Statistics and Operations Research Transactions, vol. 41, n. 2 (2017)

SORT 41 (2) July-December 2017, 191-254 DOI: 10.2436/20.8080.02.58

Hierarchical models with normal and conjugate random effects: a review

Geert Molenberghs[1,2,*], Geert Verbeke[2,1] and Clarice G.B. Demétrio[3]

Abstract

Molenberghs, Verbeke, and Demétrio (2007) and Molenberghs et al. (2010) proposed a general framework to model hierarchical data subject to within-unit correlation and/or overdispersion. The framework extends classical overdispersion models as well as generalized linear mixed models. Subsequent work has examined various aspects that lead to the formulation of several extensions. A unified treatment of the model framework and key extensions is provided. Particular extensions discussed are: explicit calculation of correlation and other moment-based functions, joint modelling of several hierarchical sequences, versions with direct marginally interpretable parameters, zero-inflation in the count case, and influence diagnostics. The basic models and several extensions are illustrated using a set of key examples, one per data type (count, binary, multinomial, ordinal, and time-to-event).

MSC: 62P10, 97K80.

Keywords: Conjugacy, frailty, joint modelling, marginalized multilevel model, mixed model, overdispersion, underdispersion, variance component, zero-inflation.

1. Introduction

Parametric or semi-parametric modelling of univariate non-Gaussian outcomes is often done within the generalized linear model (GLM) framework (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Agresti, 2002), which rests on the exponential family. Commonly encountered outcome types include categorical (binary, binomial, ordinal, etc.), count, and time-to-event outcomes, for which modelling typically, though not always, rests upon the Bernoulli, Poisson, and exponential/Weibull distributions, respectively. A key feature of exponential family distributions is the so-called *mean-*

* Corresponding author: Geert Molenberghs, I-BioStat, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium. geert.molenberghs@uhasselt.be

[1] I-BioStat, Universiteit Hasselt, B-3500 Hasselt, Belgium.

[2] I-BioStat, KU Leuven, B-3000 Leuven, Belgium.

[3] ESALQ, Universidade de Saõ Paulo, Piracicaba, Brazil.
Received: May 2017.

variance relationship, i.e., the fact that the variance is a deterministic function of the mean. For example, for Bernoulli outcomes with success probability $\mu = \pi$, the variance is $v(\mu) = \pi(1-\pi)$, for counts using Poisson assumptions $v(\mu) = \mu$ and for the exponential model $v(\mu) = \mu^2$. However, for many outcome types, empirically observed data can contradict this relationship, in the sense that the observed variance may be higher or lower than what follows from the model formulation; these are referred to as overdispersion and underdispersion, respectively. The two phenomena combined are sometimes referred to as extra-model-dispersion. Especially in the somewhat older literature, more attention was given to overdispersion than to underdispersion. Hinde and Demétrio (1998ab) provide early overviews of (semi-)parametric approaches for dealing with overdispersion. Well-known models include the beta-binomial (Skellam, 1948; Kleinman, 1973) for binary and binomial data, and the negative binomial model (Breslow, 1984; Lawless, 1987) for counts. These models can be generated by assuming the so-called natural parameter to follow a carefully chosen distribution. For example, the beta-binomial models follow from assuming the outcomes follow a binomial distribution with parameter drawn from a beta distribution; the negative binomial model follows from a Poisson model with gamma distributed parameter. The resulting models have elegant parametric expressions and are relatively easy to interpret, because the outcome and random-effects distributions are *conjugate*, a precise definition of which is given in Section 4.2. Other solutions to accommodating overdispersion include mixture modelling and specific models for zero-inflated Poisson models (Ridout, Demétrio and Hinde, 1998; Böhning, 2000; McLachlan and Peel, 2000).

Nowadays, it is very common to encounter aforementioned data types in a hierarchical context, such as resulting from multivariate, longitudinal, spatial, and clustered designs. We will generically refer to these settings as repeated measures. The data hierarchies induce association among the repeated measures, which can be captured, among others, by random effects. Especially the generalized linear mixed model (GLMM; Engel and Keen, 1994; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) has beÃ§come a popular and widespread tool, routinely implemented in a suite of standard software packages. Reviews are given in Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005). A key ingredient is a linear predictor that also incorporates normally distributed random effects. These random effects engender not only correlation among the repeated measures, but also some overdispersion. However, the empirical correlation and overdispersion present in the data may be hard to model with only a limited number of normal random effects. This is why Molenberghs et al. (2007; henceforth referred to as MVD) and Molenberghs et al. (2010; henceforth referred to as MVDV) have proposed a model family, the so-called *combined model* (CM) that combines conjugate and normal random effects, leading to highly increased flexibility for the triple of functions made up of the mean, variance, and correlation functions. Note that, for time-to-event data, not only GLMM but also the so-called frailty models (Duchateau and Janssen, 2007) have been used. These start from gamma rather than normally distributed random effects, which are conjugate to the exponential distribu-

tion, and lead to elegant expressions when combined with the Weibull distribution as well (see Section 7).

After introducing a set of key examples (Section 2) and reviewing several key ingredients in Section 3, the CM is introduced in Section 4. Sections 5-7 are devoted to the count, categorical, and time-to-event cases, respectively. In the count case, specific attention is given to the occurrence of extra-model zeroes, i.e., zero-inflated versions of the model. In the categorical case, we further distinguish between binary, binomial, and ordinal data. Of note is the rather different algebraic nature of the model with logit and that with probit link. In the time-to-event case, we also allow for censoring, and discuss some issues with the moment functions of the so-called Weibull-gamma-normal model and its sub-models. In Section 8 we describe maximum likelihood and some related estimation strategies.

In Section 9, we show how the CM and its sub-models can be used, in most cases, to derive explicit expressions for so-called manifest correlations, whereas often, for convenience, the latent correlation is considered. Usually, though, the manifest correlation is considerably smaller than its latent counterpart; hence, using the latter may lead to overly optimistic conclusions.

A typical problem arising with the GLMM, in contrast to the GLM, the linear mixed model (LMM) for Gaussian outcomes, and models with conjugate random effects is that deriving marginal expressions is not so straightforward and, related to this, that the model parameters have a hierarchical (i.e., conditional on the random effects) but not a marginal (i.e., averaged over a suitable population) interpretation. The CM evidently inherits this problem. While some progress is made for the specific cases discussed in Sections 5-7, it is still useful to take a different route: that of a so-called marginalized multilevel model, based on work of Heagerty (1999) and Heagerty and Zeger (2000). It will be referred to as the *combined marginalized multilevel model*, or COMMM.

Evidently, in line with a lot of contemporary work, it is perfectly possible to observe, for example, several longitudinal sequences simultaneously. The resulting designs are referred to as multivariate longitudinal or, more generically, joint modelling. The use of the CM in this context is reviewed in Section 11. Finally, Section 12 describes diagnostic measures based on local influence.

The review in this paper is based on work by MVD and MVDV, which is also based on Booth et al. (2003), and various extensions of all of these. Evidently, also different strands of research exist that extend the GLMM and increase its flexibility. In particular, we refer to Lee and Nelder (1996, 2001ab, 2003), Lee, Nelder, and Pawitan (2006), who proposed so-called *hierarchical generalized linear models*, accommodating many outcome and random-effects distributions, while being efficient in computational terms. In the particular case of count data, our model relates to theirs by considering log-gamma and log-normal random effects together. Regarding estimation, we focus primarily on marginal maximum likelihood estimation and Bayesian estimation, whereas Lee and Nelder employ so-called h-likelihood. In particular, we analytically integrate over the conjugate random effects and use numerical integration for the normal random effects.

Skrondal and Rabe-Hesketh (2004) brought together in a single model framework, multilevel modelling, structural equations modelling, latent variables, latent classes, and random-effects models for hierarchical data.

2. Case studies

We will describe five case studies. The outcomes are of a count, binary, binomial, ordinal, and time-to-event nature, respectively.

2.1. A clinical trial in epileptic patients

The data considered here are obtained from a randomized, double-blind, parallel group multicentre study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AEDs. The study is described in full detail in Faught et al. (1996). The randomization of epilepsy patients took place after a 12-week baseline period that served as a stabilization period for the use of AEDs, and during which the number of seizures were counted. After that period, 45 patients were assigned to the placebo group, 44 to the active (new) treatment group. Patients were then measured weekly. Patients were followed (double-blind) during 16 weeks, after which they were entered into a long-term open-extension study. Some patients were followed for up to 27 weeks. The outcome of interest is the number of epileptic seizures experienced during the most recent week. The research question is whether or not the additional new treatment reduces the number of epileptic seizures.

2.2. A clinical trial in onychomycosis

These data come from a randomized, double-blind, parallel group, multicentre study for the comparison of two oral treatments (coded as A and B) for toenail dermatophyte onychomycosis (TDO), described in full detail by De Backer et al. (1996). TDO is a common toenail infection, difficult to treat, affecting more than 2 out of 100 persons (Roberts, 1992). Anti-fungal compounds, classically used for treatment of TDO, need to be taken until the whole nail has grown out healthy. The development of new such compounds, however, has reduced the treatment duration to 3 months. The aim of the present study was to compare the efficacy and safety of 12 weeks of continuous therapy with treatment A or with treatment B. In total, 2×189 patients, distributed over 36 centres, were randomized. Subjects were followed during 12 weeks (3 months) of treatment and followed further, up to a total of 48 weeks (12 months). Measurements were taken at baseline, every month during treatment, and every 3 months afterwards, resulting in a maximum of 7 measurements per subject. At the first occasion, the treating physician indicates one of the affected toenails as the target nail, the nail which will be followed over time. We will restrict our analyses to only those patients for which

the target nail was one of the two big toenails (146 and 148 subjects, in group A and group B, respectively). One of the responses of interest was the unaffected nail length, measured from the nail bed to the infected part of the nail, which is always at the free end of the nail, expressed in *mm*. This outcome has been studied extensively in Verbeke and Molenberghs (2000). Another important outcome in this study was the severity of the infection, coded as 0 (not severe) or 1 (severe). The question of interest was whether the percentage of severe infections decreased over time, and whether that evolution was different for the two treatment groups.

2.3. Iron-deficient diets in rats

These data result from an experiment where female rats were put on iron-deficient diets (Shepard, Mackler, and Finch, 1980). This dataset has been analysed by Liang and McCullagh (1993) and Moore and Tsiatis (1991). In Agresti (2002), the data were used to estimate several logit models. Experimental rats were divided into 4 groups, one of which is a control group. The number of female rats per group (total number of fetuses per group) are: 31 (327) for placebo, 12 (118) for low dose, 5 (58) for medium dose, and 10 (104) for high dose. Weekly injections of iron supplement were to bring the rats' iron intake to normal levels. Rats in the placebo group were given placebo injection, the others got three different doses of the iron supplements. Rats were made pregnant and sacrificed 3 weeks later and the total number of fetuses and the number of dead fetuses in each litter were counted. Hemoglobin levels of the mothers were also measured.

2.4. Diabetes study

In Belgium, the diabetes project was conducted from January 2005 until December 2006, with the aim to study the effect of implementing a structured model for chronic diabetes care on the patients' clinical outcomes. General practitioners (GPs) were offered assistance and could redirect patients to the diabetes care team, consisting of a nurse educator, a dietician, an ophthalmologist, and an internal medicine doctor. For the project, two programs were implemented and GPs were randomized to one of two groups: UQIP: Usual Quality Improvement Program and AQIP: Advanced Quality Improvement Program. A total of 120 GPs took part in the study, 53 in the UQIP group and 67 in the AQIP group, including 918 and 1577 patients, respectively.

During the project, several outcomes useful to evaluate how well diabetes is controlled were measured, at the moment the program was initiated (time T_0) and one year later (T_1). The most important outcomes were HbA1c (glycosylated hemoglobin), LDL-cholesterol (low-density lipoprotein cholesterol) and SBD (systolic blood pressure). Furthermore, experts specified cut off values defining a so-called *clinical target* for each outcome: HBA1C$<$7%, LDL-cholesterol $<$ 100 mg/dl and SBD $\leq$ 130 mmHg. As a result, for a particular time point, every patient could reach between 0 and 3 clinical targets. This number was reflected in the variable *number of clinical targets*. If at least

one measurement per patient was missing, the value for the number of clinical targets was set to missing as well. The data are discussed in Borgermans et al. (2009).

2.5. Recurrent asthma attacks in children

These data have been studied in Duchateau and Janssen (2007). Asthma is occurring more and more frequently in very young children (between 6 and 24 months). Therefore, a new application of an existing anti-allergic drug is administered to children who are at higher risk to develop asthma in order to prevent it. A prevention trial is set up with such children randomized to placebo or drug, and the asthma events that developed over time are recorded in a diary. Typically, a patient has more than one asthma event. The different events are thus clustered within a patient and ordered in time. This ordering can be taken into account in the model. The data are presented in calendar time format, where the time at risk for a particular event is the time from the end of the previous event (asthma attack) to the start of the next event (start of the next asthma attack). A particular patient has different periods at risk during the total observation period which are separated either by an asthmatic event that lasts one or more days or by a period in which the patient was not under observation. The start and end of each such risk period is required, together with the status indicator to denote whether the end of the risk period corresponds to an asthma attack or not.

3. Some background

We briefly review some background on the exponential family and generalized linear models (Section 3.1), overdispersion (Section 3.2), and models with normal random effects (Section 3.3).

3.1. Generalized linear models

A random variable Y follows an exponential family distribution if the density is of the form

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

for a specific set of unknown parameters η ('natural parameter' or 'canonical parameter') and ϕ ('dispersion parameter'), and for known functions $\psi(\cdot)$ and $c(\cdot,\cdot)$. It follows that $\mathrm{E}(Y) = \mu = \psi'(\eta)$ and $\mathrm{Var}(Y) = \sigma^2 = \phi\psi''(\eta)$, with ensuing mean-variance relationship $\sigma^2 = \phi\psi''[\psi'^{-1}(\mu)] = \phi v(\mu)$, with $v(\cdot)$ the *variance function*. Commonly encountered examples and their model elements are presented in Table 1. Note that, in the normal case, there is no mean-variance relationship. In the binary case, also the

probit link is commonly encountered, whence $\eta = \Phi^{-1}(\pi)$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. As explained in Section 6.1, the probit link has appealing properties when normal random effects are introduced into the model.

In the Weibull and exponential model, the decomposition $\varphi = \lambda e^{\mu}$ is often used, allowing μ to be written as a function of covariates. Note that μ is a component of the mean function, not the mean itself. The Weibull model does not belong to the exponential family in a conventional sense, unless when y is replaced by y^{ρ}. In Table 1, $\Gamma(\cdot)$ represents the gamma function.

When not the full joint distribution but, say, the first and second moments only are specified, a semi-parametric version of the model results, for which quasi-likelihood estimation has been devised (McCullagh and Nelder, 1989; Molenberghs and Verbeke, 2005).

The generalized linear model (GLM) follows from the exponential family by assuming that a set of independent replicates Y_i with p-dimensional covariate vectors $\boldsymbol{x}_i$ $(i = 1,\ldots,N)$, follow exponential-family densities $f(y_i|\eta_i,\phi)$. Specification of the GLM is completed by modelling the means μ_i as functions of the covariate values: $\mu_i = h(\eta_i) = h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\xi})$, for a known function $h(\cdot)$, and with $\boldsymbol{\xi}$ a vector of p fixed, unknown regression coefficients. Here, $h^{-1}(\cdot)$ is called the link function. In most applications, the so-called natural link function is used, i.e., $h(\cdot) = \psi'(\cdot)$, which is equivalent to assuming $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\xi}$. In other words, it is assumed that the natural parameter satisfies a linear regression model.

3.2. Overdispersion

As stated in the introduction, and as is clear from Table 1, many standard exponential family models enforce a mean-variance relationship that may be contradicted by the data, especially for count, binomial, and time-to-event data. For binary data, such a violation can only occur when the outcomes are correlated (see Section 6).

As reviewed by Hinde and Demétrio (1998ab), an obvious way to incorporate overdispersion is by allowing $\phi \neq 1$, so that the variance becomes $\mathrm{Var}(Y) = \phi v(\mu)$. An elegant way forward is through a two-stage approach. For binary data, one would assume that $Y_i|\pi_i \sim \mathrm{Bernoulli}(\pi_i)$ and further that π_i is a random variable with $\mathrm{E}(\pi_i) = \mu_i$ and $\mathrm{Var}(\pi_i) = \sigma_i^2$. Using iterated expectations, it follows that $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \mu_i(1-\mu_i)$, underscoring that purely Bernoulli data are unable to exhibit overdispersion. The situation is different for counts. In the Poisson case, we assume that $Y_i|\zeta_i \sim \mathrm{Poi}(\zeta_i)$ and then that ζ_i is a random variable with $\mathrm{E}(\zeta_i) = \mu_i$ and $\mathrm{Var}(\zeta_i) = \sigma_i^2$. Then, it follows that $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \mu_i + \sigma_i^2$. We have not assumed a particular distributional form for the random effects π_i and ζ_i, respectively. Hence, this gives rise to a semi-parametric specification. In case it is considered advantageous to make full distributional assumptions about the random effects, common choices are the beta distribution for π_i and the gamma distribution for ζ_i; of course, these are not the only ones.

The two-stage approach is made up of considering a distribution for the outcome variable, given a random effect $f(y_i|\theta_i)$ which, combined with a model for the random effect, $f(\theta_i)$, produces the marginal model:

$$f(y_i) = \int f(y_i|\theta_i) f(\theta_i) d\theta_i. \tag{2}$$

It is easy to extend this model to the case of repeated measurements by assuming a hierarchical data structure, where now Y_{ij} denotes the jth outcome measured for cluster (subject) i, $i = 1,\ldots,N$, $j = 1,\ldots,n_i$ and $\boldsymbol{Y}_i$ is the n_i-dimensional vector of all measurements available for cluster i. In the repeated-measures case, the scalar ζ_i becomes a vector $\boldsymbol{\zeta}_i = (\zeta_{i1},\ldots,\zeta_{in_i})^{\mathrm{T}}$, with $\mathrm{E}(\boldsymbol{\zeta}_i) = \boldsymbol{\mu}_i$ and $\mathrm{var}(\boldsymbol{\zeta}_i) = \boldsymbol{\Sigma}_i$. For example, for the Poisson case, similar logic as in the univariate case produces $\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ and $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{M}_i + \boldsymbol{\Sigma}_i$, where $\boldsymbol{M}_i$ is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the diagonal. Note that a diagonal structure of $\boldsymbol{M}_i$ reflects the conditional independence assumption: all dependence between measurements on the same unit stems from the random effects. Generally, a versatile class of models results. For example, assuming that the components of $\boldsymbol{\zeta}_i$ are independent, a pure overdispersion model follows, without correlation between the repeated measures. On the other hand, assuming $\zeta_{ij} = \zeta_i$, i.e., that all components are equal, then $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{M}_i + \sigma_i^2 \boldsymbol{J}_{n_i}$, where $\boldsymbol{J}_{n_i}$ is an $n_i \times n_i$ dimensional matrix of ones. Such a structure can be seen as a general version of compound symmetry.

