymbol{\beta_r}$ is the fixed-effects coefficient vector of length $p+1$, corresponding to an intercept and p covariates, and $\boldsymbol{b}_{r,i}$ is the random-effects coefficient vector that follows a multivariate normal distribution. The probabilities of each category in the i-th subject and j-th time might be expressed as

$$\pi_{r,ij} = \begin{cases} \frac{\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta_r}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_{r,i})}{1+\sum_{h=1}^{R-1}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta_h}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_{h,i})} & \text{if} \quad 1 \leq r \leq R-1, \\ 1-\sum_{h=1}^{R-1}\pi_{h,ij} & \text{if} \quad r = R. \end{cases}$$

Estimates of $\boldsymbol{\beta}$, $\boldsymbol{D}$ and ϕ for GLMMs are obtained from maximizing the marginal likelihood, obtained by integrating out the random effects and commonly written as:

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{D}, \phi) = \prod_{i=1}^{N}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i} f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{b}_i|\boldsymbol{D}) d\boldsymbol{b}_i. \tag{3.6}$$

The key problem in maximizing (3.6) is the presence of N integrals over the random effects. Thus, numerical methods are needed, such as adaptive Gaussian quadrature (Molenberghs and Verbeke, 2005; Pinheiro and Bates, 1995) that uses the same weights and nodes as the Gauss-Hermite quadrature, but centers the nodes with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode to increase efficiency.

Although widely used, GLMMs are defined to accommodate correlation, but in practice, both overdispersion and correlation can occur simultaneously, which led Molenberghs and Verbeke (2007) to formulate a flexible and unified modeling framework, which they termed the combined model. These authors brought together two sets of random effects: the normally distributed subject-specific random effects to capture correlation and a certain amount of overdispersion, and a conjugate measurement-specific random effect on the natural parameter scale to accommodate the remaining overdispersion. Integrating out these two sets of random effects and using the generalized linear model framework, the following general family is introduced:

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \theta_{ij}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij}, \phi)\right\}, \tag{3.7}$$

with notation similar to the one used in (3.4), but now with conditional mean

$$\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \theta_{ij}) = \mu^c_{ij} = \theta_{ij}\kappa_{ij}, \tag{3.8}$$

where $\theta_{ij} \sim \mathcal{G}_{ij}(\vartheta_{ij}, \xi_{ij})$ is a conjugate random variable and $\kappa_{ij} = g(\boldsymbol{x}^T_{ij}\boldsymbol{\beta} + \boldsymbol{z}^T_{ij}\boldsymbol{b}_i)$. Finally, as before, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$. Unlike in GLMM, we now have two different notations, η_{ij} and λ_{ij}, to refer to the linear predictor and/or the natural parameter. The reason is that λ_{ij} encompasses the random variables θ_{ij}, whereas η_{ij} refers to the 'GLMM part' only. Regarding the components $\boldsymbol{\theta}_i$ and $\boldsymbol{b}_i$, three useful assumptions can be made: (1) they are independent; (2) they are correlated, implying that the collection of univariate distributions $\mathcal{G}_{ij}(\vartheta_{ij}, \xi_{ij})$ needs to be replaced with a multivariate one; and (3) they are equal to each other, useful in applications with exchangeable outcomes Y_{ij}. Obviously, parameterization (3.8) allows for random effects θ_{ij} capturing overdispersion, and formulated directly at the mean scale. The relationship between the mean and the natural parameter now is

$$\lambda_{ij} = h(\mu^c_{ij}) = h(\theta_{ij}\kappa_{ij}).$$

We can still apply standard GLM ideas to derive the mean and variance, combined with iterated expectation-based calculations. For the mean, if $\boldsymbol{\theta}_i$ and $\boldsymbol{b}_i$ are independent, it follows that

$$\mathrm{E}(Y_{ij}) = \mathrm{E}(\theta_{ij})\mathrm{E}(\kappa_{ij}) = \mathrm{E}[h^{-1}(\lambda_{ij})].$$

The work of Molenberghs *et al.* (2010) and Molenberghs *et al.* (2017) derived explicit expressions for the means, variances, and marginal densities in a number of outcome types, such as Gaussian, Poisson, and time-to-event data. Unfortunately, this is not possible for binary data modeled with a logit link and including Gaussian random effects, whether or not other random effects are present.

3.4 Combined model for nominal outcomes

In analogy with Ivanova *et al.* (2014), but using a baseline-category logit structure (3.5), with Gaussian random effects, $\boldsymbol{b}_{r,i} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, in the linear predictor, and beta random effects, $\theta_{ij} \sim \mathrm{Beta}(\vartheta, \xi)$, to capture the overdispersion (considering θ_{ij} and $\boldsymbol{b}_{r,i}$ independent), the probabilities of the proposed combined model may be written as:

$$\pi_{r,ij} = \begin{cases} \theta_{ij}\kappa_{r,ij} & \text{if} \quad 1 \leq r \leq R-1, \\ \\ 1 - \sum_{h=1}^{R-1} \theta_{ij}\kappa_{h,ij} & \text{if} \quad r = R, \end{cases}$$

and

$$\kappa_{r,ij} = \frac{\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta_r} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_{r,i})}{1 + \sum_{h=1}^{R-1}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta_h} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_{h,i})} \quad \text{if } 1 \leq r \leq R-1, \tag{3.9}$$

where $\boldsymbol{\beta_r}$ is the fixed regression coefficients for each one of the $(R-1)$ categories and $\boldsymbol{x}_{ij}$ ($\boldsymbol{z}_{ij}$) is the design vector for the fixed (random) effects. We considered here the case where θ_{ij} is constant across all categories, giving rise to a combined model with constant overdispersion.

