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

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

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

Piracicaba
2018

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 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 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.

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 Agrigultura, 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 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.