Alternatively, this repeated version of the overdispersion model can be combined with normal random effects in the linear predictor. This very specific choice was also proposed by Thall and Vail (1990) and Dean (1991) for the count case.

Marginalization (2) is general and elegant, but one has to reflect on which parameter to become random, in particular when full distributional assumptions are requested. As always, this is easy for the linear mixed model, by combining a normal hierarchical model with a normal random effect, and provided θ_i is used to express the conditional mean as a linear function of covariates. It forms the basis of the two strands of random-effects models that are potentially brought together in the combined models of Section 4: on the one hand, normal random effects can be considered with non-normal outcomes, producing the GLMM; on the other hand, gamma random effects for the Poisson model, beta random effects with binomial data, and gamma random effects for the Weibull model can be considered. This is, seemingly, a disparate collection. However, they are unified through so-called *conjugacy*, in the sense of Cox and Hinkley (1974, p. 370) and Lee et al. (2006, p. 178). The topic is also discussed by Agresti (2002). Informally, conjugacy refers to the fact that the hierarchical and random-effects densities have similar algebraic forms. Conjugate distributions produce a general and closed-form solution for the corresponding marginal distribution.

Table 1: *Conventional exponential family members and extensions with conjugate random effects.*

Element	Notation	Continuous	Binary	Count	Time to event	
Standard univariate exponential family						
Model		Normal	Bernoulli	Poisson	Exponential	Weibull
Model	$f(y)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$	$\pi^y(1-\pi)^{1-y}$	$\frac{e^{-\lambda}\lambda^y}{y!}$	$\varphi e^{-\varphi y}$	$\varphi\rho y^{\rho-1}e^{-\varphi y^\rho}$
Nat. param	η	μ	$\ln[\pi/(1-\pi)]$	$\ln\lambda$	$-\varphi$	
Mean function	$\psi(\eta)$	$\eta^2/2$	$\ln[1+\exp(\eta)]$	$\lambda=\exp(\eta)$	$-\ln(-\eta)$	
Norm. constant	$c(y,\phi)$	$-\frac{\ln(2\pi\phi)}{2}-\frac{y^2}{2\phi}$	0	$-\ln y!$	0	
(Over)dispersion	ϕ	σ^2	1	1	1	
Mean	μ	μ	π	λ	$-\varphi^{-1}$	$\varphi^{-1/\rho}\Gamma(\rho^{-1}+1)$
Variance	$\phi v(\mu)$	σ^2	$\pi(1-\pi)$	λ	φ^{-2}	$\varphi^{-2/\rho}\left[\Gamma(2\rho^{-1}+1)-\Gamma(\rho^{-1}+1)^2\right]$
Exponential family with conjugate random effects						
Model		Normal-normal	Beta-binomial	Negative binomial	Exponential-gamma	Weibull-gamma
Hier. model	$f(y\|\theta)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\theta)^2}{2\sigma^2}}$	$\theta^y(1-\theta)^{1-y}$	$\frac{e^{-\theta}\theta^y}{y!}$	$\varphi\theta e^{-\varphi\theta y}$	$\varphi\theta\rho y^{\rho-1}e^{-\varphi\theta y^\rho}$
RE model	$f(\theta)$	$\frac{1}{\sqrt{d}\sqrt{2\pi}}e^{-\frac{(\theta-\mu)^2}{2d}}$	$\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$
Marg. model	$f(y)$	$\frac{1}{\sqrt{\sigma^2+d}\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2(\sigma^2+d)}}$	$(\alpha+\beta)\frac{\Gamma(\alpha)}{\Gamma(\alpha+y)}\frac{\Gamma(\beta)}{\Gamma(\beta+1-y)}$	$\frac{\Gamma(\alpha+y)}{y!\Gamma(\alpha)}\left(\frac{\beta}{\beta+1}\right)^y\left(\frac{1}{\beta+1}\right)^\alpha$	$\frac{\varphi\alpha\beta}{(1+\varphi\beta y)^{\alpha+1}}$	$\frac{\varphi\rho y^{\rho-1}\alpha\beta}{(1+\varphi\beta y^\rho)^{\alpha+1}}$
	$h(\theta)$	θ	$\ln[\theta/(1-\theta)]$	$\ln(\theta)$	$-\theta$	$-\theta$
	$g(\theta)$	$-\frac{1}{2}\theta^2$	$-\ln(1-\theta)$	θ	$-\ln(\theta)/\varphi$	$-\ln(\theta)/\varphi$
	ϕ	σ^2	1	1	$1/\varphi$	$1/\varphi$
	γ	$1/d$	$\alpha+\beta-2$	$1/\beta$	$\varphi(\alpha-1)$	$\varphi(\alpha-1)$
	ψ	μ	$\frac{\alpha-1}{\alpha+\beta-2}$	$\beta(\alpha-1)$	$[\beta\varphi(\alpha-1)]^{-1}$	$[\beta\varphi(\alpha-1)]^{-1}$
	$c(y,\phi)$	$-\frac{1}{2}\phi y^2-\frac{1}{2}\ln\left(\frac{2\pi}{\phi}\right)$	0	$-\ln(y!)$	$\ln(\varphi)$	$\ln\left(\varphi\rho y^{\rho-1}\right)$
	$c^*(\gamma,\psi)$	$-\frac{1}{2}\gamma\psi^2-\frac{1}{2}\ln\left(\frac{2\pi}{\gamma}\right)$	$-\ln B(\gamma\psi+1,\gamma-\psi\gamma+1)$	$(1+\gamma\psi)\ln\gamma-\ln\Gamma(1+\gamma\psi)$	$\frac{\gamma+\varphi}{\varphi}\ln(\gamma\psi)-\ln\Gamma\left(\frac{\gamma+\varphi}{\varphi}\right)$	$\frac{\gamma+\varphi}{\varphi}\ln(\gamma\psi)-\ln\Gamma\left(\frac{\gamma+\varphi}{\varphi}\right)$
Mean	$\mathrm{E}(Y)$	μ	$\frac{\alpha}{\alpha+\beta}$	$\alpha\beta$	$[\varphi(\alpha-1)\beta]^{-1}$	$\frac{\Gamma(\alpha-\rho^{-1})\Gamma(\rho^{-1}+1)}{(\varphi\beta)^{1/\rho}\Gamma(\alpha)}$
Variance	$\mathrm{Var}(Y)$	σ^2+d	$\frac{\alpha\beta}{(\alpha+\beta)^2}$	$\alpha\beta(\beta+1)$	$\alpha[\varphi^2(\alpha-1)^2(\alpha-2)\beta^2]^{-1}$	$\frac{1}{\rho(\varphi\beta)^{2/\rho}\Gamma(\alpha)}\left[2\Gamma(\alpha-2\rho^{-1})\Gamma(2\rho^{-1})-\frac{\Gamma(\alpha-\rho^{-1})^2\Gamma(\rho^{-1})^2}{\rho\Gamma(\alpha)}\right]$

We will first define standard conjugacy, i.e., in models without the normal random effects and then, in Section 4, introduce a further property, *strong conjugacy*, necessary for situations where both normal and conventional conjugate random effects are present. To simplify notation, we will provide the definition at a general distribution level, with neither subject- nor measurement-specific subscripts, so that it can be applied to both univariate and longitudinal data. The hierarchical and random-effects densities are said to be conjugate if and only if they can be written in the generic forms:

$$f(y|\theta) = \exp\left\{\phi^{-1}[yh(\theta) - g(\theta)] + c(y,\phi)\right\}, \tag{3}$$

$$f(\theta) = \exp\left\{\gamma[\psi h(\theta) - g(\theta)] + c^*(\gamma,\psi)\right\}, \tag{4}$$

where $g(\theta)$ and $h(\theta)$ are functions, ϕ, γ, and ψ are parameters, and the additional functions $c(y,\phi)$ and $c^*(\gamma,\psi)$ are so-called normalizing constants. It can then be shown, upon constructing the joint distribution and then integrating over the random effect, that the marginal model resulting from (3) and (4) equals:

$$f(y) = \exp\left[c(y,\phi) + c^*(\gamma,\psi) - c^*\left(\phi^{-1}+\gamma, \frac{\phi^{-1}y + \gamma\psi}{\phi^{-1}+\gamma}\right)\right]. \tag{5}$$

Table 1 gives model elements, such as density or probability mass functions, conditional on random effects and marginalized over these, as well as the random effects distributions. For all models considered, the constants and functions featuring in (3)–(4) are listed, and finally marginal means and variances are provided. For some models, these are well known (Hinde and Demétrio, 1998ab) and/or easy to derive.

In the case of binary data, the model in Table 1 is the familiar beta-binomial model. Note that the variance still obeys the usual Bernoulli variance structure. This is entirely natural, given that we still focus on a single binary outcome, in contrast to the more conventional binomial basis model, where data of the format 'z_i successes out n_i trials' are considered. We do not consider this situation in this section, but rather leave it to Section 6. In such a case, the variance structure becomes $\pi_i(1-\pi_i)[1+\rho_i(n_i-1)]$, where ρ_i is a measure for correlation. All parameters, p_i and ρ_i, can be expressed in terms of α_i and β_i, 'cluster-specific' versions of the beta parameters.

For count data, the familiar negative-binomial model results. Unlike in the binary case, univariate counts are able to violate the mean-variance relationship inherent in the Poisson distribution, hence the great popularity of this and other types of models for overdispersion. The same applies to the exponential distribution. Of course, already the Weibull model, with its extra parameter ρ, alleviates the constraint.

The normal distribution case is a special one. Not only is it self-conjugate, also the model is not identified, unlike all others. This is because both random terms, seen from writing $Y_i = \mu_i + b_i + \varepsilon_i$, are in direct, linear relationship. In the generalized linear

context, the various random terms have no direct linear alliance. The normal case will continue to be 'the odd one out' in models to come (Sections 3.3 and 5-7).

The parameters α and β in the beta and gamma distributions are not always jointly identified. It is therefore customary to impose restrictions, such as setting one of them equal to a fixed value, e.g., $\alpha = 1$, or constraining their mean or variance, etc. Such constraints operate differently, depending on other elements present in the models. For example, the presence of additional random effects in a model for repeated measures, such as in Section 4, alters the meaning and restrictiveness of such constraints.

3.3. Models with normal random effects

The generalized linear mixed model (GLMM; Engel and Keen, 1994; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) is a straightforward extension of the linear mixed model (Verbeke and Molenberghs, 2000) to non-Gaussian hierarchical data. It is implemented in many standard software tools.

Let Y_{ij} be the jth outcome measured for cluster (subject) $i = 1,\ldots,N$, $j = 1,\ldots,n_i$ and group the n_i measurements into a vector $\boldsymbol{Y}_i$. Assume that, in analogy with Section 3.1, conditionally upon q-dimensional random effects $\boldsymbol{b}_i \sim N(\boldsymbol{0},\boldsymbol{D})$, the outcomes Y_{ij} are independent with densities:

$$f_i(y_{ij}|\boldsymbol{b}_i,\boldsymbol{\xi},\phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij},\phi)\right\}, \tag{6}$$

where

$$\eta[\psi'(\lambda_{ij})] = \eta(\mu_{ij}) = \eta[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i,\boldsymbol{\xi})] = \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i \tag{7}$$

for a known link function $\eta(\cdot)$, with $\boldsymbol{x}_{ij}^{\mathsf{T}}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, with $\boldsymbol{\xi}$ a p-dimensional vector of unknown fixed regression coefficients, and with ϕ a scale (overdispersion) parameter. Finally, let $f(\boldsymbol{b}_i|\boldsymbol{D})$ be the density of the $N(\boldsymbol{0},\boldsymbol{D})$ distribution for the random effects $\boldsymbol{b}_i$. These models closely follow the ones formulated in the top part of Table 1, with key differences that now: (a) data hierarchies are allowed for; (b) the natural parameter is written as a linear predictor, a function of both fixed and random effects.

4. Models combining conjugate and normal random effects

4.1. General model formulation

Combining overdispersion (Section 3.2) and normal random effects (Section 3.3) into the generalized linear model framework, produces the following general family:

$$f_i(y_{ij}|\boldsymbol{b}_i,\boldsymbol{\xi},\theta_{ij},\phi)=\exp\left\{\phi^{-1}[y_{ij}\lambda_{ij}-\psi(\lambda_{ij})]+c(y_{ij},\phi)\right\}, \tag{8}$$

The conditional mean follows as the product:

$$E\left(Y_{ij}|\boldsymbol{b_i},\boldsymbol{\xi},\theta_{ij}\right)=\mu_{ij}^c=\psi'(\lambda_{ij})=\theta_{ij}\kappa_{ij}, \tag{9}$$

where the random variable

$$\theta_{ij}\sim\Theta_{ij}\left(\upsilon_{ij},\sigma_{ij}^2\right), \tag{10}$$

with mean υ_{ij}, variance σ_{ij}^2, and the mean component

$$g(\kappa_{ij})=\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}+\boldsymbol{z}_{ij}^{\top}\boldsymbol{b_i} \tag{11}$$

depends on an $n_i\times p$ fixed-effects design $\boldsymbol{X}_i$ and a $n_i\times q$ random-effects design $\boldsymbol{Z}_i$ through a link function $g(\cdot)$; $\boldsymbol{\xi}$ and $\boldsymbol{b_i}\sim N(\boldsymbol{0},\boldsymbol{D})$ are fixed and random effects, respectively. The relationship between mean and natural parameter is

$$\lambda_{ij}=h(\mu_{ij}^c)=h(\theta_{ij}\kappa_{ij}). \tag{12}$$

The mean satisfies:

$$\mathrm{E}(Y_{ij})=\mathrm{E}(\theta_{ij})\,\mathrm{E}(\kappa_{ij})=\mathrm{E}[h^{-1}(\lambda_{ij})]. \tag{13}$$

Depending of the type of outcome under investigation, the distribution of θ_{ij} can be chosen appropriately.

It is computationally convenient, but not strictly necessary, to assume that the sets of random effects, $\boldsymbol{\theta}_i$ and $\boldsymbol{b}_i$, are independent. Kalema and Molenberghs (2015) and Kalema, Iddi, and Molenberghs (2016) relaxed this assumption. Regarding the components θ_{ij} of $\boldsymbol{\theta}_i$, three special cases are: (1) independence; (2) correlated, implying that the univariate distributions $\mathscr{G}_{ij}(\vartheta_{ij},\sigma_{ij}^2)$ must be replaced with a multivariate one; and (3) equal (useful in applications with exchangeable outcomes Y_{ij}).

4.2. Strong conjugacy

It is of interest to explore under what conditions Model (8) still allows for conjugacy, now that normal random effects have been introduced into the linear predictor, leading to the multiplicative factor κ_{ij} in the mean structure. To this end, MVDV considered conjugacy conditional upon the normally-distributed random effect $\boldsymbol{b}_i$. Write in simplified notation:

$$f(y|\kappa\theta) = \exp\left\{\phi^{-1}[yh(\kappa\theta) - g(\kappa\theta)] + c(y,\phi)\right\}, \tag{14}$$

generalizing (3), and retain (4). Applying the transformation theorem to (4) leads to

$$f(\theta|\gamma,\psi) = \kappa \cdot f(\kappa\theta|\widetilde{\gamma},\widetilde{\psi}),$$

where $\widetilde{\gamma}$ and $\widetilde{\psi}$ are appropriate parameters. Next, we request that the parametric form (4) be maintained:

$$f(\kappa\theta) = \exp\left\{\gamma^*[\psi^* h(\kappa\theta) - g(\kappa\theta)] + c^{**}(\gamma^*,\psi^*)\right\}, \tag{15}$$

where the parameters γ^* and ψ^* follow from $\widetilde{\gamma}$ and $\widetilde{\psi}$ upon absorption of κ. Then, the marginal model, in analogy with (5), equals:

$$f(y|\kappa) = \exp\left\{c(y,\phi) + c^{**}(\gamma^*,\psi^*) + c^{**}\left(\phi^{-1}+\gamma^*, \frac{\phi^{-1}y + \gamma^*\psi^*}{\phi^{-1}+\gamma^*}\right)\right\}. \tag{16}$$

Not every model satisfying conjugacy in the sense of Section 3.2 allows for this form of conjugacy, referred to as *strong conjugacy*. Examples include the normal, Poisson, and Weibull (and hence exponential) models with normal, gamma, and gamma random effects, respectively. A counterexample is provided by the Bernoulli, and hence also binomial, model. Because the probit model does not allow for conjugacy, it is out of the picture here, too. The latter does not preclude the existence of closed forms in the probit case, as was shown by MVDV. These authors noted that strong conjugacy stems from the random-effects distribution, not from the data model. For example, they showed, for a gamma random effect:

$$\frac{1}{\kappa} f(\theta|\alpha,\beta) = f(\kappa\beta|\alpha,\kappa\beta), \tag{17}$$

and hence a scaled version of a gamma random effect is still a gamma random effect, with invariant α and re-scaled β.

Strong conjugacy facilitates the use of standard software, which does not imply that such software cannot be used once strong conjugacy does not hold. Arguably, the derivation of analytic quantities, such as moments, and hence means, variances, and covariances, is simplified when the property holds.

All CM can be formulated using the same general principles. One simply has to combine the models formulated in Table 1 with the GLMM (6) and corresponding linear predictor (7). The effect θ is then replaced by $\theta_{ij}\kappa_{ij}$, where κ_{ij} is defined by setting $\eta = \eta_{ij}$ equal to the linear predictor whence κ_{ij} is expressed, for the respective models, as μ, π, λ, and ϕ.

5. Count data

The model elements in this case are:

$$Y_{ij} \sim \text{Poi}(\theta_{ij}\kappa_{ij}), \tag{18}$$

$$\kappa_{ij} = \exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i\right), \tag{19}$$

$$\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D}), \tag{20}$$

$$\text{E}(\boldsymbol{\theta}_i) = \text{E}[(\theta_{i1}, \ldots, \theta_{in_i})^{\top}] = \boldsymbol{\vartheta}_i, \tag{21}$$

$$\text{var}(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_i. \tag{22}$$

This model has the same structure as the one by Booth et al. (2003). In the spirit of Table 1, the θ_{ij} can be assumed to follow a gamma model, producing, what we could term, a Poisson-gamma-normal model (PGN). Recall that $\boldsymbol{b}_i$ accommodates correlation and some overdispersion, while residual overdispersion is captured by the components θ_{ij} of $\boldsymbol{\theta}_i$. Should these components be assumed dependent, then both sets of random effects capture some correlation as well as some overdispersion. In the correlated case, a multivariate extension of the gamma distribution would be needed (see, for example, Gentle, 2003).