3.5 Parameter estimation

Molenberghs and Verbeke (2007) and Molenberghs *et al.* (2010) showed that fitting the combined model is relatively easy, and that standard software tools can be used for maximum likelihood estimation in this case. A priori, fitting a combined model of the type described in Section 3.4 proceeds by integrating over the random effects. The joint distribution of the ij-th observation, assuming θ_{ij} and $\boldsymbol{b}_{r,i}$ independent, is given by

$$f(w_{ij}, \boldsymbol{b}_{r,i}, \theta_{ij}) = f(w_{ij}|\boldsymbol{b}_{r,i}, \theta_{ij})\ f(\boldsymbol{b}_{r,i})\ f(\theta_{ij}),$$

and the likelihood function can be derived as:

$$L(\boldsymbol{\beta}, \boldsymbol{D}, \vartheta, \xi) \ = \ \prod_{i=1}^{N}\int\int\prod_{j=1}^{n_i} f(w_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_{r,i}, \theta_{ij})\ f(\boldsymbol{b}_{r,i}|\boldsymbol{D})\ f(\theta_{ij}|\vartheta, \xi)\ d\boldsymbol{b}_{r,i}\ d\theta_{ij}.$$

For our proposed model, these functions are respectively, the multinomial, the normal and the beta distributions, and hence

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{D}, \vartheta, \xi) \ = \ & \prod_{i=1}^{N}\int\int\prod_{j=1}^{n_i}\prod_{h=1}^{R-1}(\theta_{ij}\kappa_{h,ij})^{w_{h,ij}} \left(1 - \sum_{h=1}^{R-1}\theta_{ij}\kappa_{h,ij}\right)^{1-\sum_{h=1}^{R-1}w_{h,ij}} \\ & \frac{1}{\sqrt{(2\pi)^{n_i}}}\frac{1}{\sqrt{|\boldsymbol{D}|}}\exp\left(-\frac{1}{2}\boldsymbol{b}_{r,i}^T\boldsymbol{D}^{-1}\boldsymbol{b}_{r,i}\right)\ \frac{\theta_{ij}^{\vartheta-1}(1-\theta_{ij})^{\xi-1}}{B(\vartheta,\xi)}\ d\boldsymbol{b}_{r,i}\ d\theta_{ij}. \end{aligned} \tag{3.10}$$

The key problem in maximizing (3.10) is the presence of N integrals over the random effects $\boldsymbol{b}_{r,i}$ and θ_{ij}, making this process time consuming and cumbersome to implement. However, we can make it simpler by integrating analytically over the beta random effects, leaving the normal random effects untouched, leading to a so-called partially marginalized density. In our case, this takes the form (details of calculations are

presented in Appendix B):

$$L(\boldsymbol{\beta}, \boldsymbol{D}, \vartheta, \xi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \prod_{h=1}^{R-1} \left(\frac{\vartheta}{\vartheta + \xi} \kappa_{h,ij} \right)^{w_{h,ij}} \left(1 - \frac{\vartheta}{\vartheta + \xi} \sum_{h=1}^{R-1} \kappa_{h,ij} \right)^{1 - \sum_{h=1}^{R-1} w_{h,ij}} \frac{1}{\sqrt{(2\pi)^{n_i}}} \frac{1}{\sqrt{|\boldsymbol{D}|}} \exp\left(-\frac{1}{2} \boldsymbol{b}_{r,i}^T \boldsymbol{D}^{-1} \boldsymbol{b}_{r,i} \right) d\boldsymbol{b}_{r,i}.$$

Then, a generic maximum likelihood routine that allows for integration over normal random effects can be used. We follow this route and use the SAS procedure NLMIXED to do this. We used the adaptive Gaussian quadrature method, which is more accurate than ordinary Gaussian quadrature (Molenberghs and Verbeke, 2005) and we chose the number Q of quadrature points by performing a numerical sensitivity analysis to check whether Q was sufficiently large. To ensure identifiability, a constraint needs to be applied, e.g., $\vartheta = e^{\delta}$ and $\xi = 1$, but it is mathematically convenient to retain them as two separate parameters, with the understanding that the constraint does apply. Thus, if the δ parameter grows large, this corresponds to diminishing overdispersion.

3.6 Simulation

A simulation study was conducted to compare the performance of GLMM and the proposed combined model. For this, nominal longitudinal data were simulated considering a baseline-category logit model, (3.5), and the following structure for the linear predictor:

$$\eta_{r,ij} = b_{r,i} + \beta_{r,0} + \beta_{r,1}\text{time}_{ij} + \beta_{r,2}\text{group}_i + \beta_{r,3}\text{time}_{ij} * \text{group}_i,$$

$$\begin{aligned} \boldsymbol{\beta}_1 &= (\beta_{1,0}; \beta_{1,1}; \beta_{1,2}; \beta_{1,3})^T = (0.1; 0.2; 0.5; 0.7)^T, \\ \boldsymbol{\beta}_2 &= (\beta_{2,0}; \beta_{2,1}; \beta_{2,2}; \beta_{2,3})^T = (0.2; 0.1; 0.4; 0.5)^T, \end{aligned}$$

$$\boldsymbol{b}_{r,i} \sim N(\boldsymbol{0}, \boldsymbol{D}),$$

$$\boldsymbol{D} = \begin{pmatrix} d_1 & c \\ c & d_2 \end{pmatrix},$$

where $\text{time}_{ij} = (t-1)/6$ for $t = 1, \ldots, 6$, $\text{group}_i = 0$ or 1 and $c = 0.5$. We simulated 200 datasets with sample sizes of 300 and 600. Both groups were simulated to be equal in size. Six scenarios with different magnitudes of random effects and overdispersion were generated to compare the behavior of the GLMM and the CM (Table 3.2). For the random

effects, we considered scenarios where the numerical values of these components were set as $d_1 = 1.0$ or 9.0 and $d_2 = 0.5$ or 4.5, attempting in this sense to generate 'weak' and 'strong' (nine times higher) values of correlation for the simulated datasets.

Table 3.2: Simulated scenarios for 200 datasets.

Scenario	Size	Variance components	Overdispersion
S1	$N = 300$ and 600	$d_1 = 1.0$, $d_2 = 0.5$	-
S2		$d_1 = 9.0$, $d_2 = 4.5$	-
S3		$d_1 = 1.0$, $d_2 = 0.5$	$\vartheta = 5$, $\xi = 1$
S4		$d_1 = 1.0$, $d_2 = 0.5$	$\vartheta = 20$, $\xi = 1$
S5		$d_1 = 9.0$, $d_2 = 4.5$	$\vartheta = 5$, $\xi = 1$
S6		$d_1 = 9.0$, $d_2 = 4.5$	$\vartheta = 20$, $\xi = 1$

To generate scenarios with overdispersion, the simulated probabilities were multiplied by values generated from a beta distribution where the shape parameters were set as $\vartheta = 5$ or 20 and $\xi = 1$. Thus, we generated values to disturb in a 'strong' or 'weak' way the probabilities generated by the model, respectively (Table 3.3).

Table 3.3: Simulated overdispersion values.

Beta parameters	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
$\vartheta = 5$ and $\xi = 1$	0.06	0.76	0.87	0.83	0 94	1
$\vartheta = 20$ and $\xi = 1$	0.48	0.93	0.97	0.95	0.98	1

The GLMM and CM were fitted to the simulated datasets using the estimation method described in Section 3.5, which was implemented in the SAS procedure NLMIXED, together with adaptive Gaussian quadrature with 10 quadrature points. Optimization took place using the quasi-Newton BFGS method. The GLM model parameters were used as starting values. For each scenario, we evaluated the maximum likelihood estimates of the parameters and then we determined the average estimates (AEs), biases and mean squared errors (MSEs). Because of identifiability, we set $\vartheta = e^{\delta}$ and $\xi = 1$ for the combined model.

In general, the simulation results indicate appropriate behavior of the models showing that the MSEs of the maximum likelihood estimators of the parameters decay towards zero as the sample size increases, as expected under standard asymptotic theory. For scenarios 1, 2, and 4, both models performed similarly, suggesting that if the dataset has just 'small' effects of correlation, overdispersion or both, the GLMM and the CM converge to analogous results (Table 3.4). It is also observed that the overdispersion parameter δ grows above the value 6 in these scenarios, correctly indicating small effects of overdispersion, once $\vartheta/(\vartheta + \xi) = \exp(\delta)/(\exp(\delta) + 1) \approx 0.98$.

Table 3.4: Results of average estimates (AE), biases and mean square errors (MSE) for the GLMM and CM based on 200 simulations for scenario S2.

		GLMM					
		$N = 300$			$N = 600$		
Parameter	True	AE	Bias	MSE	AE	Bias	MSE
$\beta_{1,0}$	0.1	0.123	0.023	0.118	0.112	0.012	0.057
$\beta_{1,1}$	0.2	0.196	-0.004	0.185	0.219	0.019	0.076
$\beta_{1,2}$	0.5	0.524	0.024	0.242	0.463	-0.037	0.127
$\beta_{1,3}$	0.7	0.689	-0.011	0.351	0.692	-0.008	0.153
$\beta_{2,0}$	0.2	0.233	0.033	0.086	0.193	-0.007	0.039
$\beta_{2,1}$	0.1	0.072	-0.028	0.144	0.149	0.049	0.061
$\beta_{2,2}$	0.4	0.402	0.002	0.130	0.428	0.028	0.077
$\beta_{2,3}$	0.5	0.478	-0.022	0.297	0.458	-0.042	0.151
d_1	9.0	9.220	0.220	2.921	9.055	0.055	1.303
d_2	4.5	4.558	0.058	0.682	4.462	-0.038	0.307
c	0.5	0.616	0.116	0.635	0.570	0.070	0.341
		CM					
		$N = 300$			$N = 600$		
Parameter	True	AE	Bias	MSE	AE	Bias	MSE
$\beta_{1,0}$	0.1	0.136	0.036	0.128	0.119	0.019	0.059
$\beta_{1,1}$	0.2	0.195	-0.005	0.187	0.221	0.021	0.077
$\beta_{1,2}$	0.5	0.572	0.027	0.244	0.466	-0.034	0.127
$\beta_{1,3}$	0.7	0.694	-0.006	0.355	0.694	-0.006	0.155
$\beta_{2,0}$	0.2	0.245	0.045	0.090	0.200	0.001	0.039
$\beta_{2,1}$	0.1	0.072	-0.028	0.145	0.150	0.050	0.062
$\beta_{2,2}$	0.4	0.405	0.005	0.132	0.431	0.031	0.078
$\beta_{2,3}$	0.5	0.483	-0.017	0.300	0.459	-0.041	0.153
d_1	9.0	9.288	0.288	2.747	9.110	0.110	1.364
d_2	4.5	4.589	0.089	0.687	4.481	-0.019	0.323
c	0.5	0.654	0.154	0.650	0.600	0.100	0.346
δ	-	9.482	-	-	9.188	-	-

Table 3.5: Results of average estimates (AE), biases and mean square errors (MSE) for the GLMM and CM based on 200 simulations for scenario S6.

		GLMM					
		$N = 300$			$N = 600$		
Parameter	True	AE	Bias	MSE	AE	Bias	MSE
$\beta_{1,0}$	0.1	-0.251	-0.351	0.215	-0.276	-0.376	0.182
$\beta_{1,1}$	0.2	0.149	-0.051	0.130	0.180	-0.020	0.072
$\beta_{1,2}$	0.5	0.380	-0.120	0.188	0.376	-0.124	0.105
$\beta_{1,3}$	0.7	0.524	-0.176	0.310	0.501	-0.199	0.201
$\beta_{2,0}$	0.2	-0.207	-0.407	0.228	-0.228	-0.428	0.220
$\beta_{2,1}$	0.1	0.068	-0.032	0.128	0.092	-0.008	0.064
$\beta_{2,2}$	0.4	0.293	-0.107	0.145	0.335	-0.065	0.064
$\beta_{2,3}$	0.5	0.353	-0.147	0.282	0.339	-0.161	0.156
d_1	9.0	6.297	-2.703	8.722	6.294	-2.706	7.916
d_2	4.5	3.509	-0.991	1.419	3.517	-0.983	1.153
c	0.5	-0.622	-1.122	1.559	-0.639	-1.139	1.455
		CM					
		$N = 300$			$N = 600$		
Parameter	True	AE	Bias	MSE	AE	Bias	MSE
$\beta_{1,0}$	0.1	0.115	0.015	0.169	0.058	-0.042	0.074
$\beta_{1,1}$	0.2	0.186	0.014	0.147	0.215	0.015	0.070
$\beta_{1,2}$	0.5	0.507	0.007	0.178	0.474	-0.026	0.107
$\beta_{1,3}$	0.7	0.688	-0.012	0.329	0.669	-0.031	0.199
$\beta_{2,0}$	0.2	0.201	0.001	0.117	0.141	-0.059	0.068
$\beta_{2,1}$	0.1	0.091	-0.009	0.122	0.117	0.017	0.054
$\beta_{2,2}$	0.4	0.398	-0.002	0.154	0.419	0.019	0.070
$\beta_{2,3}$	0.5	0.495	-0.005	0.284	0.483	-0.017	0.157
d_1	9.0	8.966	-0.034	4.965	8.636	-0.364	2.158
d_2	4.5	4.435	-0.065	1.056	4.344	-0.156	0.466
c	0.5	0.547	0.047	1.137	0.370	-0.130	0.505
δ	-	3.153	-	-	3.172	-	-

However, if there is a pronounced effect of overdispersion (S3) or if an overdispersion effect is associated with high correlations (S5 and S6), better performances were observed for the CM, mainly for the variance components (Table 3.5). Even increasing the sample size, the predicted random effects for the CM showed smaller values of bias and MSE than GLMM. Thus, there is evidence that the CM correctly models the simulated overdispersion.