This model enjoys strong conjugacy, as shown by MVDV. Continuing on the work of Zeger, Liang, and Albert (1988), and using expressions for the standard Poisson moments (Johnson, Kemp, and Kotz, 2005, p. 162), MVD derived the moments; conditional upon the random effects are:

$$\text{E}(Y_{ij}^k) = \sum_{\ell=0}^{k} S(k,\ell)(\theta_{ij}\kappa_{ij})^{\ell}, \tag{23}$$

where $S(k,\ell)$ is the so-called Stirling number of the second kind. Integrating (23) over the random effects produces:

$$\text{E}(Y_{ij}^k) = \sum_{\ell=0}^{k} S(k,\ell)\frac{\beta^{\ell}\Gamma(\alpha+\ell)}{\Gamma(\alpha)}\exp\left[\ell\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \tfrac{1}{2}\ell^2\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right]. \tag{24}$$

The mean components are:

$$\mu_{ij} = \phi_{ij}\exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \tfrac{1}{2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right), \tag{25}$$

with the variance-covariance matrix

$$\text{var}(\boldsymbol{Y}_i) = \boldsymbol{M}_i + \boldsymbol{M}_i(\boldsymbol{P}_i - \boldsymbol{J}_{n_i})\boldsymbol{M}_i, \tag{26}$$

where $\boldsymbol{M}_i$ is a diagonal matrix with the μ_{ij} along the main diagonal, and the $(j,k)^{\text{th}}$ element of $\boldsymbol{P}_i$ equals

$$p_{i,jk} = \exp\left(\tfrac{1}{2}\boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ik}\right)\cdot\frac{\sigma_{i,jk}+\phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}}\cdot\exp\left(\tfrac{1}{2}\boldsymbol{z}_{ik}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ij}\right). \tag{27}$$

MVD also derived a series-based expression for the marginal joint distribution:

$$\begin{aligned} P(\boldsymbol{Y}_i=\boldsymbol{y}_i) = \sum_{\boldsymbol{t}}\left[\prod_{j=1}^{n_i}\begin{pmatrix} y_{ij}+t_j \\ y_{ij}\end{pmatrix}\cdot\begin{pmatrix}\alpha_j+y_{ij}+t_j-1\\ \alpha_j-1\end{pmatrix}\cdot(-1)^{t_j}\cdot\beta_j^{y_{ij}+t_j}\right] \\ \times\exp\left(\sum_{j=1}^{n_i}(y_{ij}+t_j)\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}\right) \\ \times\exp\left(\frac{1}{2}\left[\sum_{j=1}^{n_i}(y_{ij}+t_j)\boldsymbol{z}_{ij}^{\mathsf{T}}\right]\boldsymbol{D}\left[\sum_{j=1}^{n_i}(y_{ij}+t_j)\boldsymbol{z}_{ij}\right]\right). \end{aligned} \tag{28}$$

In the above equation, the vector-valued index $\boldsymbol{t}=(t_1,\ldots,t_{n_i})^{\mathsf{T}}$ ranges over all non-negative integer vectors.

In Section 9, the benefit of having closed-form expressions will show when deriving quantities such as marginal correlations.

Kalema and Molenberghs (2015) and Kalema, Iddi, and Molenberghs (2016) showed how the combined model formulation can be used to generate correlated count data. Neyens, Faes, and Molenberghs (2012) adapted the framework to accommodate overdispersion in counts that arise in a spatial context.

5.1. A clinical trial in epileptic patients

We will analyse the epilepsy data, introduced in Section 2.1. Let Y_{ij} represent the number of epileptic seizures patient i experiences during week j of the follow-up period. Also, let t_{ij} be the time-point at which Y_{ij} has been measured, $t_{ij}=1,2,\ldots$ until at most 27. Consider the combined model (18)–(22), with specific choices

$$\ln(\kappa_{ij}) = \begin{cases} (\xi_{00}+b_i)+\xi_{01}t_{ij} & \text{if placebo} \\ (\xi_{10}+b_i)+\xi_{11}t_{ij} & \text{if treated,}\end{cases} \tag{29}$$

where the random intercept b_i is assumed to be zero-mean normally distributed with variance d. We consider special cases (a) the ordinary Poisson model (P--), (b) the negative-binomial model (PG-), (c) the Poisson-normal model (P-N), together with (d)

Table 2: *Epilepsy study. Parameter estimates (standard error) in (1) Poisson model (P--), (2) negative-binomial model (PG-), (3) Poisson-normal model P-N), and (4) combined model (PGN), as well as their zero-inflated counterparts ZI(P--), ZI(PG-), ZI(P-N), ZI(PGN).*

		Combined models		Negative-binomial models	
Effect	Par.	ZI(PGN)	(PGN)	ZI(PG-)	(PG-)
Interc. plac.	ξ_{00}	0.947(0.167)	0.911(0.176)	1.236(0.110)	1.259(0.0.112)
Slope plac.	ξ_{01}	−0.016(0.008)	−0.025(0.008)	−0.007(0.011)	−0.013(0.011)
Interc. treatm.	ξ_{10}	0.836(0.172)	0.656(0.178)	1.397(0.110)	1.475(0.109)
Slope treatm.	ξ_{11}	−0.006(0.007)	−0.012(0.008)	−0.022(0.011)	−0.035(0.010)
Neg.-bin. par.	α_1	0.245(0.025)	2.464(0.211)	1.787(0.100)	0.527(0.026)
SD non-zero part RE	$\sqrt{d_1}$	0.997(0.085)	1.063(0.087)	–	–
Infl. Interc.	γ_0	−4.581(0.641)	–	−7.106(1.334)	–
Infl. slope	γ_1	0.092(0.034)	–	0.292(0.066)	–
SD zero part RE	$\sqrt{d_2}$	2.533(0.440)	-	–	–
Corr. RE	ρ	−0.096(0.153)	–	–	–
Pred. prob. zeros		0.352	0.321	0.185	0.158
−2log-likelihood		5317.9	5417.0	6318.9	6326.1
		Poisson-normal models		Poisson models	
Effect	Par.	ZI(P-N)	(P-N)	ZI(P--)	(P--)
Interc. plac.	ξ_{00}	0.903(0.155)	0.818(0.168)	1.485(0.043)	1.266(0.0.042)
Slope plac.	ξ_{01}	−0.004(0.005)	−0.014(0.004)	−0.007(0.005)	−0.0.013(0.004)
Interc. treatm.	ξ_{10}	0.908(0.159)	0.648(0.170)	1.806(0.040)	1.453(0.038)
Slope treatm.	ξ_{11}	−0.007(0.005)	−0.012(0.004)	−0.025(0.014)	−0.033(0.004)
SD non-zero part RE	$\sqrt{d_1}$	0.971(0.082)	1.076(0.086)	–	–
Infl. Interc.	γ_0	−3.712(0.500)	–	−0.659(4.699)	–
Infl. slope	γ_1	0.095(0.025)	–	−3.291(4.444)	–
SD zero part RE	$\sqrt{d_2}$	2.222(0.343)	–	–	–
Corr. RE	ρ	−0.154(0.157)	–	–	–
Pred. prob. zeros		0.338	0.263	0.014	0.046
−2log-likelihood		5845.1	6271.9	10912	11590

the combined model (PGN). Estimates (standard errors) are presented in Table 2. The table also contains zero-inflated versions, that will be discussed in Section 5.2. Clearly, both the negative-binomial model and the Poisson-normal model are important improvements, in terms of the likelihood, relative to the ordinary Poisson model. This should come as no surprise since the latter unrealistically assumes there is neither overdispersion nor correlation within the outcomes, while clearly both are present. In addition, when considering the combined model, there is a very strong improvement in fit when gamma and normal random effects are simultaneously allowed for. This strongly affects the point and precision estimates of such key parameters as the slope difference and the slope ratio. There is also an impact on hypothesis testing. The Poisson model leads to unequivocal significance for both the difference ($p = 0.0008$) and ratio ($p = 0.0038$),

whereas for the Poisson normal this is not the case for the difference of the slopes ($p = 0.7115$), while some significance is maintained for the ratio ($p = 0.0376$). Because the Poisson-normal is commonly used, it is likely that in practice one would decide in favor of a treatment effect when considering the slope ratio. This is no longer true with the negative-binomial model, where the p-values change to $p = 0.01310$ and $p = 0.2815$, respectively. Of course, one must not forget that, while the negative-binomial model accommodates overdispersion, the θ_{ij} random effects are assumed independent, implying independence between repeated measures. Again, this is not realistic and therefore the combined model is a more viable candidate, corroborated further by the aforementioned likelihood comparison. This model produces non-significant p-values of $p = 0.2260$ and $p = 0.1591$, respectively.

Thus, in conclusion, whereas the conventionally used and broadly implemented Poisson-normal model would suggest a significant effect of treatment, our combined model issues a message of caution, because there is no evidence whatsoever regarding a treatment difference.

Molenberghs and Verbeke (2005, Ch. 19), considered a (P-N) model with random intercepts as well as random slopes in time. It is interesting to note that, when allowing for such an extension in our models, the random slopes improve the fit of the (P-N) model with random intercept, but not of the combined one with random intercept (details not shown). As a consequence, the combined model with random intercept is the best fitting one. At the same time, note that fitting such a model establishes that the presence of a conjugate random effect does not preclude the consideration of normal random effects beyond random intercepts. The data were analysed by Booth et al. (2003), too.

Let us now turn to the correlation functions. Given that the gamma random effects are assumed independent, we only need to consider the Poisson-normal and combined cases; the versions with and without random slopes are considered. Because the fixed-effects structure is not constant but rather depends on time, MVD formulated a correlation function. In the (P-N) case with random intercepts only, and for the placebo group, based on the parameter estimates in Table 2, they obtained:

$$\mathrm{Corr}(Y(t),Y(s)) = \frac{35.58 \cdot 0.99^{t+s}}{\sqrt{(4.04 \cdot 0.99^{t} + 35.58 \cdot 0.97^{t}) \cdot (4.04 \cdot 0.99^{s} + 35.58 \cdot 0.97^{s})}},$$

where $Y(t)$ represents the outcome for an arbitrary subject at time t. Calculations in all other cases are similar. The smallest and largest values for the correlation functions, for both arms, for both the Poisson-normal and combined models, and for both choices of the random-effects structure are given in Table 3. When only random intercepts are considered, the correlations range over a narrow interval; they are rather high and there is little difference between the Poisson-normal and combined models. However, turning to the models with random intercepts and random slopes, several differences become apparent. First, the values exhibit a much broader range between their smallest and

largest values. Second, the range is somewhat over-estimated by the Poisson-normal model, which then narrows when we switch to the combined model, thereby incorporating overdispersion effects, random intercepts, and random slopes. Thus, the random slope allows for the correlation to range over a considerable interval, while the overdispersion effect prevents the range from becoming overly wide.

Table 3: *Epilepsy study. Observed smallest and largest values for the correlation function, for the Poisson-normal and combined models, and for both treatment arms. The time pair for which the values are observed is shown too. (RI: random intercept; RS: random slope.)*

		Smallest value		Largest value	
Model	Arm	ρ	time pair	ρ	time pair
Poisson-normal, RI	placebo	0.8577	26 & 27	0.8960	1 & 2
Poisson-normal, RI	treatment	0.8438	26 & 27	0.8794	1 & 2
Combined, RI	placebo	0.8259	26 & 27	0.8981	1 & 2
Combined, RI	treatment	0.8383	26 & 27	0.8744	1 & 2
Poisson-normal, RI+RS	placebo	0.2966	1 & 27	0.9512	26 & 27
Poisson-normal, RI+RS	treatment	0.2936	1 & 27	0.9530	26 & 27
Combined, RI+RS	placebo	0.4268	1 & 27	0.9281	26 & 27
Combined, RI+RS	treatment	0.4225	1 & 27	0.9329	26 & 27

Within each model, there is relatively little difference between the placebo and treated groups, although the difference is a bit more pronounced in the combined model. Further, the correlation range within every group is relatively narrow. The most noteworthy feature, unquestionably, is the large discrepancy between both models. This is because the (P-N) model forces the correlation and overdispersion effects to stem from a single additional parameter, the random-intercept variance d. Thus, considerable overdispersion also forces the correlation to increase, arguably beyond what is consistent with the data. In the combined model, in contrast, there are *two* additional parameters, giving proper justice to both correlation and overdispersion effects. It was already clear from the above discussion and that in MVD that the combined model is an important improvement. This now clearly manifests itself in the correlation function, too.

The above underscores the need for the combined model. Some indication came, for example, from the correlation functions in the epilepsy case. It is useful to perform formal comparison of all nested models, using Wald statistics, for each of the three cases. A summary is given in Table 4. Note that, owing to the familiar boundary problem that occurs when testing for variance components, mixtures of a χ^2_0 and χ^2_1 were used, instead of the conventional χ^2_1 (Molenberghs and Verbeke, 2007).

Table 4: *Epilepsy, onychomycosis, and asthma studies. Wald test results for comparison of nested models.*

Null model	Alternative model	Z-value	p-value
	Epilepsy study		
Poisson	Negative-binomial	20.68	<0.0001
Poisson	Poisson-normal	6.27	<0.0001
Negative-binomial	Combined	6.10	<0.0001
Poisson-normal	Combined	11.66	<0.0001
	Onychomycosis study		
Logistic	Beta-binomial	17.91	<0.0001
Logistic	Logistic-normal	10.53	<0.0001
Beta-binomial	Combined	4.28	<0.0001
Logistic-normal	Combined	8.01	<0.0001
	Asthma study		
Exponential	Exponential-gamma	8.54	<0.0001
Exponential	Exponential-normal	10.63	<0.0001
Exponential-gamma	Combined	8.54	<0.0001
Exponential-normal	Combined	3.99	<0.0001

For our case study, it is clear that: (a) independence is strongly rejected in favour of both a model with normal random effects or a model with conjugate random effects; (b) on top of one set of random effects, there is a clear need for the other set as well, hence providing very strong evidence for the proposed combined model. The evidence is extremely convincing. The table also contains results for two more case studies that will be discussed in detail in subsequent sections.

These findings, taken together, imply that the data exhibit, at the same time, within-subject correlation and overdispersion, in such a way that a single model feature cannot capture both simultaneously.

5.2. Additional zeroes

It is not uncommon when count data are collected to observe more zeroes than predicted by the model assumed, whether of a simple Poisson nature, or more elaborate, such as the combined model considered here. This feature, often referred to as zero inflation, then needs to be accommodated, in addition to correlation and/or overdispersion. Such data are often fitted by using either hurdle (Mullahy, 1986; Greene, 1994) or zero-inflated models (ZI; Lambert, 1992). In the context of the CM, additional zeroes were studied by Kassahun et al. (2014a) and Iddi and Molenberghs (2013).

We will first describe the hurdle (H) and zero-inflation (ZI) approaches for univariate data, and then turn to hierarchical versions. The hurdle model is a two-part model, whereby the first part is a binary model for the count value to be either zero or positive. Given that the value is positive, a count distribution, say f_i, is truncated at zero and fitted

to the second part. Suppose Y_i is a univariate count outcome, and π_i is the probability of the i^{th} observation to be in the zero state. The hurdle model then takes the form:

$$p(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \begin{cases} \pi_i & \text{if } y_i = 0, \\ (1-\pi_i)\frac{f_i(y_i|\lambda_i)}{1-f_i(0|\lambda_i)} & \text{if } y_i > 0. \end{cases} \tag{30}$$

An alternative approach is a zero-inflated model, which assumes zeros to come from two processes. The first process generates only zeros with probability π_i for observation i, say, while the second process generates counts with probability $1-\pi_i$. The ZI model is:

$$p(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \begin{cases} \pi_i + (1-\pi_i)f_i(0|\lambda_i) & \text{if } y_i = 0, \\ (1-\pi_i)f_i(y_i|\lambda_i) & \text{if } y_i > 0. \end{cases} \tag{31}$$

Here, π_i and λ_i are functions of covariates. Link functions, such as the logit or probit, can be used for π_i, with the log link commonly used for λ_i.

Kassahun et al. (2014a) extended the combined model to take zero-inflation into account. The ZI version of the CM (ZICOM) is given by

$$p(Y_{ij} = y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij}) = \begin{cases} \pi_{ij} + (1-\pi_{ij})f_i(0|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{if } y_{ij} = 0, \\ (1-\pi_{ij})f_i(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{if } y_{ij} > 0. \end{cases} \tag{32}$$

The ZI component $\pi_{ij} = \pi(\boldsymbol{x}_{2ij}^{\mathsf{T}}\boldsymbol{\gamma} + \boldsymbol{z}_{2ij}^{\mathsf{T}}\boldsymbol{b}_{2i})$ is modelled using a Bernouilli model: in the simplest case with only an intercept, but potentially containing known regressors $\boldsymbol{x}_{2ij}$ and $\boldsymbol{z}_{2ij}$, a vector of zero-inflation coefficients $\boldsymbol{\gamma}$ to be estimated, as well as random effects $\boldsymbol{b}_{2i}$. Common link functions, such as the logit or probit, can be used. Note that $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$, and $\boldsymbol{b}_i$ in Section 4 are now replaced by $\boldsymbol{x}_{1ij}$, $\boldsymbol{z}_{1ij}$, and $\boldsymbol{b}_{1ij}$, respectively, for the non-zero count part. The regressors in the count and zero-inflation component can either be overlapping, a subset of the regressors can be used for the zero-inflation, or entirely different regressors for the two parts can be used. In many cases, but of course not always, a simple random-intercept model is adequate, where $\boldsymbol{b}_{1i} = b_{1i}$, $\boldsymbol{b}_{2i} = b_{2i}$, and $\boldsymbol{z}_{1ij} = \boldsymbol{z}_{2ij} = 1$. The variance-covariance matrix of the random effects, assumed normally distributed, is denoted by D, as before. The model is denoted as ZI(PGN), as an obvious extension with earlier notational conventions. Three obvious special cases are ZI(P-N), ZI(PG-), and ZI(P--). Also, all four models without zero inflation are special cases as well. The conditional mean and variance of the ZI(PGN) are:

$$\mathrm{E}(Y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) = \theta_{ij}\kappa_{ij}(1-\pi_{ij}), \tag{33}$$

$$\mathrm{Var}(Y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) = \theta_{ij}\kappa_{ij}(1-\pi_{ij})[1 + \theta_{ij}\kappa_{ij}(\pi_{ij} + 1/\alpha)]. \tag{34}$$

It can be seen that the conditional variance is inflated as a result of either overdispersion in the data (parameter α), or as a result of ZI (parameter π_{ij}), or both.

Further model developments that allow for extra zeroes are reported in Sections 10 and 11.

5.3. A clinical trial in epileptic patients

We re-analyse the epilepsy data, introduced in Section 2.1 and analysed before in Section 5.1. Let Y_{ij} represent the number of epileptic seizures that patient i experiences during week j of the follow-up period. Also, let t_{ij} be the time-point at which Y_{ij} has been recorded. Consider parameterization (29), but now accounting for zero inflation, assuming that counts are generated from a (P-N) process with λ_{ij} as in (29), or from a (PGN) process with mean $\lambda_{ij} = \theta_{ij}\kappa_{ij}$, and now κ_{ij} specified as in (29). The ZI probability (π_{ij}) is modelled as $\text{logit}(\pi_{ij}) = \gamma_0 + b_{2i} + \gamma_1 t_{ij}$. The data are analysed with the ZI(PGN), ZI(PG-), ZI(P-N), ZI(P--). One can compare the results with the non-ZI counterpart. Parameter estimates and predicted probabilities of zeros are presented in Table 2, alongside the non-ZI counterparts. Clearly, in terms of likelihood comparison, the zero-inflated versions performed much better, resulting in a substantial improvement in fit.