From Table 3.6, we show the convergence rate for the two models when 200 datasets were simulated. It was observed that the proportion of converging sets is lower in the CM than in the GLMM. This can be attributed to sensitivity to the starting values. In

Table 3.6: Convergence rates for the GLMM and the CM in six simulated scenarios with 200 datasets

			Scenario					
	Model	Size	S1	S2	S3	S4	S5	S6
Rate (%)	GLMM	300	98.5	100.0	97.3	98.5	100.0	100.0
		600	100.0	100.0	99.5	100.0	100.0	100.0
	CM	300	97.5	100.0	91.0	95.5	100.0	100.0
		600	100.0	100.0	99.0	100.0	100.0	99.5

practice, this points to the need for carefully selecting starting values. We suggest, when convergence problems arise, to start the analysis with the GLM or GLMM estimates and, if necessary, to start from various starting values.

3.7 Analysis of the grazing management data

We analyzed the grazing management data, introduced in Section 3.2. Note that the data were analyzed before in Menarin and Lara (2017), using extended generalized estimating equations that use local odds ratios to explain the dependence among the categories (Touloumis *et al.*, 2013). Here, emphasis was placed on conditional interpretation of the models. Let $Y_{ij} = 1, 2, 3$ be the types of vegetation (weed, bare ground and tussock, respectively) that target point i, $(i = 1, \ldots, 640)$, reached at season j, $(j = 1, \ldots, 6)$. Thus, under a baseline-category logit model, (3.5), the GLMM and the CM can be written as:

$$\text{logit 1:} \;\; = \ln\left(\frac{\pi_{1,ij}}{\pi_{3,ij}}\right), \quad \text{logit 2:} \;\; = \ln\left(\frac{\pi_{2,ij}}{\pi_{3,ij}}\right),$$

where $\pi_{r,ij}$ is the probability of the i-th point being classified in the r-th category in season j. The first logit is a log-odds between weeds and tussocks and the second logit is the log-odds between bare grounds and tussocks. As before, both models were fitted with SAS procedure NLMIXED using adaptive Gaussian quadrature. We performed a sensitivity analysis increasing the number of quadrature points up to 5, when the estimates showed stability, being the maximization made by quasi-Newton BFGS method. The fixed and random effects were selected using backward selection, starting first with the random

effects. The selected linear predictor was given by:

$$\begin{aligned}\eta_{r,ij} &= b_i + \beta_{r,0} + \overbrace{\beta_{r,1}X_{11i} + \beta_{r,2}X_{12i} + \beta_{r,3}X_{13i}}^{\text{blocks}} + \overbrace{\beta_{r,4}X_{2i}}^{\text{pre-grazing}} \\ &+ \overbrace{\beta_{r,5}X_{31i} + \beta_{r,6}X_{32i} + \beta_{r,7}X_{33i} + \beta_{r,8}X_{34i} + \beta_{r,9}X_{35i}}^{\text{seasons}} \\ &+ \overbrace{\beta_{r,10}X_{2i}X_{31i} + \ldots + \beta_{r,14}X_{2i}X_{35i}}^{\text{pre} \times \text{seasons}},\end{aligned}$$

where $b_i \sim N(0, d)$, and X's are dummy covariates for blocks $(X_{11i}, \ldots, X_{13i})$, pre-grazing management (X_{2i}) and seasons $(X_{31i}, \ldots, X_{35i})$. To ensure identifiability, we take the last level of each covariate as reference (Block4=0, Pre:100=0 and Summer2=0) and for the CM we set $\vartheta = e^{\delta}$ and $\xi = 1$.

Table 3.7: Grazing management data. Parameter estimates (standard errors) from the regression coefficients in the GLMM and CM. Estimation was done by maximum likelihood using numerical integration over the normal and beta random effects, if present.

		GLMM		CM	
Effects	Par.	logit 1	logit 2	logit 1	logit 2
Intercept	$\beta_{r,0}$	-2.826(0.332)	-0.053(0.129)	-2.467(0.392)	0.325(0.247)
Block 1	$\beta_{r,1}$	0.623(0.190)	0.117(0.099)	0.736(0.198)	0.220(0.129)
Block 2	$\beta_{r,2}$	0.567(0.190)	-0.024(0.098)	0.683(0.193)	0.032(0.124)
Block 3	$\beta_{r,3}$	-0.681(0.252)	0.072(0.097)	-0.633(0.248)	0.169(0.120)
Pre(95%)	$\beta_{r,4}$	0.725(0.364)	-0.605(0.168)	0.936(0.379)	-0.602(0.209)
Summer1	$\beta_{r,5}$	0.514(0.370)	-0.810(0.170)	0.580(0.390)	-0.677(0.211)
Autumn	$\beta_{r,6}$	-0.309(0.444)	-0.345(0.162)	-0.199(0.453)	-0.342(0.205)
Winter	$\beta_{r,7}$	-0.136(0.426)	-0.364(0.163)	-0.044(0.438)	-0.352(0.206)
Early spring	$\beta_{r,8}$	0.308(0.403)	-0.286(0.169)	0.286(0.432)	-0.217(0.216)
Late spring	$\beta_{r,9}$	0.969(0.365)	-0.082(0.163)	1.029(0.395)	-0.023(0.218)
Pre(95%) × summer1	$\beta_{r,10}$	-0.389(0.464)	0.898(0.243)	-0.347(0.487)	0.930(0.305)
Pre(95%) × autumn	$\beta_{r,11}$	0.346(0.525)	0.279(0.237)	0.397(0.532)	0.274(0.295)
Pre(95%) × winter	$\beta_{r,12}$	-0.701(0.551)	0.429(0.236)	-0.633(0.547)	0.397(0.290)
Pre(95%) × early spring	$\beta_{r,13}$	-0.670(0.504)	0.120(0.243)	-0.697(0.530)	0.133(0.300)
Pre(95%) × late spring	$\beta_{r,14}$	-1.678(0.496)	0.147(0.237)	-1.632(0.513)	0.132(0.301)
Random effect	d	0.020(0.042)		0.013(0.062)	
Overdispersion	δ	-		1.374(0.437)	
-2loglik		6602.7		**6590.9**	
AIC		6664.7		**6654.9**	
BIC		6803.0		**6797.7**	

The results of both models are presented in Table 3.7. The estimates are very similar, but there is a reduction in the variance component for the CM, a pronounced value

of the overdispersion parameter ($\delta = 1.374$) and also a clear improvement in terms of the likelihood. To compare the models, the likelihood ratio test can be used. The difference in the deviance is 11.8, however, care has to be taken when comparing such models because of the special status of the variance components. Based on the work by Stram and Lee (1994); Self and Liang (1987), the likelihood ratio test statistic does not follow asymptotically the conventional chi-squared null distribution in this case, but rather a mixture of chi-squared distributions. For the hypothesis $H_0 : \theta_{ij} = 1$, this is a 50:50 mixture of a χ^2_0 (the degenerate chi-squared distribution at 0) and χ^2_1, often denoted as $\chi^2_{0:1}$. Thus, we obtain $p = P(\chi^2_{0:1} \geq 11.8) = 0.5P(\chi^2_0 \geq 11.8) + 0.5P(\chi^2_1 \geq 11.8) = 0.0003$, showing that the inclusion of the overdispersion parameter was important to model the data.

These results are quite similar to Menarin and Lara (2017), however, using the GEE approach an evidence of significant post-grazing management effect was reported, while it was not observed for the GLMM and the CM. It should be said that once should not consider the framework as best fit, but more as an elegant way of dealing with overdispersion and hierarchical structure simultaneously. For this experiment, overdispersion is something reasonable to happen since it is a field experiment that can suffer with several environmental changes and also because some types of vegetation can occur in an aggregate pattern inside paddocks, e.g., the weeds dispersion or bare ground occurrence.

3.8 Concluding remarks

In this Chapter, we have proposed a model for overdispersed, repeated nominal data. The model combines the baseline-category logit assumption to handle the nominal nature of the outcome, with normal random effects in the linear predictor to deal with correlation across repeated measures, and beta random effects to account for overdispersion. Similar models were proposed by Molenberghs and Verbeke (2007), Molenberghs *et al.* (2010), Molenberghs *et al.* (2012), Ivanova *et al.* (2014), and Molenberghs *et al.* (2017) for count data, binary and binomial data, time-to-event and ordinal outcomes. The model is easy to formulate and can be fitted in almost a routine fashion using, for example, the SAS procedure NLMIXED.

A limited simulation study was conducted to examine the behavior of the combined model relative to their more conventional GLMM counterparts. Both models performed well, but when there is a pronounced effect of overdispersion or if the overdispersion effect is associated with high correlations between the repeated measurements, better performance was observed for the CM, mainly for the variance components.

We applied the GLMM and the CM in agricultural experimental data to model the probability of occurrence of three types of vegetation (weeds, bare ground and tussocks). Comparing both models, a strong evidence is found in favor of the CM. It means that an extra parameter to model the overdispersion is needed in order to handle the environmental and biological changes that contribute extra-variability to the data.

The next steps of this research is to develop the CM for nominal outcomes with probit link and also extend the conjugated random effects in order to capture an overdispersion parameter for each category using, e.g., the Dirichlet distribution.

References

Agresti, A., 2010 *Categorical data analysis*. Wiley, New York.

Clayton, D., 1992 Repeated ordinal measurements: a generalised estimating equation approach pp. 1–11.

Demétrio, C. G. B., J. Hinde, and R. A. Moral, 2014 Models for Overdispersed Data in Entomology. In *Ecological modeling applied to entomology*, Springer.

Diggle, P. J., P. J. Heagerty, K. Y. Liang, and S. L. Zeger, 2002 *Analysis of longitudinal data*. Oxford University Press, New York.

Grunwald, G. K., S. L. Bruce, L. Jiang, M. Strand, and N. Rabinovitch, 2011 A statistical model for under- or overdispersed clustered and longitudinal count data. Biometrical Journal **53**: 578–594.

Hartzel, J., A. Agresti, and B. Caffo, 2001 Multinomial logit random effects models. Statistical Modelling **1**: 81–102.

Hedeker, D., 2003 A mixed-effects multinomial logistic regression model. Statistics in Medicine **1446**: 1433–1446.

Hinde, J. and C. G. B. Demétrio, 1998 Overdispersion: Models and estimation. Computational Statistics and Data Analysis **27**: 151–170.

Ivanova, A., G. Molenberghs, and G. Verbeke, 2014 A model for overdispersed hierarchical ordinal data. Statistical Modelling **14**: 399–415.

Lara, I. A. R. D., J. P. Hinde, A. C. D. Castro, and I. J. O. da Silva, 2017 A proportional odds transition model for ordinal responses with an application to pig behaviour. Journal of Applied Statistics **4763**: 1031–1046.

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

Lipsitz, L. R., K. Kim, and L. Zhao, 1994 Repeated categorical data using generalized estimating equations. Statistics in medicine **13**: 1149–1163.

Menarin, V. and I. A. R. Lara, 2017 Longitudinal model for categorical data applied in an agriculture experiment about elephant grass. Scientia Agricola **74**: 265–274.

Molenberghs, G. and G. Verbeke, 2005 *Models for discrete longitudinal data*. Springer-Verlang, New York.

Molenberghs, G. and G. Verbeke, 2007 Likelihood Ratio, Score, andWald Tests in a Constrained Parameter Space. The American Statistician **61**: 22–27.

Molenberghs, G., G. Verbeke, and C. G. B. Demétrio, 2017 Hierarchical models with normal and conjugate random effects: a review. SORT-Statistics and Operations Research Transactions **41**: 191–254.

Molenberghs, G., G. Verbeke, C. G. B. Demétrio, and A. M. C. Vieira, 2010 A Family of Generalized Linear Models for Repeated Measures with Normal and Conjugate Random Effects. Statistical Science **25**: 325–347.

Molenberghs, G., G. Verbeke, S. Iddi, and C. G. B. Demétrio, 2012 A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. Journal of Multivariate Analysis **111**: 94–109.

Morel, J. G. and K. J. Koehler, 1995 A one-step Gauss-Newton estimator for modelling categorical data with extraneous variation. Journal of the Royal Statistical Society. Series C (Applied Statistics) **44**: 187–200.

Morel, J. G. and N. K. Nagaraj, 1993 A finite mixture distribution for modelling multinomial extra variation. Biometrika **80**: 363–371.

Mosimann, J. E., 1962 On the compound multinomial distribution, the multivariate β - distribution, and correlations among proportions. Biometrika **49**: 65–82.

Neerchal, N. K. and J. G. Morel, 1998 Large cluster results for two parametric multinomial extra variation models. Journal of the American Statistical Association **93**: 1078–1087.

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

P, M. and J. A. Nelder, 1983 *Generalized linear models*. Chapman {&} Hall, London, first edition.

Pereira, L. E. T., A. J. Paiva, E. V. Geremia, and S. C. Silva, 2015a Grazing management and tussock distribution in elephant grass. Grass and Forage Science pp. 406–417.