The ZI(PG-) is an important improvement relative to the ZI(P--), while much more improvement is gained in the case of the ZI(P-N). Moreover, the ZI(PGN) leads to a substantially improved fit. Further, we observe that, omitting either the overdispersion or the correlation underestimates the predicted probability of zeros, which becomes worse when both are omitted at the same time. The ZI(PGN), fitted without random effects in the zero-inflation part, results in -2log-likelihood of 5386.8, and predicted probability of zeros equal to 0.3271. This implies that inclusion of random effects in the zero-inflation part tends to have little impact on the predicted probability of zeros. However, based on likelihood comparison, model fit improves considerably. This same phenomenon is also evident in the ZI(P-N) fitted with random effects included only in the non-zero count part (-2log-likelihood is 5971.9, and predicted probability of zeros 0.3112).

None of the zero-inflated models suggests evidence of significance in slope difference and slope ratio, except for the ZI(P--), where significance is maintained for the slope difference ($p = 0.004$). However, the latter, unrealistically, omits correlation and overdispersion. The zero-inflation regression coefficients can be interpreted as model coefficients for the proportion of extra zeros, and are statistically significant in all except the ZI(P--). Evidently, models can be extended further. For example, one could consider a version with where the ZI component is specific to treatment arm.

6. Categorical data

Categorical data come in various forms, and we usefully distinguish between them. Building on MVDV, Molenberghs et al. (2012) laid out the combined-model framework and various ramifications for the binary and binomial cases. An overview will be given in Sections 6.1 and 6.2 for the binary cases with logit and probit links, respectively, and in Section 6.3 for binomial data. The iron deficiency case study is analysed in Section 6.4. An application of the binary version of the model to the Jimma Infant study was reported in Kassahun et al. (2012). A binomial application is described in Del Fava et al. (2014). Ivanova, Molenberghs, and Verbeke (2014) developed a version of the combined model to handle ordinal data, which is the basis for Section 6.5.

6.1. Bernoulli-type models for binary data with logit link

Similar to the Poisson case in Section 5, a natural binary-data counterpart to (18)–(19) is

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij} = \theta_{ij}\kappa_{ij}), \tag{35}$$

$$\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i\right)}{1 + \exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i\right)}, \tag{36}$$

completing the specification with (20)–(22). Unlike in the Poisson case, closed forms for neither the mean nor the variance follow when normal random effects are present. When only overdispersion random effects are included, especially when they are assumed to follow a beta distribution, as in Table 1, conjugacy applies. However, the beta distribution does not allow for the multiplicative invariance as (17), precluding strong conjugacy.

When the overdispersion random effects are assumed to be equal: $\theta_{ij} = \theta_i$, then the beta-binomial model follows if no normal random effects are present.

Explicitly considering $\theta_{ij} \sim \text{Beta}(\alpha, \beta)$, then $\phi_{ij} = \text{E}(\theta_{ij}) = \alpha/(\alpha+\beta)$, and

$$\sigma_{ij}^2 = \text{var}(\theta_{ij}) = \sigma_{i,jj} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

$$\sigma_{i,jk} = \text{cov}(\theta_{ij}, \theta_{ik}) = \rho_{ijk}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Observe that there are two correlations: ρ_{ijk}, which described the correlation between draws from the beta distribution and $(\alpha+\beta+1)^{-1}$. It is of course possible to let α and β vary with *i* and/or *j*. In such cases, the above and below expressions will change somewhat, but computations are straightforward.

Using the general expressions, the above results can be used to derive approximate expressions for means and variance-covariance elements. For the special case of no normal random effects, but maintaining the fixed effects in (36), i.e.,

$$\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}\right)}{1+\exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}\right)}, \tag{37}$$

we obtain

$$\mathrm{E}(Y_{ij}) = \frac{\alpha}{\alpha+\beta}\kappa_{ij}, \tag{38}$$

$$\mathrm{Var}(Y_{ij}) = \frac{\alpha}{\alpha+\beta}\kappa_{ij} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \kappa_{ij}^2,$$

$$\mathrm{Cov}(Y_{ij},Y_{ik}) = \rho_{ijk}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\kappa_{ij}\kappa_{ik}.$$

If we further make exchangeability assumptions, i.e., $\kappa_{ij} = \kappa_{ik} \equiv \kappa_i$ and $\rho_{ijk} = \rho_i$, further simplification follows. Finally, setting $\kappa_i = 1$, the conventional beta-binomial follows. It is then easy to derive the resulting binomial version by defining:

$$Z_i = \sum_{i=1}^{n_i} Y_{ij}. \tag{39}$$

Simple algebra then shows:

$$\mathrm{E}(Z_i) = n_i\frac{\alpha}{\alpha+\beta} = n_i\pi_i,$$

$$\mathrm{Var}(Z_i) = n_i\frac{\alpha\beta}{(\alpha+\beta)^2}\left\{1+(n_i-1)\frac{1}{\alpha+\beta+1}\right\} = n_i\pi_i(1-\pi_i)\left\{1+(n_i-1)\widetilde{\rho}_i\right\},$$

with $\widetilde{\rho}_i$ the beta-binomial correlation. Hence, the conventional beta-binomial model follows.

While the logit link defeats closed-form expressions when normal random effects are introduced, this is different with the probit link. The random-effects probit model has received some attention in earlier decades (Schall, 1991; Guilkey and Murphy, 1993; Hedeker and Gibbons, 1994; McCulloch, 1994; Gibbons and Hedeker, 1997; Renard, Molenberghs, and Geys 2004).

6.2. Bernoulli-type models for binary data with probit link

Introducing the probit version of the model, while at the same time assuming that the overdispersion parameters are beta distributed, comes down to:

$$\kappa_{ij} = \Phi_1(\boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\intercal}\boldsymbol{b}_i), \tag{40}$$

$$\theta_{ij} \sim \text{Beta}(\alpha, \beta). \tag{41}$$

Like before, α and β could be allowed to vary with i and/or j.

It now follows that the joint distribution can be written as (see MVDV):

$$f_{n_i}(\boldsymbol{y}_i = \boldsymbol{1}) = \left(\frac{\alpha}{\alpha+\beta}\right)^{n_i} \cdot \Phi_{n_i}(X_i\boldsymbol{\xi}; \boldsymbol{L}_{n_i}^{-1}), \tag{42}$$

with

$$\boldsymbol{L}_{n_i} = \boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\left(\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^{\intercal}\boldsymbol{Z}_i\right)^{-1}\boldsymbol{Z}_i^{\intercal}. \tag{43}$$

Note that (42) is the joint probability only for the outcome $(1,\ldots,1)^{\intercal}$, a so-called success probability. However, given that the dimension n_i is arbitrary, all other probabilities can be derived by appropriate contrasts of success probabilities. Precisely,

$$f_{n_i}[\boldsymbol{y}_i = \boldsymbol{m}_i = (m_{i1},\ldots,m_{in_i})^{\intercal}] = \sum_{\boldsymbol{s}\supset\iota(\boldsymbol{m}_i)} \text{sgn}(\boldsymbol{s})\Phi_{\#\boldsymbol{s}}\left(\widetilde{X}_i^{(\boldsymbol{s})}\boldsymbol{\xi}; \boldsymbol{L}_{(\boldsymbol{s})}^{-1}\right) \cdot \left(\frac{\alpha}{\alpha+\beta}\right)^{\#\boldsymbol{s}}, \tag{44}$$

with $\iota(\boldsymbol{m}_i) = \lambda(m_{i1},\ldots,m_{in_i})$ the set of places for which $m_{ij} = 1$,

$$\text{sgn}(\boldsymbol{s}) = \begin{cases} 1 & \text{if } \#\boldsymbol{s} - \#\iota(\boldsymbol{m}_i) \text{ is even,} \\ 0 & \text{otherwise,} \end{cases}$$

$\widetilde{X}_i^{(\boldsymbol{s})}$ contains the rows from X_i with row number in $\boldsymbol{s}$, and $\boldsymbol{L}_{(\boldsymbol{s})}$ is the $\#s$-dimensional matrix built from the appropriate sub-matrices of these used in (43). The above developments straightforwardly generalize when (41) is replaced with $\theta_{ij} \sim \text{Beta}(\alpha_j, \beta_j)$.

Next, the means, variances, and covariances can be derived from (42), by evaluating it for the one- and two-dimensional cases. We find:

$$\text{E}(Y_{ij}) = \frac{\alpha}{\alpha+\beta} \cdot \Phi_1(\boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\xi}; \boldsymbol{L}_1^{-1}) = \frac{\alpha}{\alpha+\beta} \cdot \Phi_1(|\boldsymbol{I} + \boldsymbol{D}\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}^{\intercal}|^{-1/2}\boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\xi}), \tag{45}$$

$$\text{Var}(Y_{ij}) = \frac{\alpha}{\alpha+\beta} \cdot \Phi_1(\boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\xi}; \boldsymbol{L}_1^{-1}) \cdot \left[1 - \frac{\alpha}{\alpha+\beta}.\Phi_1(\boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\xi}; \boldsymbol{L}_1^{-1})\right], \tag{46}$$

$$\mathrm{Cov}(Y_{ij},Y_{ik}) = \left(\frac{\alpha}{\alpha+\beta}\right)^2 \cdot \left\{ \Phi_2\left[\begin{pmatrix} \boldsymbol{x}_{ij}^\top \\ \boldsymbol{x}_{ik}^\top \end{pmatrix}\boldsymbol{\xi}, \boldsymbol{L}_{2jk}^{-1}\right] - \Phi_1(\boldsymbol{x}_{ij}^\top\boldsymbol{\xi};\boldsymbol{L}_{1j}^{-1})\Phi_1(\boldsymbol{x}_{ik}^\top\boldsymbol{\xi};\boldsymbol{L}_{1k}^{-1}) \right\}, \tag{47}$$

where

$$\boldsymbol{L}_{2jk} = \boldsymbol{I}_2 - \begin{pmatrix} \boldsymbol{z}_{ij}^\top \\ \boldsymbol{z}_{ik}^\top \end{pmatrix}\left[\boldsymbol{D}^{-1} + \begin{pmatrix} \boldsymbol{z}_{ij}^\top \\ \boldsymbol{z}_{ik}^\top \end{pmatrix}(\boldsymbol{z}_{ij}\ \boldsymbol{z}_{ik})\right]^{-1}(\boldsymbol{z}_{ij}\boldsymbol{z}_{ik}).$$

The rightmost density in (45) is the standard normal one. Evidently, (42) and (44) lead, not only to the mean, variance, and covariance expressions, but also to the higher-order moments.

MVDV noted that the existence of closed-form expressions for the probit case opens a window of opportunity for the logit case. Indeed, the well-known approximation formulae, linking the normal and logistic densities, prove useful here. As shown in Johnson and Kotz (1970, p. 6) and used in Zeger et al. (1988):

$$\frac{e^y}{1+e^y} \approx \Phi_1(cy), \tag{48}$$

with $c = (16\sqrt{3})/(15\pi)$. Applied to (35)–(36), it follows that

$$\pi_{ij} \sim \theta_{ij}\frac{\exp\left(\boldsymbol{x}_{ij}^\top\boldsymbol{\xi}+\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i\right)}{1+\exp\left(\boldsymbol{x}_{ij}^\top\boldsymbol{\xi}+\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i\right)} \approx \theta_{ij}\Phi_1[c(\boldsymbol{x}_{ij}^\top\boldsymbol{\xi}+\boldsymbol{z}_{ij}^\top\boldsymbol{b}_i)]. \tag{49}$$

Applying (49) to (42), yields

$$f_{n_i}(\boldsymbol{y}_i = \boldsymbol{1}) \approx \left(\frac{\alpha}{\alpha+\beta}\right)^{n_i} \cdot \Phi_{n_i}\left(cX_i\boldsymbol{\xi};\widetilde{\boldsymbol{L}}_{n_i}^{-1}\right), \tag{50}$$

with

$$\widetilde{\boldsymbol{L}}_{n_i} = \boldsymbol{I}_{n_i} - c^2\boldsymbol{Z}_i\left(\boldsymbol{D}^{-1}+\boldsymbol{Z}_i^\top\boldsymbol{Z}_i\right)^{-1}\boldsymbol{Z}_i^\top.$$

For the expectation, we find, based on (49) and (45):

$$\mathrm{E}(Y_{ij}) \approx \frac{\alpha}{\alpha+\beta}\cdot\Phi_1\left(|\boldsymbol{I}+c^2\boldsymbol{D}\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}^\top|^{-1/2}c\boldsymbol{x}_{ij}^\top\boldsymbol{\xi}\right), \tag{51}$$

with similar expressions for the variance and covariance terms. Upon estimating the parameters within the probit approximation paradigm, back-transformation to the original

logit scale is possible, using expressions such as (49) and (51). This opens perspectives for alternative estimation methods for the combined model with logit link, with the important special case of the normal-logistic GLMM.

In the Bernoulli case, calculating the moments is extremely simple. Indeed, the Bernoulli moments are all identical. The conditional moments are all $\mathrm{E}(Y_{ij}^k|\theta_{ij},\boldsymbol{b}_i) = \theta_{ij}\kappa_{ij}$ $(k = 1,2,\ldots)$. Hence, they all reduce to (38). In the probit case, they are equal to (45).

6.2.1. A clinical trial in onychomycosis

We present the MVDV analysis of the binary onychomycosis data, introduced in Section 2.2. For the logit, consider the model:

$$\begin{aligned} Y_{ij}|(b_i) &\sim \text{Bernoulli}(\pi_{ij}), \\ \text{logit}(\pi_{ij}) &= \xi_1(1-T_i) + b_i + \xi_2(1-T_i)t_{ij} + \xi_3 T_i + \xi_4 T_i t_{ij}, \end{aligned} \tag{52}$$

where T_i is the treatment indicator for subject i, t_{ij} is the time-point at which the jth measurement is taken for the ith subject, and $b_i \sim N(0,d)$. Parameter estimates for the logistic model, with and without the normal random effect on the one hand, and with and without the beta-binomial component on the other hand, as described in Section 6.1, are presented in Table 5. Observe that the model becomes hard to fit when the beta random

Table 5: *Onychomycosis study. Parameter estimates (standard errors) for the regression coefficients in (1) the logistic model, (2) the beta-binomial model, (3) the logistic-normal model, and (4) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

Effect	Par.	Logistic	Beta-binomial
Intercept treatment A	ξ_0	−0.5571 (0.1090)	17.9714 (1482.6)
Slope treatment A	ξ_1	−0.1769 (0.0246)	5.2454 (12970.0)
Intercept treatment B	ξ_2	−0.5335 (0.1122)	18.6744 (2077.13)
Slope treatment B	ξ_3	−0.2549 (0.0309)	4.7775 (12912.0)
Std. dev random effect	$\sqrt{d}$	—	—
Ratio	α/β	—	3.6739 (0.2051)
−2log-likelihood		1812	1980
Effect	**Par.**	**Logistic-normal**	**Combined**
Intercept treatment A	ξ_0	−1.6299 (0.4354)	−1.6042 (4.0263)
Slope treatment A	ξ_1	−0.4042 (0.0460)	−6.4783 (1.4386)
Intercept treatment B	ξ_2	−1.7486 (0.4478)	−16.2079 (3.5830)
Slope treatment B	ξ_3	−0.5634 (0.0602)	−8.0745 (1.5997)
Std. dev random effect	$\sqrt{d}$	4.0150 (0.3812)	60.8835 (14.2237)
Ratio	α/β	—	0.2805 (0.0350)
−2log-likelihood		1248	1240

effects are present, which is seen from estimates and standard errors in both the beta-binomial model as well as the combined model. To understand this, we must observe that the conjugate random effects in the Bernoulli case, unlike in the Poisson, binomial, and Weibull cases, cannot add to the variability, only to the correlation structure. This means that there is considerably less information available than in the other cases. This does not mean that the beta random effects are unnecessary, but rather that they challenge the stable estimation of other model parameters.

6.3. Models for binomial data with logit and probit link

Molenberghs et al. (2012) supplemented the study of the binary case with the binomial one. Starting from the Bernoulli expressions (35) and (36) but now for three rather than two levels, they got:

$$Y_{ijk} \sim \text{Bernoulli}(\pi_{ijk} = \theta_{ijk}\kappa_{ijk}), \tag{53}$$

$$\kappa_{ijk} = \frac{\exp\left(\boldsymbol{x}_{ijk}^{\intercal}\boldsymbol{\xi} + \boldsymbol{z}_{ijk}^{\intercal}\boldsymbol{b}_i\right)}{1+\exp\left(\boldsymbol{x}_{ijk}^{\intercal}\boldsymbol{\xi} + \boldsymbol{z}_{ijk}^{\intercal}\boldsymbol{b}_i\right)}, \tag{54}$$

where i stands for the independent block, as before, j for occasion, and k for the repeats of the Bernoulli trials. It is natural to define $Z_{ij} = \sum_{k=1}^{m_{ij}} Y_{ijk}$. Also here, there are no closed-form expressions for the moments when a logit link is used, but they do exist for the probit case. The data consists of an array of successes $\boldsymbol{z}_i = (z_{i1},\ldots,z_{in_i})^{\intercal}$ out of $\boldsymbol{m}_i = (m_{i1},\ldots,m_{in_i})^{\intercal}$ trials. It is also convenient to provide for multi-indices $\boldsymbol{t} = (t_1,\ldots,t_{n_i})^{\intercal}$ and for vectors of the parameters $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_{n_i})^{\intercal}$ and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_{n_i})^{\intercal}$. The joint distribution can then be written as:

$$f(\boldsymbol{z}_i|\boldsymbol{m}_i,\boldsymbol{\xi},\boldsymbol{D},\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{\boldsymbol{t}=\boldsymbol{0}}^{\boldsymbol{m}_i-\boldsymbol{z}_i}\left[\prod_{j=1}^{n_i}\frac{(-1)^{t_j}}{B(\alpha_j,\beta_j)}\binom{m_{ij}}{z_{ij}}\binom{m_{ij}-z_{ij}}{t_j}B(z_{ij}+\alpha_j+t_j,\beta_j)\right]\times$$
$$\times\,\Phi_{\sum_j t_j}\left[(X_i(t)\boldsymbol{\xi};\boldsymbol{L}(t)^{-1}\right]. \tag{55}$$

Here, $\boldsymbol{X}_i(t)$ is the design matrix, built from X_i, with row j in X_i replicated t_j times. The design matrix X_i is built similarly, and then, in analogy with (43),

$$\boldsymbol{L}(t) = \boldsymbol{I}_{\sum_j t_j} - \boldsymbol{Z}_i(t)\left[\boldsymbol{D}^{-1} + \boldsymbol{Z}_i(t)^{\intercal}\boldsymbol{Z}_i(t)\right]^{-1}\boldsymbol{Z}_i(t)^{\intercal}. \tag{56}$$

6.4. Iron-deficient diets in rats

We turn to the data in Section 2.3. Because the probability of a fetus dying varies from litter to litter, the total variance of the proportions will be greater than that predicted by a binomial model, even when covariates are accounted for. Hence, overdispersion and correlation need to be accommodated.

Construct predictor function $\eta_i = \xi_0 + \xi_2 x_{2i} + \xi_3 x_{3i} + \xi_4 x_{4i}$ with $x_{g_i} = 1$ if litter i belongs to group g and 0 otherwise. The placebo group figures as a reference category. Further, let $Z_i = \sum_{j=1}^{n_i} Y_{ij} \sim \text{Binomial}(n_i, \pi_i)$ be the number of dead fetuses out of n_i in litter i. Five models are considered: (a) the binomial model, $\text{logit}(\pi_i) = \eta_i$; (b) the GLMM: $\text{logit}(\pi_i) = \eta_i + b_i$, where $b_i \sim N(0,d)$; (c) the beta-binomial model, $\text{logit}(\mu_i) = \eta_i$, where $\pi_i \sim \text{Beta}(\alpha,\beta)$, and $\mu_i = \text{E}(\pi_i)$; (d) the beta-binomial model with normal random effects: for $b_i \sim N(0,d)$, $\text{logit}(\mu_i) = \eta_i$, and π_i and μ_i as in the beta-binomial; (e) in the combined model: $\text{logit}(\kappa_i) = \eta_i + b_i$ where $\pi_i = \theta_i \kappa_i$, $\theta_i \sim \text{Beta}(\alpha,\beta)$, and $b_i \sim N(0,d)$. The constraint $\alpha\beta \equiv 1$ is imposed in the latter case.

The results of the various models are presented in Table 6. We observe that the two models that simultaneously account for overdispersion and correlation perform better than the others. The classical beta-binomial model with normal random effects has the same double negative log-likelihood as the combined model. This is the case only for cross-sectional data; even though their hierarchical formulations are different, they marginally coincide in this case. That said, the parameters have a different meaning, as they are to be interpreted conditionally on the assumed random-effects structure. Differences may be very noticeable when binomial measurements are collected repeatedly over time or in an otherwise hierarchical fashion.

Between these two, the estimates' precision is best in the combined model. Owing to conjugacy, the mean model and overdispersion parameter estimators are less correlated, leading to increased precision, even though the effect is modest.

Table 6: *Iron-deficiency study. Parameter estimates (standard errors) for (1) the binomial model, (2) the GLMM, (3) the beta-binomial model, (4) the conventional beta-binomial model with random effect in the linear predictor, and (5) the combined model.*

Effect	Par.	Binomial	GLMM	BB	BB-normal	Combined
Intercept	ξ_0	1.14(0.13)	1.80(0.36)	1.35(0.25)	1.79(0.38)	1.80(0.36)
Group2	ξ_2	−3.32(0.33)	−4.52(0.74)	−3.11(0.50)	−4.49(0.80)	−4.51(0.74)
Group3	ξ_3	−4.48(0.73)	−5.86(1.19)	−3.87(0.81)	−5.81(1.30)	−5.85(1.19)
Group4	ξ_4	−4.13(0.48)	−5.60(0.92)	−3.93(0.67)	−5.57(0.97)	−5.59(0.92)
Std. dev. RE	$\sqrt{d}$	—	1.54(0.29)	—	1.52(0.37)	1.53(0.29)
Overdispersion		—	—	0.24(0.06)	0.005(0.051)	0.0005(0.0018)
−2log-likelihood		244.9	183.9	186.9	183.8	183.8

6.5. Ordinal data: a combined proportional odds-beta-normal model

The ordinal case was studied by Ivanova et al. (2014). Assume the ordinal outcome Y_{ij} can take values $r = 1, \ldots, R$, and replace it by a set of R dummies:

$$Z_{r,ij} = \begin{cases} 1 & \text{if } Y_{r,ij} = r, \\ 0 & \text{otherwise,} \end{cases}$$

for $r = 1, \ldots, R$. Evidently, there are redundant dummies, but any subset of $R-1$ components is not. Group the dummies into vectors $\boldsymbol{Z}_{ij}$ and $\boldsymbol{Z}_i$ for a specific subject i and occasion j, and for a specific subject i, respectively. We assume a multinomial distribution $\boldsymbol{Z}_{ij} \sim \text{multinomial}(\boldsymbol{\pi}_{ij})$, with $\boldsymbol{\pi}_{ij} = (\pi_{1,ij}, \ldots, \pi_{r,ij}, \ldots, \pi_{R,ij})$. The multinomial distribution at a given occasion is determined by the modelling choice for the ordinal outcome. Under a proportional odds assumption, using normal random effects $\boldsymbol{b}_i \sim N(0, D)$ in the linear predictor, and beta random effects $\theta_{ij} \sim \text{Beta}(\alpha_j, \beta_j)$ to capture further overdispersion, the probabilities can be written as:

$$\pi_{r,ij} = \begin{cases} \theta_{ij}\kappa_{1,ij} & \text{if } r = 1, \\ \theta_{ij}(\kappa_{r,ij} - \kappa_{r-1,ij}) & \text{if } 1 < r < R, \\ 1 - \theta_{ij}\kappa_{R-1,ij} & \text{if } r = R. \end{cases} \tag{57}$$

where

$$\kappa_{r,ij} = \frac{\exp\left(\xi_{0r} + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i\right)}{1 + \exp\left(\xi_{0r} + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i\right)}. \tag{58}$$

Here, $\xi_{01} \leq \cdots \leq \xi_{0,R-1}$ are intercepts, $\boldsymbol{\xi}$ are fixed regression coefficients, and $\boldsymbol{x}_{ij}$ ($\boldsymbol{z}_{ij}$) is the design vector for the fixed (random) effects at occasion j. Also here, some choices in the above can be relaxed and/or altered. For example, like before, the α_j and β_j parameters, describing the beta distribution, need not be dependent on j. To ensure identifiability, a constraint needs to be applied to it, e.g., $\alpha_j\beta_j = 1$, but it is mathematically convenient to retain them as two separate parameters, with the understanding that the constraint does apply. Finally, the θ_{ij} within a subject are assumed different from each other and independent. One could allow them to be correlated, or even constant across subjects. This will not be considered here.

As argued in MVDV, MVID, and Molenberghs et al. (2012), closed-form expressions for marginal means, variances, covariances, and even the entire marginal distribution, i.e., integrated over both sets of random effects, cannot be derived in the binary case with logit link and normal random effects (regardless of the overdispersion random effects). Evidently, the same will be true for the ordinal case. If necessary, numerical integration or other Monte Carlo methods can be used to derive such marginal quantities.

6.6. Diabetes study

We describe the analysis of the diabetes study (Section 2.4), reported in Ivanova et al. (2014). Let $Y_{ij} = 0, \ldots, 3$ be the number of clinical targets patient i reached at occasion j. Also, let $t_{ij} = 0, 1$ be the time point at which the jth measurement was taken. Consider the combined proportional odds logistic regression model:

$$\text{logit}[P(Y_{ij} \leq r | t_{ij}, X_i)] = \xi_{0r} + b_i + \xi_1 t_{ij} + \xi_2 X_i,$$

$(r = 0, \ldots, 3)$, where the random intercept b_i is assumed $N(0, d)$ distributed, and X_i is an indicator for group. The beta random effect is re-parameterized such that

$$\nu = \frac{e^{\delta}}{1 + e^{\delta}} = \frac{\alpha}{\alpha + \beta},$$

thus simultaneously avoiding identifiability and range violation issues. The parameter δ is the one entered into the likelihood function. We consider (1) the ordinary proportional odds model, (2) the proportional odds model with beta overdispersion effect, (3) the proportional odds model with random normal effect, and (4) the combined model. Estimates (standard errors) are presented in Table 7. Clearly, there is no significant im-

Table 7: *Diabetes study. Parameter estimates (standard errors) from the regression coefficients in (1) the ordinary proportional odds model, (2) the proportional odds model with beta overdispersion effect, (3) the proportional odds model with random normal effect, together with (4) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

Effect	Par.	PO	PO-Beta
Intercept 0	ξ_{00}	−0.7130 (0.0662)	−1.7129 (0.0662)
Intercept 1	ξ_{01}	0.2668 (0.0560)	0.2667 (0.0560)
Intercept 2	ξ_{02}	2.0279 (0.0648)	2.0277 (0.0650)
Slope time	ξ_1	−0.7614 (0.0575)	−0.7610 (0.0575)
Slope group	ξ_2	−0.2053 (0.0587)	−0.2053 (0.0587)
Std. dev. RE	$\sqrt{d}$	—	—
Beta parameter	δ	—	13.1622 (390.44)
−2 log-likelihood		10588.18	10588.18
Effect	Par.	PO-Normal	PO-Beta-Normal
Intercept 0	ξ_{00}	−2.3201 (0.0100)	−2.3201 (0.0999)
Intercept 1	ξ_{01}	0.3336 (0.0818)	0.3335 (0.0818)
Intercept 2	ξ_{02}	2.7727 (0.1035)	2.7728 (0.1035)
Slope time	ξ_1	−1.0268 (0.0659)	−1.0268 (0.0659)
Slope group	ξ_2	−0.2605 (0.0912)	−0.2605 (0.0912)
Std. dev. RE	$\sqrt{d}$	1.5105 (0.0729)	1.5205 (0.0729)
Beta parameter	δ	—	15.4925 (246.55)
−2 log-likelihood		10320.39	10320.39

provement, neither when we switch from model (1) to model (2), nor when we move from (3) to (4). The estimate for the beta-parameter δ is large and has a very large standard error. This indicates that there is probably no overdispersion in the data.

7. Time-to-event data

MVDV, using their general framework, also focused on time-to-event data, combining the Weibull model with normal and gamma random effects. The model extends both the GLMM and the gamma frailty model. Molenberghs et al. (2015) extended the approach to allow for censoring. In what follows, we will give an overview of these developments. Efendi and Molenberghs (2013) paid particular attention to various estimation strategies. Abrams et al. (2017) integrated this framework in the modelling of current-status data, in the context of infectious diseases modelling.

Molenberghs and Verbeke (2011a), using closed-form expressions for the model's moments, pointed to both probabilistic as well as data-analytic implications of using (gamma) frailty models. We give a brief summary of these in Section 7.2.

The general Weibull model for repeated measures, with both gamma and normal random effects can be expressed as

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}_i,\boldsymbol{b}_i)=\prod_{j=1}^{n_i}\lambda\rho\theta_{ij}y_{ij}^{\rho-1}e^{\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}+\boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i}e^{-\lambda y_{ij}^{\rho}\theta_{ij}e^{\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}+\boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i}}, \tag{59}$$

$$f(\boldsymbol{\theta}_i)=\prod_{j=1}^{n_i}\frac{1}{\beta_j^{\alpha_j}\Gamma(\alpha_j)}\theta_{ij}^{\alpha_j-1}e^{-\theta_{ij}/\beta_j}, \tag{60}$$

$$f(\boldsymbol{b}_i)=\frac{1}{(2\pi)^{q/2}|\boldsymbol{D}|^{1/2}}e^{-\frac{1}{2}\boldsymbol{b}_i^{\top}\boldsymbol{D}^{-1}\boldsymbol{b}_i}. \tag{61}$$

A few observations are in place. First, setting $\rho=1$ leads to the special case of an exponential time-to-event distribution. Second, the classical gamma frailty model (i.e., no normal random effects) and the Weibull-based GLMM (i.e., no gamma random effects) follow as special cases. Third, strong conjugacy applies. This is definitely true for the exponential model, but carries over to the Weibull model, using the transformation Y_{ij}^{ρ}. It is equally possible to derive this result by merely re-writing the factor $\phi=\lambda\kappa$. Fourth, the above expressions are derived for a two-parameter gamma density. It is customary in a gamma frailty context (Duchateau and Janssen, 2007) to set $\alpha_j\beta_j=1$, for reasons of identifiability. In this case, (60) is replaced by

$$f(\boldsymbol{\theta}_i)=\prod_{j=1}^{n_i}\frac{1}{\left(\frac{1}{\alpha_j}\right)^{\alpha_j}\Gamma(\alpha_j)}\theta_{ij}^{\alpha_j-1}e^{-\alpha_j\theta_{ij}}, \tag{62}$$

Alternatively, assuming $\alpha_j = 1$ and $\beta_j = 1/\delta_j$, one could write

$$f(\boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \delta_j e^{-\delta_j \theta_{ij}}, \tag{63}$$

implying that the gamma density is reduced to an exponential one.

MVDV derived a multi-index series formulation of the marginal joint distribution:

$$f(\boldsymbol{y}_i) = \sum_{(m_1,\ldots,m_{n_i})} \prod_{j=1}^{n_i} \frac{(-1)^{m_j}}{m_j!} \frac{\Gamma(\alpha_j + m_j + 1)\beta_j^{m_j+1}}{\Gamma(\alpha_j)} \lambda^{m_j+1} \rho y_{ij}^{(m_j+1)\rho-1}$$
$$\times \exp\left\{(m_j+1)\left[\boldsymbol{x}_{ij}^\top \boldsymbol{\xi} + \tfrac{1}{2}(m_j+1)\cdot \boldsymbol{z}_{ij}^\top \boldsymbol{D}\boldsymbol{z}_{ij}\right]\right\}. \tag{64}$$

In case censorship applies, it is easy to integrate (64) over the interval $[C_{ij}, +\infty[$ or, in a multivariate fashion, over the cube $[\boldsymbol{0}, \boldsymbol{C}_i]$:

$$F(\boldsymbol{C}_i) = \sum_{(m_1,\ldots,m_{n_i})} \prod_{j=1}^{n_i} \frac{(-1)^{m_j}}{(m_j+1)!} \frac{\Gamma(\alpha_j + m_j + 1)\beta_j^{m_j+1}}{\Gamma(\alpha_j)} \lambda^{m_j+1} C_{ij}^{(m_j+1)\rho}$$
$$\times \exp\left\{(m_j+1)\left[\boldsymbol{x}_{ij}^\top \boldsymbol{\xi} + \tfrac{1}{2}(m_j+1)\cdot \boldsymbol{z}_{ij}^\top \boldsymbol{D}\boldsymbol{z}_{ij}\right]\right\}. \tag{65}$$

Evidently, if censorship applies to some but not all of the times within the vector, then the integration can be restricted to these, and the corresponding contribution will be an amalgamation of components taken from (64) and (65).

MVDV also derived the following moment expression, with mean, variance, and covariance expressions:

$$\mathrm{E}(Y_{ij}^k) = \frac{\alpha_j B(\alpha_j - k/\rho, k/\rho + 1)}{\lambda^{k/\rho}\beta_j^{k/\rho}} \exp\left(-\frac{k}{\rho}\boldsymbol{x}_{ij}^\top \boldsymbol{\xi} + \frac{k^2}{2\rho^2}\boldsymbol{z}_{ij}^\top \boldsymbol{D}\boldsymbol{z}_{ij}\right), \tag{66}$$

$$\mathrm{E}(Y_{ij}) = \frac{\alpha_j B(\alpha_j - 1/\rho, 1/\rho + 1)}{\lambda^{1/\rho}\beta_j^{1/\rho}} \exp\left(-\frac{1}{\rho}\boldsymbol{x}_{ij}^\top \boldsymbol{\xi} + \frac{1}{2\rho^2}\boldsymbol{z}_{ij}^\top \boldsymbol{D}\boldsymbol{z}_{ij}\right), \tag{67}$$

$$\mathrm{Var}(Y_{ij}) = \frac{\alpha_j}{\lambda^{2/\rho}\beta_j^{2\rho}} \exp\left(-\frac{2}{\rho}\boldsymbol{x}_{ij}^\top \boldsymbol{\xi} + \frac{1}{\rho^2}\boldsymbol{z}_{ij}^\top \boldsymbol{D}\boldsymbol{z}_{ij}\right)$$
$$\times \left[B(\alpha_j - 2/\rho, 2/\rho + 1)\exp\left(\frac{1}{\rho^2}\boldsymbol{z}_{ij}^\top \boldsymbol{D}\boldsymbol{z}_{ij}\right) - \alpha_j B\left(\alpha_j - \frac{1}{\rho}, \frac{1}{\rho} + 1\right)^2\right], \tag{68}$$

$$\begin{aligned}\text{Cov}(Y_{ij},Y_{ik}) = &\frac{\alpha_j\alpha_k}{\lambda^{2/\rho}\beta_j^{1/\rho}\beta_k^{1/\rho}}\exp\left[-\frac{1}{\rho}(\boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\xi}+\boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\xi})\right]\\ &\times B\left(\alpha_j-\frac{1}{\rho},\frac{1}{\rho}+1\right)B\left(\alpha_k-\frac{1}{\rho},\frac{1}{\rho}+1\right)\\ &\times\exp\left[\frac{1}{2\rho^2}(\boldsymbol{z}_{ij}^{\intercal}\boldsymbol{D}\boldsymbol{z}_{ij}+\boldsymbol{z}_{ik}^{\intercal}\boldsymbol{D}\boldsymbol{z}_{ik})\right]\left[\exp\left(\frac{1}{\rho^2}\boldsymbol{z}_{ij}^{\intercal}\boldsymbol{D}\boldsymbol{z}_{ik}\right)-1\right]. \qquad (69)\end{aligned}$$

7.1. Recurrent asthma attacks in children

MVDV analysed the times-to-event, introduced in Section 2.5. They considered an exponential model, i.e., a model of the form (59) with $\rho = 1$, and further a predictor of the form:

$$\kappa_{ij} = \xi_0 + b_i + \xi_1 T_i,$$

where T_i is an indicator for treatment and $b_i \sim N(0,d)$. Results from fitting all four models (with/without normal random effect; with/without gamma random effect) can be found in Table 8. A formal assessment of the treatment effect from all four models is given in Table 9. The treatment effect ξ_1 is stably identifiable in all four models. As can be seen from Table 9, the treatment effects are similar in strengths, but including both random effects reduces the evidence, relative to the exponential model. Needless to say that too parsimonious an association structure might lead to liberal test behaviour.