Pereira, L. E. T., A. J. Paiva, E. V. Geremia, and S. C. Silva, 2015b Regrowth patterns of elephant grass (Pennisetum purpureum Schum) subjected to strategies of intermittent stocking management. Grass and Forage Science **70**: 195–204.

Pinheiro, J. C. and D. M. Bates, 1995 Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics **4**: 12–35.

Self, S. G. and K. Y. Liang, 1987 Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. Journal of the American Statistical Association **82**: 605–610.

Stram, D. O. and J. W. Lee, 1994 Variance Components Testing in the Longitudinal Mixed Effects Model. Biometrics pp. 1171–1177.

Touloumis, A., A. Agresti, and M. Kateri, 2013 GEE for Multinomial Responses Using a Local Odds Ratios Parameterization. Biometrics **69**: 633–640.

4 FINAL CONSIDERATIONS

In this thesis, we aimed to explore and develop flexible statistical models to analyze agricultural datasets. We notice that such data are rich, involving several sources of variability that usually classical models are not able to handle. In Chapter 2, a model that uses random smoothing splines was used to assess the biological control of the insect *Diaphorina citri* using an entomopathogenic fungus *Isaria fumosorosea* ESALQ-1296. We described the main aspects of this model and compared with the standard models for count data. This methodology brings more information to the data analysis, describing properly the fluctuations of the insects over days.

In Chapter 3, we developed a combined model that takes into account overdispersion and clustering through two separate sets of random effects. This study was motivated by an experiment that aims to model the probability of occurrence of three types of vegetation in a longitudinal experiment with grass pasture and dairy cows. The analysis undertaken for this dataset showed that the extended framework increased in model fit, when comparing it to traditional generalized linear mixed model framework. Therefore, it is important to note that aspects like overdispersion and hierarchical structure need to be taken into account when making appropriate predictions and conclusions. Since general conclusions cannot be made on a few data analysis, simulation studies are put forward to explore the extended framework in detail. To conclude, it should be said that one should not consider the framework as best fit, but more as an elegant way of dealing with overdispersion and hierarchical structure simultaneously.

APPENDICES

Appendix A: SAS code for Chapter 2

We carried out all programming in SAS version 9.4. For negative binomial and Poisson-normal models the following codes were used:

```
proc glimmix data=dc;
title 'Negative binomial';
class treat(ref='Control') block;
model counts = block treat time treat*time time2 treat*time2  / solution
dist=negbinomial;
output out=NB /allstats;
run;

proc glimmix data=dc method=laplace;
title 'Poisson-normal';
class treat(ref='Control') block plot;
model counts = block treat time treat*time time2 treat*time2  / solution
dist=poisson;
random  intercept  / solution subject=plot;
output out=PN /allstats;
run;
```

To fit the generalized linear mixed models with random smoothing splines, we first constructed the second degree B-spline basis with `proc transreg`. We chose 6 knots (`nkots=6`) using the rule of thumb described in Sect. 3.3. This results in 6 + 2 terms, `time_0-time_8`, that were stored in the data set `basis`.

```
proc transreg data=dc;
model identity(counts) = bspline(time/degree=2 nknots=6);
output out=basis predicted;
run;
```

We merged the resulting transformations of time with the original data set, which contains the outcome (counts), the plot number (plot) and the covariates (treat, block, time and time2).

```
data final;
merge dc basis;
keep counts plot treat block time time2 time_0-time_8;
run;
```

We used PROC GLIMMIX to fit the generalized linear mixed models with random smoothing splines as described in Section 3.3 as follows:

```
proc glimmix data=final method=laplace;
title 'Prs';
class treat(ref='Control') block plot ;
model counts = block treat time treat*time time2 treat*time2 / solution
dist=poisson;
random time_0-time_8 / type=toep(1) solution subject=plot;
output out=Prs /allstats;
run;

proc glimmix data=final method=laplace;
title 'NBrs';
class treat(ref='Control') block plot ;
model counts = block treat time treat*time time2 treat*time2  / solution
dist=negbinomial;
random time_0-time_8 / type=toep(1) solution subject=plot;
output out=NBrs /allstats;
run;
```

The specification `type=toep(1)` is the same as $\tau^2\mathbf{I}$, and it is useful to specify the same variance component for several effects.

Appendix B: CM algebraic development of Chapter 3

The partially marginalized density function of the combined model was obtained by integrating analytically over the beta random effects, leaving the normal random effects untouched. To do this, we need to consider the category to which the outcome belongs in order to proceed with the integration over the beta random effect. To simplify notation, let us consider the case where 3 categories are analyzed. Leaving the Gaussian effects

untouched, i.e., treating them as constants, we can rewrite this expression as

$$\begin{aligned} f(w_{r,ij}|\boldsymbol{b}_{r,i}) &= \int (\theta_{ij}\kappa_{1,ij})^{w_{1,ij}} \ (\theta_{ij}\kappa_{2,ij})^{w_{2,ij}} \\ &\quad (1-\theta_{ij}\kappa_{1,ij}-\theta_{ij}\kappa_{2,ij})^{1-w_{1,ij}-w_{2,ij}} \ \frac{\theta_{ij}^{\vartheta-1}(1-\theta_{ij})^{\xi-1}}{B(\vartheta,\xi)} \ d\theta_{ij}. \end{aligned}$$

Thus, if the outcome belongs to the first category, e.i., $W_{r,ij}$ is equals to 1 if $Y_{ij}=1$ and 0 otherwise, the following expression is obtained

$$\begin{aligned} f(w_{1,ij}=1|\boldsymbol{b}_{r,i}) &= \int_0^1 (\theta_{ij}\kappa_{1,ij})^1 \ \frac{\theta_{ij}^{\vartheta-1}(1-\theta_{ij})^{\xi-1}}{B(\vartheta,\xi)} \ d\theta_{ij}, \\ &= \frac{\kappa_{1,ij}}{B(\vartheta,\xi)}\int_0^1 \theta_{ij}^{(\vartheta-1)+1}(1-\theta_{ij})^{\xi-1}d\theta_{ij} \\ &= \kappa_{1,ij}\frac{B(\vartheta+1,\xi)}{B(\vartheta,\xi)} \\ &= \kappa_{1,ij} \ \frac{\Gamma(\vartheta+1)\Gamma(\xi)}{\Gamma(\vartheta+\xi+1)} \ \frac{\Gamma(\vartheta+\xi)}{\Gamma(\vartheta)\Gamma(\xi)} \\ &= \kappa_{1,ij} \ \vartheta \ \frac{\Gamma(\vartheta)\Gamma(\xi)}{\Gamma(\vartheta+\xi+1)} \ \frac{\Gamma(\vartheta+\xi)}{\Gamma(\vartheta)\Gamma(\xi)} \\ &= \kappa_{1,ij} \ \vartheta \ \frac{\Gamma(\vartheta+\xi)}{(\vartheta+\xi)\Gamma(\vartheta+\xi)} \\ &= \kappa_{1,ij} \ \frac{\vartheta}{\vartheta+\xi}. \end{aligned}$$

Similar results apply to $r=2$, but multiplied, of course, by their respective κ. For the last category ($r=3$), the expression is given by

$$\begin{aligned} f(w_{3,ij}=1|\boldsymbol{b}_{r,i}) &= \int_0^1 \ (1-\theta_{ij}\kappa_{1,ij}-\theta_{ij}\kappa_{2,ij})^1 \ \frac{\theta_{ij}^{\vartheta-1}(1-\theta_{ij})^{\xi-1}}{B(\vartheta,\xi)} \ d\theta_{ij} \\ &= \int_0^1 \ \frac{\theta_{ij}^{\vartheta-1}(1-\theta_{ij})^{\xi-1}}{B(\vartheta,\xi)} \\ &\quad -(\theta_{ij}\kappa_{1,ij}) \ \frac{\theta_{ij}^{\vartheta-1}(1-\theta_{ij})^{\xi-1}}{B(\vartheta,\xi)} - (\theta_{ij}\kappa_{2,ij}) \ \frac{\theta_{ij}^{\vartheta-1}(1-\theta_{ij})^{\xi-1}}{B(\vartheta,\xi)} \ d\theta_{ij} \\ &= 1-\kappa_{1,ij}\frac{\vartheta}{\vartheta+\xi}-\kappa_{2,ij}\frac{\vartheta}{\vartheta+\xi} \\ &= 1-\frac{\vartheta}{\vartheta+\xi}(\kappa_{1,ij}+\kappa_{2,ij}). \end{aligned}$$

In this sense, the partially marginalized likelihood function of the combined model considering 3 categories is given by

$$L(\boldsymbol{\beta}, \boldsymbol{D}, \vartheta, \xi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \left(\frac{\vartheta}{\vartheta + \xi} \kappa_{1,ij} \right)^{w_{1,ij}} \left(\frac{\vartheta}{\vartheta + \xi} \kappa_{2,ij} \right)^{w_{2,ij}} \left(1 - \frac{\vartheta}{\vartheta + \xi} \sum_{h=1}^{R-1} \kappa_{h,ij} \right)^{1-w_{1,ij}-w_{2,ij}} \frac{1}{\sqrt{(2\pi)^{n_i}}} \frac{1}{\sqrt{|\boldsymbol{D}|}} \exp\left(-\frac{1}{2} \boldsymbol{b}_{r,i}^T \boldsymbol{D}^{-1} \boldsymbol{b}_{r,i} \right) d\boldsymbol{b}_{r,i}.$$

Appendix C: SAS code for Chapter 3

We carried out all programming in SAS version 9.4. To fit the GLMM and the CM, the following codes were used:

```
proc nlmixed data=dados2 method=gauss qpoints=5 tech=quanew;
title 'GLMM';
parms beta01=-2.9 beta02=-0.03 beta11=0.6 beta12=0.12 beta21=0.54
beta22=-.005 beta31=-0.69 beta32=0.07 beta41=0.67 beta42=-0.56 beta61=0.5
beta62=-0.85 beta71=-0.34 beta72=-0.32 beta81=-0.16 beta82=-0.34 beta91=0.32
beta92=-0.29 beta101=0.96 beta102=-0.07 beta111=-0.4 beta112=0.89
beta121=0.33 beta122=0.29 beta131=-0.72 beta132=0.45 beta141=-0.66
beta142=0.12 beta151=-1.69 beta152=0.16 var1=.1;

xb1= beta01 + beta11*block1 + beta21*block2 + beta31*block3 + beta41*pre95 +
beta61*month1 + beta71*month2 + beta81*month3 + beta91*month4 + beta101*month5+
beta111*month1pre95 + beta121*month2pre95 + beta131*month3pre95 +
beta141*month4pre95 +beta151*month5pre95;

xb2= beta02 + beta12*block1 + beta22*block2 + beta32*block3 + beta42*pre95 +
beta62*month1 + beta72*month2 + beta82*month3 + beta92*month4 + beta102*month5+
beta112*month1pre95 + beta122*month2pre95 + beta132*month3pre95 +
beta142*month4pre95 + beta152*month5pre95;

eta1=xb1+b1;
eta2=xb2+b1;

den= 1+exp(eta1)+exp(eta2);

k1=exp(eta1)/den;
```

```
k2=exp(eta2)/den;
k3=1/den;

if (y=1) then lik=k1;
else if (y=2) then lik=k2;
else if (y=3) then lik=k3;

ll=log(lik) ;
model y ~ general(ll);
random b1 ~ normal(0,var1) subject=idpoint;
ods output ParameterEstimates=a FitStatistics=b;
predict  xb1 out=output_fixed1;
predict  xb2 out=output_fixed2;
predict  eta1 out=output_random1;
predict  eta2 out=output_random2;
run;

PROC NLMIXED DATA=dados2 method=gauss qpoints=5 tech=quanew;
title 'CM';

parms beta01=-2.9 beta02=-0.03 beta11=0.6 beta12=0.12 beta21=0.54
beta22=-.005 beta31=-0.69 beta32=0.07 beta41=0.67 beta42=-0.56 beta61=0.5
beta62=-0.85 beta71=-0.34 beta72=-0.32 beta81=-0.16 beta82=-0.34 beta91=0.32
beta92=-0.29 beta101=0.96 beta102=-0.07 beta111=-0.4 beta112=0.89
beta121=0.33 beta122=0.29 beta131=-0.72 beta132=0.45 beta141=-0.66
beta142=0.12 beta151=-1.69 beta152=0.16 var1=.1;

xb1= beta01 + beta11*block1 + beta21*block2 + beta31*block3 + beta41*pre95 +
beta61*month1 + beta71*month2 + beta81*month3 + beta91*month4 + beta101*month5+
beta111*month1pre95 + beta121*month2pre95 + beta131*month3pre95 +
beta141*month4pre95 +beta151*month5pre95;

xb2= beta02 + beta12*block1 + beta22*block2 + beta32*block3 + beta42*pre95 +
beta62*month1 + beta72*month2 + beta82*month3 + beta92*month4 + beta102*month5+
beta112*month1pre95 + beta122*month2pre95 + beta132*month3pre95 +
beta142*month4pre95 + beta152*month5pre95;
```

```
eta1=xb1+b1;
eta2=xb2+b1;

den=1+exp(eta1)+exp(eta2);

K1=exp(eta1)/den;
K2=exp(eta2)/den;

nu=exp(delta)/(1+exp(delta));

if (y=1) then p=nu*K1;
else if (y=2) then p=nu*K2;
else if (y=3) then p=1-(nu*K1 + nu*K2);

LogLike = LOG(p);
model y ~ general(LogLike);
random b1~ normal(0,var1) subject=Idpoint;
run;
```