Table 8: *Asthma study. Parameter estimates (standard errors) for the regression coefficients in (1) the exponential model, (2) the exponential-gamma model, (3) the exponential-normal model, and (4) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

Effect	Par.	Exponential	Exponential-gamma
Intercept	ξ_0	−3.3709(0.0772)	−3.9782(15.354)
Treatment effect	ξ_1	−0.0726(0.0475)	−0.0755(0.0605)
Shape parameter	λ	0.8140(0.0149)	1.0490(16.106)
Std. dev. random effect	$\sqrt{d}$	—	—
Gamma parameter	γ	—	3.3192(0.3885)
−2log-likelihood		18,693	18,715
Effect	Par.	Exponential-normal	Combined
Intercept	ξ_0	−3.8095(0.1028)	3.9923(20.337)
Treatment effect	ξ_1	−0.0825(0.0731)	−0.0887(0.0842)
Shape parameter	λ	0.8882(0.0180)	0.8130(16.535)
Std. dev. random effect	$\sqrt{d}$	0.4097(0.0386)	0.4720(0.0416)
Gamma parameter	γ	—	6.8414(1.7146)
−2log-likelihood		18,611	18,629

Table 9: *Asthma study. Wald test results for the assessment of treatment effect.*

Model	Z value	p-value
Exponential	−1.5283	0.1264
Exponential-gamma	−1.1293	0.2588
Exponential-normal	−1.2480	0.2120
Combined	−1.0534	0.2921

7.2. Probabilistic and data-analytic issues with frailty models and their combined-model extensions

Based on moment expression (66), Molenberghs and Verbeke (2011a) observed that there can be a problem with models combining Weibull outcomes with gamma random effects, as well as with several extensions and sub-models. In particular, they established a connection with the so-called log-logistic distribution (Shoukri, Mian and Tracy, 1988), a transformation of the logistic distribution to the half line with only a finite number of finite moments.

To make their point, they started from a univariate Weibull distribution with gamma random effects (adding the normal random effects to the linear predictor does not substantially change anything), for which all expressions are given in the last column of Table 1. Like before, setting $\alpha\beta = 1$, and using formulation (62), the gamma and marginal distributions are written as:

$$f(\theta) = \frac{1}{\left(\frac{1}{\alpha}\right)^{\alpha}\Gamma(\alpha)}\theta^{\alpha-1}e^{-\alpha\theta}, \tag{70}$$

$$f(y) = \frac{\varphi\rho y^{\rho-1}\alpha^{\alpha+1}}{(\alpha+\varphi y^{\rho})^{\alpha+1}}. \tag{71}$$

Molenberghs and Verbeke (2011a) term this Case I. They also considered Case II, obtained by setting $\alpha = 1$ and $\beta = 1/\delta$, line in (63), henceforth, Case II:

$$f(\theta) = \delta e^{-\delta\theta}, \tag{72}$$

$$f(y) = \frac{\varphi\rho y^{\rho-1}\delta}{(\delta+\varphi y^{\rho})^2}. \tag{73}$$

Here, the gamma distribution has been replaced by its exponential special case and (73) is the log-logistic distribution (Bennett, 1983; Collett, 2003).

The moments follow from (66). For the general case, with α and β free parameters, for Case I, and for Case II, they are, respectively:

$$\text{General}: \mathrm{E}(Y^k) = \frac{\alpha B(\alpha - k/\rho, k/\rho + 1)}{(\beta\varphi)^{k/\rho}}, \tag{74}$$

$$\text{Case I}: \mathrm{E}(Y^k) = \left(\frac{\alpha}{\varphi}\right)^{k/\rho} \frac{k}{\rho} B(\alpha - k/\rho, k/\rho), \tag{75}$$

$$\text{Case II (log-logistic)}: \mathrm{E}(Y^k) = \frac{k}{\rho}\left(\frac{\delta}{\varphi}\right)^{k\rho} \cdot \Gamma(1 - k/\rho) \cdot \Gamma(k/\rho). \tag{76}$$

The moments (74) are finite if and only if $k < \alpha\rho$. Hence, if $\alpha\rho$ is small, there is a risk that even lower-order moments do not exist, which evidently is problematic. Molenberghs and Verbeke (2011a) gave an example, using data from Duchateau and Janssen (2007). For certain methods of estimation in the context of the Weibull-Gamma frailty model, this would imply that regularity conditions are not satisfied. For the log-logistic case, this becomes $k < \rho$. The moments have been presented by Rinne (2009, p. 157) as well, though without reference to the irregularity issue.

8. Estimation

MVD and MVDV showed that fitting the combined model is relatively easy, and that standard software tools, such as the SAS procedure NLMIXED, can be used for maximum likelihood estimation in this case. More generically, any sufficiently flexible likelihood maximization tool that allows for normally distributed random effects can be used to this effect. This can typically be done with relatively little programming effort. Efendi and Molenberghs (2013) expanded upon this for the specific case of time-to-event data, and supplemented maximum likelihood with pairwise likelihood and Bayesian estimation. Their simulations indicated that, while maximum likelihood can be faster than pairwise likelihood, the latter has somewhat better convergence properties.

A priori, fitting a combined model of the type described in Section 4, proceeds by integrating over the random effects. The likelihood contribution of subject i is

$$f_i(\boldsymbol{y}_i|\boldsymbol{\vartheta}, \boldsymbol{D}, \boldsymbol{\vartheta}_i, \boldsymbol{\Sigma}_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\vartheta}, \boldsymbol{b}_i, \boldsymbol{\theta}_i)\, f(\boldsymbol{b}_i|\boldsymbol{D})\, f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \boldsymbol{\Sigma}_i)\, d\boldsymbol{b}_i\, d\boldsymbol{\theta}_i. \tag{77}$$

Here, $\boldsymbol{\vartheta}$ groups all parameters in the conditional model for Y_i. From (77) the likelihood derives as:

$$\begin{aligned} \boldsymbol{L}(\boldsymbol{\vartheta}, \boldsymbol{D}, \boldsymbol{\vartheta}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{N} f_i(\boldsymbol{y}_i|\boldsymbol{\vartheta}, \boldsymbol{D}, \boldsymbol{\vartheta}_i, \boldsymbol{\Sigma}_i) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\vartheta}, \boldsymbol{b}_i, \boldsymbol{\theta}_i)\, f(\boldsymbol{b}_i|\boldsymbol{D})\, f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \boldsymbol{\Sigma}_i)\, d\boldsymbol{b}_i\, d\boldsymbol{\theta}_i. \end{aligned} \tag{78}$$

The key problem in maximizing (78) is the presence of N integrals over the random effects $\boldsymbol{b}_i$ and $\boldsymbol{\theta}$. It is widely claimed that the absence of a closed-form solution precludes an analytical-integration based solution (Molenberghs and Verbeke, 2005), explaining the popularity of Taylor-series expansion based methods, such as PQL and MQL, Laplace approximation, and numerical-integration based methods. These have been implemented in, for example, the SAS procedures GLIMMIX and NLMIXED. Several of the series expansion methods tend to exhibit bias, an issue taken up in Breslow and Lin (1995), and suggesting the use of alternative methods.

However, thanks to our results in Section 4, further progress can be made. Closed-form integration, apart from the normal case, is within reach for the Poisson, probit, and Weibull cases. Now, some closed forms involve series expansions, and may be either time consuming or cumbersome to implement. This notwithstanding, a variety of alternative approaches are possible.

Let us turn to the Poisson case. While closed-form expressions can be used to implement maximum likelihood estimation, with numerical accuracy governed by the number of terms included in the series, one can also proceed by what we will term partial marginalization. By this we refer to integrating (18)–(22) over the gamma random effects only, leaving the normal random effects untouched. The corresponding probability is:

$$f(y_{ij}|\boldsymbol{b}_i) = \left(\begin{array}{c} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{array} \right) \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j} \right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j} \right)^{\alpha_j} \kappa_{ij}^{y_{ij}}, \tag{79}$$

where $\kappa_{ij} = \exp[\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i]$. Note that, with this approach, we assume that the gamma random effects are independent within a subject. This is fine, given the correlation is induced by the normal random effects.

Similarly, for the Weibull case we obtain

$$f(y_{ij}|\boldsymbol{b}_i) = \frac{\lambda\kappa_{ij}e^{\mu_{ij}}\rho y_{ij}^{\rho-1}\alpha_j\beta_j}{(1+\lambda\kappa_{ij}e^{\mu_{ij}}\beta_j y_{ij}^{\rho})^{\alpha_j+1}}. \tag{80}$$

Now, in the survival case it is evidently very likely that censoring occurs. Focusing on right-censored data, it is then necessary to integrate the marginal density over the survival time within the interval $[0, C_i]$. The corresponding cumulative distribution is given in (65). In the spirit of (80), the partial marginalization of a censored component takes the form:

$$f(C_{ij}|\boldsymbol{b}_i) = \int_{C_{ij}}^{+\infty} f(y_{ij}|\boldsymbol{b}_i)dy_{ij} = \frac{1}{(1+\lambda\kappa_{ij}e^{\mu_{ij}}C_{ij}^{\rho})^{\alpha_j}}. \tag{81}$$

The concept of partial integration always applies whenever strong conjugacy holds. Indeed, an expression of the form (16) corresponds to integrating over the conjugate ran-

dom effect θ, while leaving the normally distributed random effect embedded in the predictor, κ in this notation. Recall that, while expressions of the type (16) appear to be for the univariate case, they extend without problem to the longitudinal setting as well.

Because there is lack of strong conjugacy, the logit case defies the mere exploitation of conjugacy, such as the negative binomial form (79) and the Weibull-gamma frailty form (80). Nevertheless, it is easy to derive, for this case:

$$f(y_{ij}|\boldsymbol{b}_i) = \frac{1}{\alpha_j+\beta_j} \cdot (\kappa_{ij}\alpha_j)^{y_{ij}} \cdot [(1-\kappa_{ij})\alpha_j+\beta_j]^{1-y_{ij}}. \tag{82}$$

For all of these, it is straightforward to obtain the fully marginalized probability by numerically integrating the normal random effects out of (79), (80), and (82), using a tool such as the SAS procedure NLMIXED that allows for normal random effects in arbitrary, user-specified models.

For the specific case of the marginalized probit model, the computational challenge stems from the presence of a multivariate normal integral of the form (42), a phenomenon also known from the fully marginally specified multivariate probit model (Ashford and Sowden, 1970; Lesaffre and Molenberghs, 1991; Molenberghs and Verbeke, 2005). Specific to the context of the probit models with random effects, Zeger et al. (1988) derived the marginal mean function, needed for their application of generalized estimating equations as a fitting algorithm for the marginalized probit model.

In the ordinal case, the partially marginalized density at occasion *j* for subject *i* takes the form:

$$f(y_{ij}|\boldsymbol{b}_i) = \frac{\alpha_j}{\alpha_j+\beta_j} \cdot (\kappa_{1,ij})^{z_{1,ij}} \cdot \prod_{r=2}^{R-1} (\kappa_{r,ij}-\kappa_{r-1,ij})^{z_{r,ij}} \cdot \left(\frac{\alpha_j+\beta_j}{\alpha_j} - \kappa_{R-1,ij}\right)^{z_{R,ij}}.$$

From these, the likelihood can be constructed by assembling all contributions over subjects and repeated measurements within subjects.

MVDV discussed a number of alternative estimation strategies. These include pseudo-likelihood (or: pairwise likelihood; Aerts et al., 2002; Molenberghs and Verbeke, 2005), Bayesian inferences, non-parametric maximum likelihood (Booth et al., 2003: Aitkin, 1999; Alfò and Aitkin, 2000). Also, hierarchical generalized linear models (Lee and Nelder, 1996; Lee et al., 2006) can be used. They also referred to transformation-based methods, whereby non-normal random effects are transformed to normal ones, or vice versa (Liu and Yu, 2008; Nelson et al., 2006).

An important point is that not all parameters may be simultaneously identifiable. For example, the gamma-distribution parameters in the Poisson case, α and β, are not simultaneously identifiable when the linear-predictor part is also present, because there is aliasing with the intercept term. Therefore, one can set, for example, β equal to a constant, removing the identifiability problem. It is then clear that α, in the univariate case, or the set of α_j in the repeated-measures case, describe the additional overdispersion, in

addition to what stems from the normal random effect(s). A similar phenomenon also plays in the binary case, where both beta-distribution parameters are not simultaneously estimable.

In addition, also Bayesian estimation and inference can be considered. Ghebretinsae et al. (2013) considered a Bayesian version in the time-to-event case. Ghebretinsae et al. (2012) presented a Bayesian joint CM. Efendi and Molenberghs (2013) juxtaposed likelihood-based and Bayesian estimation. The performance of the Bayesian method for the count case was assessed, using simulations, by Aregay, Shkedy, and Molenberghs (2013) and Rizzato et al. (2016). Aregay, Shkedy, and Molenberghs (2015) compared model versions with additive and multiplicative random effects. On a related note, Iddi et al. (2014) examined empirical Bayes estimation for the combined model.

9. Implication for computation of correlation and derived quantities

As we have seen, the combined model allows for closed-form expressions for moments, and hence for means and variances, for the normal, Poisson, probit, and Weibull cases, with a combination of normal random effects on the one hand, supplemented on the other hand with conjugate random effects, taking a normal, gamma, beta, and gamma form, respectively. The obvious one missing from the list is the logit model, but then the logit-probit connection, as discussed in Section 6.2, comes to the rescue.

These closed-form moments enable easy calculations of such derived quantities as correlations. For the count case, this was done by Vangeneugden et al. (2011), while Vangeneugden et al. (2014) focused on the binary setting.

For the count combined model, Vangeneugden et al. (2011) used the following derivation. The mean vector $\boldsymbol{\mu}_i = \mathrm{E}(\boldsymbol{Y}_i)$ has components:

$$\mu_{ij} = \phi_{ij} \exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \tfrac{1}{2}\boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ij}\right), \tag{83}$$

and the variance-covariance matrix is given by

$$\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{M}_i + \boldsymbol{M}_i(\boldsymbol{P}_i - \boldsymbol{J}_{n_i})\boldsymbol{M}_i, \tag{84}$$

where $\boldsymbol{\phi}_i$ is the mean vector of the overdispersion random effects, with components ϕ_{ij}, $\boldsymbol{\Sigma}_i$ is the variance-covariance matrix of the overdispersion random effects, with components σ_{ij}, and $\boldsymbol{M}_i$ is a diagonal matrix with elements μ_{ij}. Further, the $(j,k)^{\text{th}}$ element of $\boldsymbol{P}_i$ equals

$$p_{i,jk} = \exp\left(\tfrac{1}{2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ik}\right) \cdot \frac{\sigma_{i,jk} + \phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}} \cdot \exp\left(\tfrac{1}{2}\boldsymbol{z}_{ik}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right). \tag{85}$$

Evidently, from this variance-covariance structure, the correlations immediately follow.

For the binary combined model, with probit link, the means, variances, and covariances were given in (45)–(47). When the logit link is used, no similar closed form exist. One can proceed by approximating the logit function via the probit function, or by using Taylor-series-based expressions. Details on these can be found in Vangeneugden et al. (2014).

The availability of closed-form correlation and other moment-based functions is useful in a number of contexts. For example, when studying psychometric reliability and generalizability (Vangeneugden et al., 2008; 2010), the correlation function is the basic building block. Correlation functions are also used in the context of surrogate marker evaluation from clinical-trial data (Alonso et al., 2017). Milanzi et al. (2015) used developments of this type to underscore the difference between manifest and latent correlations, for example when reliability measures are calculated in item response theory.

10. Marginalized versions of the combined model

As is clear from Sections 4–7, for many though not all versions of the CM there are explicit moment expressions and quantities derived there from. Nevertheless, they are algebraically involved, chiefly due to the non-conjugate nature of the normal random effects. To simplify the derivation of marginal quantities, such as effect measures, mean functions, etc., it is sensible to turn to the methodology of Heagerty (1999) and Heagerty and Zeger (2000), who modified the GLMM so that the first-order moments, i.e., the mean functions, are directly marginally interpretable. They originally focused on the logistic-normal model for binary longitudinal data, but they and others then extended the framework to other data types and link functions. The method specifies, at first sight contrary to intuition, a separate model for the marginal and conditional means. But this works thanks to a connector function that depends on covariates, marginal parameters, and the random-effects specification. Hence, both a marginal and conditional interpretation of the parameters can be maintained. The model, called the marginalized multilevel model (MMM), also allows for the use of maximum likelihood and Bayesian inferences, which is useful when data are incomplete.

To bring together the flexibility of the CM and the marginal interpretability of the MMM, Iddi and Molenberghs (2012ab) developed the *combined overdispersed and marginalized multilevel model* (COMMM). They focused on binary data and to some extent on counts. Kassahun et al. (2014b) studied further the count data case. The time-to-event case was studied by Efendi, Molenberghs, and Iddi (2014). Molenberghs et al. (2013) and Kenward and Molenberghs (2016) established connections between various ways of deriving marginally interpretable random-effects models, of which the MMM

idea is one. Iddi and Molenberghs (2013) and Kassahun et al. (2014b) combined the MMM idea, for counts, with the occurrence of zero-inflation.

The rest of this section is organized in the following way. In Section 10.1, the general MMM and COMMM methodology is given. The analysis of the epilepsy, onychomycosis, and asthma cases studies is presented in Sections 10.2–10.4. In Section 10.5, we show how further zero inflation in the count case can be added.

10.1. Methodology

The general formulation of the CM was given in Section 4. The other building block that we need is the general marginalized multilevel model (MMM), after which both will be merged.

The general marginalized multilevel model due to Heagerty (1999) can be written as:

$$g_1(\mu_{ij}^m) = \boldsymbol{x}_{ij}^\top \boldsymbol{\xi}^m, \tag{86}$$

$$g_2(\mu_{ij}^c) = \Delta_{ij} + \boldsymbol{z}_{ij}^\top \boldsymbol{b_i}, \tag{87}$$

$$\boldsymbol{b_i} \sim F_b(\boldsymbol{0}, \boldsymbol{D}), \tag{88}$$

$$Y_{ij}^c = Y_{ij} | \boldsymbol{b_i} \sim F_{Y^c}\left(\mu_{ij}^c, \upsilon\right). \tag{89}$$

The two link functions g_1 and g_2 can be different, although frequently they will be identical and then denoted by g. Further F_b is an arbitrary distribution. Here, υ is a dispersion parameter, similar to the overdispersion parameter ϕ in the exponential family. The marginal mean $\mu_{ij}^m = \mathrm{E}(Y_{ij})$ is made to depend on an $n_i \times p$ matrix of p linear predictors $\boldsymbol{X}_i$ through a link function $g(\cdot)$. Further, the conditional mean $\mu_{ij}^c = \mathrm{E}(Y_{ij}|\boldsymbol{b_i})$ relates to the random variable $\boldsymbol{b_i}$ with distribution (88) and the function Δ_{ij} connects the marginal and conditional means through the same link function; the latter aspect could be relaxed if desired. The conditional response distribution is given by F_{Y^c}. The function Δ_{ij} is obtained from the solution to the integral equation

$$\mu_{ij}^m = g^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\xi}^m) = \int_b g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^\top \boldsymbol{b_i}) dF_b. \tag{90}$$

For example, when the link function is logit and the distribution of the random effect is normal, the expression of Δ_{ij} is obtained from:

$$\mathrm{expit}(\boldsymbol{x}_{ij}^\top \boldsymbol{\xi}^m) = \int_b \mathrm{expit}(\Delta_{ij} + \boldsymbol{z}_{ij}^\top \boldsymbol{b_i}) \varphi(\boldsymbol{b_i}|\boldsymbol{0}, \boldsymbol{D}) d\boldsymbol{b_i}.$$

Here, $\mathrm{expit}(\eta) = e^\eta/(1+e^\eta)$. Griswold and Zeger (2004) expanded the model by relaxing the common link function assumed for both the marginal and conditional model specification. For example, using a logistic-probit-normal model:

$$\begin{aligned}
\text{logit}(\mu_{ij}^m) &= \boldsymbol{x}_{ij}^\top \boldsymbol{\xi}^m, \\
\Phi^{-1}(\mu_{ij}^c) &= \Delta_{ij} + \boldsymbol{z}_{ij}^\top \boldsymbol{b_i}, \\
\boldsymbol{b_i} &\sim F_b\left(\boldsymbol{0}, \boldsymbol{D}\right), \\
Y_{ij}^c | \boldsymbol{b_i} = Y_{ij} &\sim F_{Y^c}\left(\mu_{ij}^c, \upsilon\right).
\end{aligned}$$

(90) becomes:

$$\Delta_{ij} = \left(\sqrt{1 + \boldsymbol{z}_{ij}^\top \boldsymbol{D} \boldsymbol{z}_{ij}}\right) \cdot \Phi^{-1}\{\text{expit}(\boldsymbol{x}_{ij}^\top \boldsymbol{\xi}^m)\}. \tag{91}$$

The logit-probit-normal is more attractive than the logit-logit normal version in the sense that, for example, the marginal parameters will enjoy the odds ratio interpretation while at the same time retaining the computational advantage associated with the probit-normal relationship. Of course, when both link functions are of probit form, (91) becomes:

$$\Delta_{ij} = \left(\sqrt{1 + \boldsymbol{z}_{ij}^\top \boldsymbol{D} \boldsymbol{z}_{ij}}\right) \cdot \boldsymbol{x}_{ij}^\top \boldsymbol{\xi}^m. \tag{92}$$

For count data, a log-log-normal specification leads to

$$\Delta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\xi}^m - \boldsymbol{z}_{ij}^\top \boldsymbol{D} \boldsymbol{z}_{ij}/2. \tag{93}$$

Note from this expression that, in particular for a random intercept model, i.e., one where $\boldsymbol{z}_{ij}^\top \boldsymbol{b_i} = b_i$ with $b_i \sim N\left(0, \tau^2\right)$, then $\boldsymbol{z}_{ij}^\top \boldsymbol{D} \boldsymbol{z}_{ij} = \sqrt{1+\tau^2}$, which implies that only fixed intercept parameters will be affected in the MMM model compared to their counterparts in the conditional GLMM model. For a general random-effects design $\boldsymbol{z}_{ij}^\top \boldsymbol{b_i}$, this will not be the case. The expression for Δ_{ij}, in the case of probit-probit-normal, log-log-gamma model and the logistic-logistic-Bridge MMM can be found in Griswold and Zeger (2004).

Iddi and Molenberghs (2012), and Efendi et al. (2014) combined the MMM with the CM, by combining (9), (10), and (11) from the CM with (86), (88), and (89) from the MMM in the following way:

$$\begin{aligned}
g(\mu_{ij}^m) &= \boldsymbol{x}_{ij}^\top \boldsymbol{\xi}^m \\
g(\kappa_{ij}) &= \Delta_{ij} + \boldsymbol{z}_{ij}^\top \boldsymbol{b_i} \\
\mu_{ij}^c &= \theta_{ij} \kappa_{ij} \\
\theta_{ij} &\sim \Theta_{ij}\left(\tau_{ij}, \sigma_{ij}^2\right) \\
\boldsymbol{b_i} &\sim F_b\left(\boldsymbol{0}, \boldsymbol{D}\right) \\
Y_{ij}^c = (Y_{ij} | \theta_{ij}, \boldsymbol{b_i}) &\sim F_{Y^c}\left(\mu_{ij}^c, \upsilon\right).
\end{aligned}$$

Note that the response distribution is now conditioned on two sets of random effects, namely the overdispersion and longitudinal ones. This implies that the expression for Δ_{ij} will change slightly. Because $\mu^c_{ij} = \mathrm{E}(Y_{ij}|\theta_{ij}, \boldsymbol{b_i})$, the function Δ_{ij} will then be obtained from the integral equation

$$\begin{aligned}\mu^m_{ij} = g^{-1}(\boldsymbol{x}^\intercal_{ij}\boldsymbol{\xi}^m) &= \int_b \int_\theta \theta_{ij} g^{-1}(\Delta_{ij} + \boldsymbol{z}^\intercal_{ij}\boldsymbol{b_i}) d\Theta_\theta dF_b \\ &= \int_b \mathrm{E}(\theta_{ij}) g^{-1}(\Delta_{ij} + \boldsymbol{z}^\intercal_{ij}\boldsymbol{b_i}) dF_b. \end{aligned} \tag{94}$$

These authors showed that for the logistic-probit-normal model with beta distribution for the overdispersion parameter, i.e., $\theta_{ij} \sim \mathrm{Beta}(\alpha_{1j}, \beta_{2j})$, (94) becomes

$$\Delta_{ij} = \left(\sqrt{1 + \boldsymbol{z}^\intercal_{ij}\boldsymbol{D}\boldsymbol{z}_{ij}}\right) \cdot \Phi^{-1}\{(1 + c_j) \cdot \mathrm{expit}(\boldsymbol{x}^\intercal_{ij}\boldsymbol{\xi}^m)\},$$

where $c_j = \beta_{2j}/\alpha_{1j}$, which can serve as one of several possible constraints, given that the model is now over-parameterized. For the log-log-normal MMM model with $\theta_{ij} \sim \mathrm{Gamma}(\alpha_{1j}, \alpha_{2j})$,

$$\Delta_{ij} = -\log(\alpha_{1j}\alpha_{2j}) + \boldsymbol{x}^\intercal_{ij}\boldsymbol{\xi}^m - \boldsymbol{z}^\intercal_{ij}\boldsymbol{D}\boldsymbol{z}_{ij}/2.$$

The fully marginalized joint distribution can be obtained from integrating out the two random effects. Less effort is needed here because the expressions for the marginal distribution are similar to those found in Molenberghs et al. (2010), except for replacing κ_{ij} with $\kappa_{ij} = g^{-1}(\Delta_{ij} + \boldsymbol{z}^\intercal_{ij}\boldsymbol{b_i})$.

Efendi et al. (2014) showed that, in the particular case of the Weibull-gamma-normal model, the integral equation leads to:

$$\Delta_{ij} = -\log(\alpha_j\beta_j) + \boldsymbol{x}^\intercal_{ij}\boldsymbol{\xi}^m - \boldsymbol{z}^\intercal_{ij}\boldsymbol{D}\boldsymbol{z}_{ij}/2. \tag{95}$$

Should there be no gamma random effects, then the first term on the right hand side of (95) simply drops.

Parameter estimation conveniently proceeds by using the partially marginalized distribution method, explained in Section 8. Only here, the conditional distribution is partly specified through the marginal mean function, which is passed on to the conditional mean function via the connector function.

10.2. A clinical trial in epileptic patients

Building further on the models fitted in Sections 5.1 and 5.3, assume Y_{ij} to follow a Poisson distribution with marginal mean

$$\log(\pi_{ij}^m) = \begin{cases} \beta_{00} + \beta_{01} t_{ij} & \text{if placebo,} \\ \beta_{10} + \beta_{11} t_{ij} & \text{if treatment.} \end{cases} \tag{96}$$

Write the conditional model $\log(\pi_{ij}^c) = \Delta_{ij} + b_i$, with $b_i \sim N(0,d)$ and Δ_{ij} the connector. If also overdispersion is present, consider the COMMM version with then $\pi_{ij}^c = \theta_{ij}\exp(\Delta_{ij} + b_i)$ where $\theta_{ij} \sim \text{Gamma}(\alpha_1, \alpha_2)$ and impose constraint $\alpha_2 = 1/\alpha_1$.

Table 10: *Epilepsy study. Comparison of the log-log-normal MMM with the combined gamma and log-log-normal MMM.*

Effect	Par.	CM Gamma and log-normal	MMM Log-Log-normal	COMMM Gamma and Log-Log normal
Interc. plac.	β_{00}	0.9112(0.1755)	1.3960 (0.1887)	1.4757 (0.1962)
Slope plac.	β_{01}	−0.0248(0.0077)	−0.0143 (0.0044)	−0.0248 (0.0077)
Interc. treatm.	β_{10}	0.6555(0.1782)	1.2256 (0.1901)	1.2200 (0.1970)
Slope treatm.	β_{11}	−0.0118(0.0075)	−0.0120 (0.0043)	−0.0118 (0.0075)
SD RE	$\sqrt{d}$	1.0625(0.0871)	1.0755 (0.0857)	1.0625 (0.0871)
Neg.-bin. par.	α_1	2.4640(0.2113)	—	2.4640 (0.2113)
Neg.-bin. par.	$\alpha_2 = \frac{1}{\alpha_1}$	0.4059(0.0348)	—	0.4059 (0.0348)
−2 log-likelihood		−7664	−6810	−7664

Parameter estimates and standard errors for the log-log-normal MMM and the gamma-log-log-normal COMMM model are presented in Table 10. Observe that the parameter estimates for the two models are very similar, with the same holding for the standard errors. The log-log-normal model improves when the gamma random effect is introduced, as seen from a likelihood ratio comparison. This crucially affects inferences about the difference between the slopes as well as the ratio of the slopes. For the log-log-normal model, the difference of the slopes $\beta_{11} - \beta_{01}$ was found not to be significantly different form zero while the ratio of the slopes β_{11}/β_{01} showed a significant difference from one ($p = 0.7111$ and $p = 0.0376$, respectively). On the other hand both the slope difference ($p = 0.2260$) and ratio ($p = 0.1591$) showed non-significance in the combined model. To understand this, two things need to be borne in mind. First, the above demonstrates that, due to more careful modelling of the association and dispersion structures, inferences about functions of the model parameters may be erroneous in the simpler model, underscoring that care must be taken regarding conclusions based on the simpler model. Indeed, it would lead to a significant treatment difference, whereas the more general combined model showed no evidence for treatment difference. Similar observations were also made by MVD, where the combined Poisson-Gamma-normal showed a strong improvement of the Poisson GLMM model, underscoring the importance of introducing the gamma random effect. Second, and very important, one should not directly compare the estimates in the marginalized and the conditional version. Indeed, in the MMM model, treatment effects, slopes, etc. have a marginal interpretation. In addi-

tion, we can examine the results of fitting a combined beta and log-normal model, which is purely conditionally specified. The interpretation of the latter should be considered at the individual level, or at least for a change between two patients with different covariate profile (e.g., treated versus non-treated), but with the same level of the random effect.

We note from these results that for a random intercept model, only the intercepts parameters are affected but all other parameters remain the same compared to the combined Gamma and log-log-normal model. These would, however, not be the same, for example, for a random intercept and slope model. Given that the log link was used for both marginal and conditional models, we see further that the log-likelihood remains the same across both combined models.

10.3. A clinical trial in onychomycosis

Also here, both the conditional as well as the marginal mean are specified:

$$\begin{aligned} Y_{ij}|b_i &\sim \text{Bernoulli}(\pi^c_{ij}),\\ \Phi^{-1}\left(\pi^c_{ij}\right) &= \Delta_{ij} + b_i,\\ b_i &\sim N(0,d),\\ \text{logit}(\pi^m_{ij}) &= \beta_0 + \beta_1 X_i + \beta_2 t_{ij} + \beta_3 X_i t_{ij}. \end{aligned}$$

Recall that X_i is an indicator for the treatment applied to subject i, t_{ij} is the time at which the jth measurement is taken. For the COMMM model, the conditional mean model is specified as $\pi^c_{ij} = \theta_{ij}\Phi(\Delta_{ij} + b_i)$ where $\theta_{ij} \sim \text{Beta}(\alpha_1, \alpha_2)$ and Φ^{-1} is the probit link. The constraint $c = \alpha_2/\alpha_1$ was imposed.

From the results presented in Table 11, it is again clear that introducing the beta random effect improves significantly the model fit when comparing the log-likelihoods (smaller AIC). Parameter estimates from both models are slightly different, but a much more dramatic effect is seen in precision estimation. For many, but not all parameters, the extended model yields a higher precision. Furthermore, we observed that whereas the broader model encompassing both overdispersion and correlation concludes that there is no effect of the evolution of treatment (β_3) on the response with p-value of $p = 0.0790$, the MMM model results in a significant treatment evolution ($p = 0.0155$). Also presented in Table 11 are the results for a combined beta and probit-normal model whose parameters have a conditional interpretation. The treatment evolution was found to be significant with $p = 0.0343$. By comparing the two combined models, which both account for overdispersion and correlation simultaneously but with different interpretation of parameters, we may conclude that, while there is a significant treatment evolution given subjects, there is no evidence of population average treatment evolution.

Table 11: *A clinical trial in onychomycosis. Comparison of logistic-probit-normal MMM with the combined Beta and logistic-probit-normal MMM.*

Effect	Par.	CM Beta and probit-normal	MMM logistic-probit-normal	COMMM Beta and logistic-probit-normal
Interc.	β_0	−0.7285(0.8622)	−0.6154 (0.1493)	−0.4762 (0.0408)
Treatment	β_1	−0.7404(1.1816)	−0.0382 (0.2120)	−0.1858 (0.1240)
Time	β_2	−0.9109(0.2321)	−0.1529 (0.0190)	−0.1832 (0.0241)
Interaction	β_3	−0.3989(0.1876)	−0.0702 (0.0288)	−0.0691 (0.0392)
SD RE	$\sqrt{d}$	8.6763(1.9535)	2.1061 (0.1904)	8.8901 (0.0152)
Beta-bin. par.	α_2/α_1	0.2828(0.0372)	—	0.2769 (0.0363)
−2 log-likelihood		1259.9	1265.2	1254.0

10.4. Recurrent asthma attacks in children

We now turn to the recurrent asthma data, described in Section 2.5. For each of the 226 patients, their treatment allocation and repeated time-to-event outcomes, the time between the end of the previous to onset of the next attack, Y_{ij} is recorded; the outcome is subject to censoring. Also here, the combined model and its marginalized version are presented next to each other. Regarding the normal random-effects structure, a random intercept b_{i1} (with variance σ_i^2) and a random slope b_{i2} (with variance σ_e^2) is included. While this could be relaxed, both random effects are assumed to be independently normally distributed. Model fitting is done using both full and pairwise likelihood. Parameter estimates (standard errors) are presented in Table 12.

Full likelihood estimates between the ordinary and marginalized models are similar. Treatment effect is not significant. Because marginalization does not change the likelihood, the likelihood ratios are invariant to this operation (Griswold and Zeger, 2004). Because we now include two normally distributed random effects, the connector function (95) uses a different vector z_{ij}. This now implies that the treatment effect estimate changes upon marginalization, although the change is minor.

Table 12: *Asthma study. Original and marginalized combined model results. 'WGN' refers to the Weibull-gamma-normal model, whilst 'C' and 'CM' means censored and censored-marginalized, respectively.*

Effect	Par	WGN-C	WGN-CM	WGN-C	WGN-CM
		Full likelihood		*Pairwise likelihood*	
		Estimate(s.e.)	Estimate(s.e.)	Estimate(s.e.)	Estimate(s.e.)
Treatment	ξ	−0.113(0.106)	−0.111(0.102)	−0.127(0.105)	−0.127(0.105)
Shape	λ	0.014(0.001)	0.017(0.001)	0.025(0.002)	0.027(0.003)
Conj.RE	α	3.566(0.632)	3.566(0.632)	4.583(0.708)	4.584(0.708)
s.d. norm. R.int.	σ_i	0.560(0.068)	0.560(0.068)	0.445(0.039)	0.445(0.039)
s.d. norm. R.eff.	σ_e	0.077(0.734)	0.077(0.741)	11E-4(11E-4)	20E-6(20E-6)
−2 log-likelihood		16649	16649		

Turning attention to results using pairwise likelihood estimation, it is found that the estimates before and after marginalization are still similar. We also see that the estimate of the random slope parameter is virtually zero in all cases, although more pronounced in the pairwise-likelihood case. This does not contradict the results from full likelihood, where this component was non-significant, although the numerical behaviour is quite different.

In the four versions presented in the table, the conjugate random effect parameter is statistically significant. This is important and underscores that neither the standard GLMM nor the available marginalized model of Griswold and Zeger (2004) fit the data adequately.

10.5. Adding zero inflation to the COMMM in the case of counts

In line with Kassahun et al. (2014b), the above construction can be combined with the concepts of Section 5.2, where additional zeroes are allowed for in the CM for count data.

We have to be careful regarding the correct logic. We first specify the model fully hierarchically, derive its marginal mean function, model the former including connector functions, and the latter in the usual parametric way.

10.5.1. Zero-inflation

Dropping indices to diminish notational clutter at this point, the conditional model specification is:

$$P(\boldsymbol{Y}=\boldsymbol{y}|\theta,\boldsymbol{b}) = \begin{cases} \pi^c + (1-\pi^c) f(0|\lambda^c) & y=0, \\ (1-\pi^c) f(y|\lambda^c) & y>0, \end{cases} \tag{97}$$

$$\pi^c = \Phi(\Delta_1 + \boldsymbol{z}_1^{\mathsf{T}} \boldsymbol{b}_1), \tag{98}$$

$$\lambda^c = \theta \exp(\Delta_2 + \boldsymbol{z}_2^{\mathsf{T}} \boldsymbol{b}_2), \tag{99}$$

$$\theta \sim \text{Gamma}(\alpha,\beta), \tag{100}$$

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} \right]. \tag{101}$$

It now follows:

$$\text{E}(Y|\theta,\boldsymbol{b}) = [\pi^c + (1-\pi^c) f(0|\lambda^c)] \cdot 0 + \sum_{y=1}^{\infty} y \frac{e^{-\lambda^c} (\lambda^c)^y}{y!} = (1-\pi^c)\lambda^c. \tag{102}$$

We then require that the marginal mean is of the form:

$$\mathrm{E}(Y) = (1-\pi^m)\lambda^m. \tag{103}$$

The fact that calculating the mean form (102) results in the form (103) does not imply that the marginal model behind (97)–(101) is equal to (103). In fact, as stated before, we know this is not true.

Focusing on the mean functions, as we should, leads to the requirement:

$$\int\int (1-\pi^c)\lambda^c f(\theta) f(\boldsymbol{b}) d\theta d\boldsymbol{b} = (1-\pi^m)\lambda^m. \tag{104}$$

It looks like this is straightforward, but there is a caveat: π^c and λ^c are connected through correlated random effects. In the special but relevant case that $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are uncorrelated, and hence that $\boldsymbol{D}_{12} = 0$, we can solve the system:

$$\int \pi^c f(\boldsymbol{b}_1) d\boldsymbol{b}_1 = \pi^m, \tag{105}$$

$$\int\int \lambda^c f(\boldsymbol{b}_2) f(\theta) d\boldsymbol{b}_2 \, d\theta = \lambda^m. \tag{106}$$

Now, (105) is the classical binary connector function integral equation; (106) is the counterpart for the Poisson case.