To perform the simulation study where an overdispersion effect is included, we developed the following MACRO:

```
*-----Parameters-----;

*category 1;
%let beta10=0.1;
%let beta11=0.2;
%let beta12=0.5;
%let beta13=0.7;

*category 2;
%let beta20=0.2;
%let beta21=0.1;
%let beta22=0.4;
%let beta23=0.5;
```

```
* Number of time points, number of ids, sigma2
(normal random effect);
%let ntime=6;
%let nid=300;

%let d11=9;
%let d22=4;
%let d12=0.5;

%let alpha=20;
%let beta=1;

*create objects to store the estimates and the dataset;

data glmm;
input Parameter$1-6 Estimate StandardError DF tValue Probt
Alpha Lower Upper Gradient dataset;
datalines;
;
run;

data cm;
input Parameter$1-6 Estimate StandardError DF tValue Probt
Alpha Lower Upper Gradient dataset;
datalines;
;
run;

data dat;
input trt time id y dataset;
datalines;
;
run;
*/
```

```
*-----MACRO-----;
%macro simu(start,stop,par1,par2,dat1);

%do k=&start %to &stop;

proc iml;

call randseed(&k);

beta10=&beta10;
beta11=&beta11;
beta12=&beta12;
beta13=&beta13;

beta20=&beta20;
beta21=&beta21;
beta22=&beta22;
beta23=&beta23;

ntime=&ntime;
nid=&nid;

d11=&d11;
d22=&d22;
d12=&d12;

int1=J(&nid,&ntime,&beta10);
time1=&beta11*J(&nid,1,1)*shape(((1:&ntime)-1)/&ntime,1,&ntime);
trt1=J(&nid/2,&ntime,0) // J(&nid/2,&ntime,&beta12);
interaction1 =J(&nid/2,&ntime,0)
//shape((&beta13*((1:&ntime)-1)/&ntime),&nid/2,&ntime) ;

int2=J(&nid,&ntime,&beta20);
time2=&beta21*J(&nid,1,1)*shape(((1:&ntime)-1)/&ntime,1,&ntime);
trt2=J(&nid/2,&ntime,0) // J(&nid/2,&ntime,&beta22);
interaction2 =J(&nid/2,&ntime,0)
```

```
//shape((&beta23*((1:&ntime)-1)/&ntime),&nid/2,&ntime) ;

fixef1= int1 + time1 + trt1 + interaction1;
fixef2= int2 + time2 + trt2 + interaction2;

Mean = {0, 0}; /* population means */
Cov = {&d11 &d12, /* population covariances */
&d12 &d22};

rand = randnormal(&nid, Mean, Cov);
rand1=rand[,1]*J(1,&ntime,1);
rand2=rand[,2]*J(1,&ntime,1);

eta1=fixef1+rand1;
eta2=fixef2+rand2;

kappa1= exp(eta1)/(1+exp(eta1)+exp(eta2));
kappa2= exp(eta2)/(1+exp(eta1)+exp(eta2));

theta = j(&nid,&ntime, .);
call randgen(theta, "beta", &alpha, &beta);

pi1=kappa1#theta;
pi2=kappa2#theta;
pi3=1-pi1-pi2;

a=j(&nid,&ntime,.);

do i=1 to &nid;
do j=1 to &ntime;
a[i,j]=randmultinomial(1,1,pi1[i,j]||pi2[i,j]||pi3[i,j])*{1,2,3};
end;
end;

y=colvec(a);
time=(colvec(repeat(shape(((1:&ntime)-1)/&ntime,1,&ntime),1,nid)));
```

```
id=(colvec(repeat(t(1:nid),1,ntime)));
trt=(colvec(repeat(t(0:1),1,(nid/2)*(ntime))));

create mydata var {id time trt y};
append;
close mydata;
quit;

data mydata;
set mydata;
dataset=&k;
run;
proc print data=mydata;run;

*-----model estimation-----;
proc nlmixed data=mydata method=gauss qpoints=10;
title'glmm';
parms beta10=.1 beta11=.2 beta12=.5 beta13=.7
beta20=.2 beta21=.1 beta22=.4 beta23=.5;

eta1= beta10 + beta11*time + beta12*trt + beta13*trt*time + b1;
eta2= beta20 + beta21*time + beta22*trt + beta23*trt*time + b2;

den= 1+exp(eta1)+exp(eta2);

k1=exp(eta1)/den;
k2=exp(eta2)/den;
k3=1/den;

if (y=1) then lik=k1;
else if (y=2) then lik=k2;
else if (y=3) then lik=k3;

ll=log(lik) ;
model y ~ general(ll);
random b1 b2 ~ normal([0,0],[var1,cov,var2]) subject=id;
```

```
ods output ParameterEstimates=a;
run;
*----------;
proc nlmixed data=mydata method=gauss qpoints=10;
title'cm beta';
parms beta10=.1 beta11=.2 beta12=.5 beta13=.7
beta20=.2 beta21=.1 beta22=.4 beta23=.5;

eta1= beta10 + beta11*time + beta12*trt + beta13*trt*time + b1;
eta2= beta20 + beta21*time + beta22*trt + beta23*trt*time + b2;

den= 1+exp(eta1)+exp(eta2);

k1=exp(eta1)/den;
k2=exp(eta2)/den;

nu=exp(delta)/(1+exp(delta));

if (y=1) then lik=nu*k1;
else if (y=2) then lik=nu*k2;
else if (y=3) then lik=1-(nu*k1 + nu*k2);

ll=log(lik) ;
model y ~ general(ll);
RANDOM b1 b2 ~ NORMAL([0,0],[var1,cov,var2]) SUBJECT=id;
ods output ParameterEstimates=b;
RUN;
*----------;

data a; set a;
dataset=&k;
run;

data b; set b;
dataset=&k;
run;
```

```
proc append base=&par1 data=a force;
run;

proc append base=&par2 data=b force;
run;

proc append base=&dat1 data=mydata force;
run;

proc datasets library=work nolist;
delete a b mydata;
run;

%end;
%mend;

%simu(1,200,glmm,cm,dat);
```