In case $\boldsymbol{D}_{12} \neq 0$, the integral equation takes the form:

$$\int\int\int (1-\pi^c)\lambda^c f(\theta) f(\boldsymbol{b}_1) f(\boldsymbol{b}_2|\boldsymbol{b}_1) d\theta \, d\boldsymbol{b}_1 \, d\boldsymbol{b}_2 = (1-\pi^m)\lambda^m. \tag{107}$$

Given that

$$\boldsymbol{b}_2|\boldsymbol{b}_1 \sim N\left(\boldsymbol{D}_{21}\boldsymbol{D}_{11}^{-1}\boldsymbol{b}_1, \boldsymbol{D} = \boldsymbol{D}_{22} - \boldsymbol{D}_{21}\boldsymbol{D}_{11}^{-1}\boldsymbol{D}_{12}\right),$$

and with some straightforward algebra, we obtain the following intermediate step:

$$\mathrm{E}(\theta) e^{\Lambda_2 + \frac{1}{2}\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{D}\boldsymbol{z}_2} \int (1-\pi^c) e^{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{D}_{21}\boldsymbol{D}_{11}^{-1}\boldsymbol{b}_1} f(\boldsymbol{b}_1) d\boldsymbol{b}_1 = (1-\pi^m)\lambda^m.$$

This, in turn, leads to

$$\mathrm{E}(\theta) e^{\Lambda_2 + \frac{1}{2}\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{D}_{22}\boldsymbol{z}_2} \int (1-\pi^c) f(\boldsymbol{b}_1; \mu = \boldsymbol{D}_{12}\boldsymbol{z}_2) d\boldsymbol{b}_1.$$

Upon applying a final transformation ($\widetilde{\boldsymbol{b}}_1 = \boldsymbol{b}_1 - \boldsymbol{D}_{12}\boldsymbol{z}_2 \sim N(\boldsymbol{0},\boldsymbol{D}_{11})$), we find that the Poisson connector remains the same, but for the binary connector, we need to solve:

$$\pi^c = \Phi(\Delta_1 + \boldsymbol{z}_1^{\mathsf{T}}\widetilde{\boldsymbol{b}}_1 + \boldsymbol{z}_1^{\mathsf{T}}\boldsymbol{D}_{12}\boldsymbol{z}_2).$$

Of course, this is equal to the standard binary connector problem, but merely with a shift applied to Δ_1.

10.5.2. Hurdle models

Using the same simplified notation as before, we now have:

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta,\boldsymbol{b}) = \begin{cases} \pi^c & y = 0, \\ (1-\pi^c)\frac{f(y|\lambda^c)}{1-f(0|\lambda^c)} & y > 0, \end{cases}, \tag{108}$$

with the rest of the model specified by (98)–(101). It now follows:

$$\mathrm{E}(Y|\theta,\boldsymbol{b}) = \pi^c \cdot 0 + \frac{1-\pi^c}{1-f(0|\lambda^c)}\sum_{y=1}^{\infty} f(y|\lambda^c) = \frac{1-\pi^c}{1-f(0|\lambda^c)}\cdot\lambda^c = \frac{1-\pi^c}{1-e^{-\lambda^c}}\cdot\lambda^c. \tag{109}$$

Also here, we require conditional mean (109) to take the same form marginally:

$$\mathrm{E}(Y) = (1-\pi^m)\cdot\frac{\lambda^m}{1-e^{-\lambda^m}}.$$

When $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are independent, we find the classical connector integral equation for the binary component:

$$\int \pi^c f(\boldsymbol{b}_1)d\boldsymbol{b}_1 = \pi^m.$$

For the count connector function, we need to solve:

$$\iint \frac{\lambda^c}{1-e^{-\lambda^c}} f(\boldsymbol{\theta})f(\boldsymbol{b}_2)d\theta d\boldsymbol{b}_2 = \frac{\lambda^m}{1-e^{-\lambda^m}}.$$

More explicitly,

$$\iint \frac{\theta e^{\Delta_2+\boldsymbol{z}_2^{\mathsf{T}}\boldsymbol{b}_2}}{1-e^{-\left[\theta e^{\Delta_2+\boldsymbol{z}_2^{\mathsf{T}}\boldsymbol{b}_2}\right]}} f(\boldsymbol{\theta})f(\boldsymbol{b}_2)d\theta d\boldsymbol{b}_2 = \frac{e^{\boldsymbol{x}_2^{\mathsf{T}}\boldsymbol{\xi}}}{1-e^{-e^{\boldsymbol{x}_2^{\mathsf{T}}\boldsymbol{\xi}}}}.$$

Of course, also here, a further modification is needed when the two normal random effects are correlated. In line with what we find in the zero-inflated case, we now have:

$$\iiint \Phi(\Delta_1+\boldsymbol{z}_1^{\mathsf{T}}\boldsymbol{b}_1)\cdot\frac{\theta e^{\Delta_2+\boldsymbol{z}_2^{\mathsf{T}}\boldsymbol{b}_2}}{1-e^{-\left[\theta e^{\Delta_2+\boldsymbol{z}_2^{\mathsf{T}}\boldsymbol{b}_2}\right]}}f(\boldsymbol{\theta})f(\boldsymbol{b}_1)f(\boldsymbol{b}_2|\boldsymbol{b}_1)d\theta d\boldsymbol{b}_1 d\boldsymbol{b}_2=\Phi(\boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{\gamma})\cdot\frac{e^{\boldsymbol{x}_2^{\mathsf{T}}\boldsymbol{\xi}}}{1-e^{-e^{\boldsymbol{x}_2^{\mathsf{T}}\boldsymbol{\xi}}}}.$$

However now, the denominator under the integrand implies that simplification is less straightforward, and hence a Newton-Raphson approach for the pair (Δ_1,Δ_2) is an obvious way forward. Note that in the zero-inflated case, we were able to derive intuitive expressions for Δ_1 and Δ_2, but these are not unique, given that there is one integral equation with two tuning parameters. Thus, at best, one can find an algebraic expression for Δ_1, because even in the uncorrelated random-effects case, there is no closed form for the count connector. Therefore, we can simply set one of the two equal to zero, $\Delta_1\equiv 0$, say, and then solve the reduced integral equation for Δ_2.

11. Joint modelling of several outcomes

The common recording of not one but several longitudinal sequences is common practice nowadays. The use of normal random effects in the combined model allows one to simultaneously analyse several longitudinal sequences, which do not even need to be of the same type.

Iddi and Molenberghs (2012a) made use of this possibility to jointly model a continuous and a binary longitudinal sequence. Kassahun et al. (2015) jointly modelled a continuous and a zero-inflated count sequence. Njeru Njagu et al. (2016) considered the case where repeated time-to-event outcomes are coupled with a longitudinal outcome of various types (continuous, binary, count) as well as the joint modelling of a continuous and binary outcome. Ivanova, Molenberghs, and Verbeke (2016) allow for ordinal outcomes as well. Ghebretinsae et al. (2012) used CM joint modelling to analyse comet assay data.

To give an example, let us consider Case 1 of Njeru Njagu et al. (2016), where a linear mixed model for the continuous outcome is coupled with a Weibull-gamma-normal model for the time-to-event outcome. The joint model, conditional on both the normal and gamma random effects, takes the form:

$$\begin{aligned}f(\boldsymbol{t}_i,\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\psi}_i)=&\prod_k \lambda_k\rho_k t_{ik}^{\rho_k-1}\psi_{ik}e^{\mu_{ik}+d_{ik}}e^{-\lambda_k t_{ik}^{\rho_k}\psi_{ik}e^{\mu_{ik}+d_{ik}}}\\&\times\frac{1}{(2\pi)^{\frac{n_i}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}e^{\frac{-1}{2}(\boldsymbol{y}_i-X_i\boldsymbol{\xi}-Z_i\boldsymbol{b}_i)^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-X_i\boldsymbol{\xi}-Z_i\boldsymbol{b}_i)},\end{aligned}\tag{110}$$

with $\boldsymbol{\Sigma}_i$ an n_i by n_i diagonal covariance matrix with diagonal elements σ^2. Also, $\boldsymbol{t}_i$ is the set of p_i survival times for cluster i, while $\mathbf{y}_i$ is the vector of n_i continuous outcomes. Moreover, $d_{ik} = \boldsymbol{w}_{ik}^{\top}\boldsymbol{b}_i$, where $\boldsymbol{w}_{ik}^{\top}$ is a vector of scale factors. Here, the index k refers to the kth survival time in cluster i. For the scale and shape parameters in the baseline hazard, we consider a more general case, where both λ and ρ are allowed to vary between members of a cluster. The continuous and survival processes are assumed independent, conditional on the shared normal random effects. Note that the shared random effect in the way considered here is generic. For example, one can choose $\boldsymbol{z}_{ij}$ and $\boldsymbol{w}_{ij}$ such that some random effects are present in the normal-outcome linear predictors, with others influencing the Weibull predictor, and a third set influencing both. As such, our paradigm encompasses both shared as well as correlated random effects.

12. Influence diagnostics

Because of the relative novelty of the CM and its extensions, development regarding model assessment and diagnostic tools has been limited. Rakhmawati et al. (2017) presented local influence diagnostic tools for the count-data CM. Rakhmawati et al. (2016ab) extended this to allow for zero inflation and incomplete data, respectively.

Local influence was presented by Cook (1986). The impact of individuals and measurements on the analysis is assessed by comparing standard maximum likelihood estimates with those resulting from slightly perturbing the contribution of an individual or measurement. The method is to be contrasted with global influence (case deletion), where impact is assessed by simply deleting an individual or measurement. While conceptually a bit technical, it is easy and fast to use in practice and in several cases it leads to interpretable components of influence. Lesaffre and Verbeke (1998) introduced influence assessment for the linear mixed model. Ouwens, Tan, and Berger (2001) applied local influence to the Poisson-normal model. Rakhmawati et al. (2017) followed their ideas, but with extensions in three directions. First, they provided closed-form expressions, based on an analytical form for the marginal likelihood function, as well as based on an integral form for the said likelihood. Second, they considered three important cases: binary, count, and time-to-event. Third, they started from the combined model, rather than merely from the GLMM.

The general theory behind so-called case-weighted likelihood is as follows. Let the log-likelihood for the generalized linear mixed model or its combined extension take the form

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}), \tag{111}$$

in which $\ell_i(\boldsymbol{\theta})$ is the contribution of the ith individual to the log-likelihood. Let

$$\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^{N} \omega_i \ell_i(\boldsymbol{\theta}), \tag{112}$$

and denote the perturbed version of $\ell(\boldsymbol{\theta})$, depending on an N-dimensional vector $\boldsymbol{\omega}$ of weights, assumed to belong to an open subset Ω of $\mathbb{R}^N$. The original log-likelihood (111) follows for $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = (1,1,\ldots,1)^{\mathsf{T}}$. Let $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator for $\boldsymbol{\theta}$, obtained by maximizing $\ell(\boldsymbol{\theta})$, and let $\widehat{\boldsymbol{\theta}}_\omega$ denote the estimator for $\boldsymbol{\theta}$ under $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$. Cook (1986) proposed to measure the distance between $\widehat{\boldsymbol{\theta}}_\omega$ and $\widehat{\boldsymbol{\theta}}$ by the likelihood displacement: $\mathrm{LD}(\boldsymbol{\omega}) = 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\right)$. $\mathrm{LD}(\boldsymbol{\omega})$ will be large if $\ell(\boldsymbol{\theta})$ is strongly curved at $\widehat{\boldsymbol{\theta}}$. A graph of $\mathrm{LD}(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ brings out information on the influence of case-weight perturbations. The graph is the geometric surface formed by the values of the $(N+1)$-dimensional vector

$$\boldsymbol{\xi}(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\omega} \\ \mathrm{LD}(\boldsymbol{\omega}) \end{pmatrix}$$

as $\boldsymbol{\omega}$ varies throughout Ω. Following Cook (1986) and Verbeke and Molenberghs (2000), we will refer to $\boldsymbol{\xi}(\boldsymbol{\omega})$ as an influence graph.

Cook (1986) derived a convenient computational scheme. Let $\boldsymbol{\Delta}_i$ be the s-dimensional vector of second-order derivatives of $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$, w.r.t. ω_i and all components of $\boldsymbol{\theta}$, and evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Also, write Δ for the $s \times r$ matrix with $\boldsymbol{\Delta}_i$ in the ith column. Let $\ddot{L}$ denote the $s \times s$ matrix of second derivatives of $\ell(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. For any unit vector $\boldsymbol{h}$ in Ω, it follows that:

$$C_h = 2\left| \boldsymbol{h}^{\mathsf{T}} \boldsymbol{\Delta}^{\mathsf{T}} \ddot{\boldsymbol{L}}^{-1} \boldsymbol{\Delta} \boldsymbol{h} \right|. \tag{113}$$

Various choices for $\boldsymbol{h}$ have received attention. First, as will be done here, one can focus on subject i only, by choosing $\boldsymbol{h} = \boldsymbol{h_i}$, the zero vector with a sole 1 in the ith position. Local influence then is

$$C_i \equiv C_{h_i} = 2\left| \boldsymbol{\Delta}_i^{\mathsf{T}} \ddot{\boldsymbol{L}}^{-1} \boldsymbol{\Delta}_i \right|. \tag{114}$$

Second, $\boldsymbol{h} = \boldsymbol{h}_{\max}$ can be considered, the direction of maximal normal curvature (Verbeke and Molenberghs 2000). Expressions can be derived when only a sub-vector of the parameter vector is of interest as well. We refer to Rakhmawati et al. (2017) for details.

These authors derived interpretable expressions for several cases. For example, for the probit-normal case they showed that

$$||\boldsymbol{\Delta}_i||^2 = \left(\sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij}\right)\left(\sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij}\right)^{\intercal} + \sum_{k,l}\left\{-\tfrac{1}{2}(\boldsymbol{D}^{-1})_{kl} + \frac{1}{2}(\boldsymbol{D}^{-1}\boldsymbol{D}^{-1})_{kl}\mathrm{Var}(\boldsymbol{b_i})\right\}^2.$$

Let $C_i = C_{1i} + C_{2i}$ with:

$$C_{1i} = 2||\ddot{\boldsymbol{L}}^{-1}||\ ||\boldsymbol{r}_i^{\intercal}\boldsymbol{x}_i||^2\cos(\varphi_i), \tag{115}$$

$$C_{2i} = \tfrac{1}{2}||\ddot{\boldsymbol{L}}^{-1}||\ ||(\boldsymbol{D}^{-1})_{kl} - (\boldsymbol{D}^{-1}\boldsymbol{D}^{-1})_{kl}\mathrm{Var}(\boldsymbol{b_i})||^2\cos(\varphi_i), \tag{116}$$

where $\boldsymbol{r}_i^{\intercal}\boldsymbol{x}_i = \sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij}$. Note that C_{1i} and C_{2i} are the contributions of subject i to local influence C_i from $\boldsymbol{\beta}$ and $\boldsymbol{D}$, respectively. Now, C_{1i} and C_{2i} were shown to equal:

$$C_{1i} = 2||\ddot{\boldsymbol{L}}^{-1}||\ ||\boldsymbol{x}_i\boldsymbol{x}_i^{\intercal}||\ ||\boldsymbol{r}_i||^2\cos(\alpha_i)\cos(\varphi_i), \tag{117}$$

$$\begin{aligned} C_{2i} = \tfrac{1}{2}||\ddot{\boldsymbol{L}}^{-1}||\cos(\varphi_i) \times \big[&\mathrm{tr}\left\{(\boldsymbol{D}^{-1})_{kl}^2\right\} - \mathrm{tr}\left\{2(\boldsymbol{D}^{-1})_{kl}(\boldsymbol{D}^{-1}\boldsymbol{D}^{-1})_{kl}\mathrm{Var}(\boldsymbol{b_i})\right\}) \\ &+\mathrm{tr}\left\{(\boldsymbol{D}^{-1}\boldsymbol{D}^{-1})_{kl}^2\mathrm{Var}(\boldsymbol{b_i})^2\right\}\big], \end{aligned} \tag{118}$$

where $\cos(\alpha_i)$ is the angle between $\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^{\intercal})$ and $\mathrm{vec}(\boldsymbol{r}_i\boldsymbol{r}_i^{\intercal})$, and φ_i is the angle between $\mathrm{vec}(-\ddot{\boldsymbol{L}}^{-1})$ and $\mathrm{vec}(\boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^{\intercal})$. Hence, the interpretable components of C_i in the case of the Poisson-normal model can be described using the 'length of the fixed effect' ($||\boldsymbol{x}_i\boldsymbol{x}_i^{\intercal}||$), the 'squared length of the residual' ($||\boldsymbol{r_i}||^2$), and the 'squared of random effect variability' ($\mathrm{Var}(\boldsymbol{b_i})^2$).

Rakhmawati et al. (2017) derived similar expressions for the probit-normal, logit-normal and Weibull-normal models.

12.1. A clinical trial in epileptic patients

We start from the Poisson-normal (P-N) and Poisson-gamma-normal (PGN) models studied before:

$$\ln(\lambda_{ij}) = \begin{cases} (\xi_{00} + b_i) + \xi_{01}t_j & \text{if placebo} \\ (\xi_{10} + b_i) + \xi_{11}t_j & \text{if treated,} \end{cases} \tag{119}$$

where Y_{ij} represent the number of epileptic seizures patient i experienced during week j, t_j is the time point at which Y_{ij} was measured, and with random intercept $b_i \sim N(0,d)$. Parameter estimates are given in Table 13. Index plots (versus patient ID) for various local influence analyses are given in Figure 2. The top row of the plot represents the total local influence, with subsequent rows representing influence for sub-vectors: fixed effects, random-intercept variance d, and, for the (PGN), the overdispersion parameter

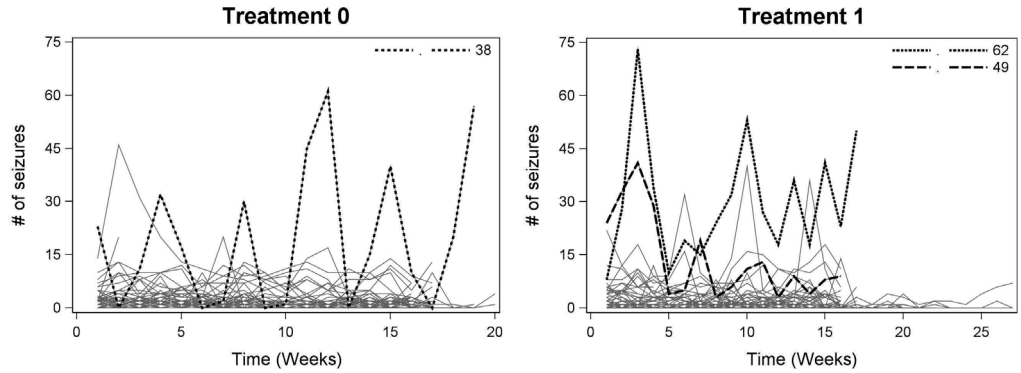

Figure 1: *Epilepsy data. Individual profiles.*

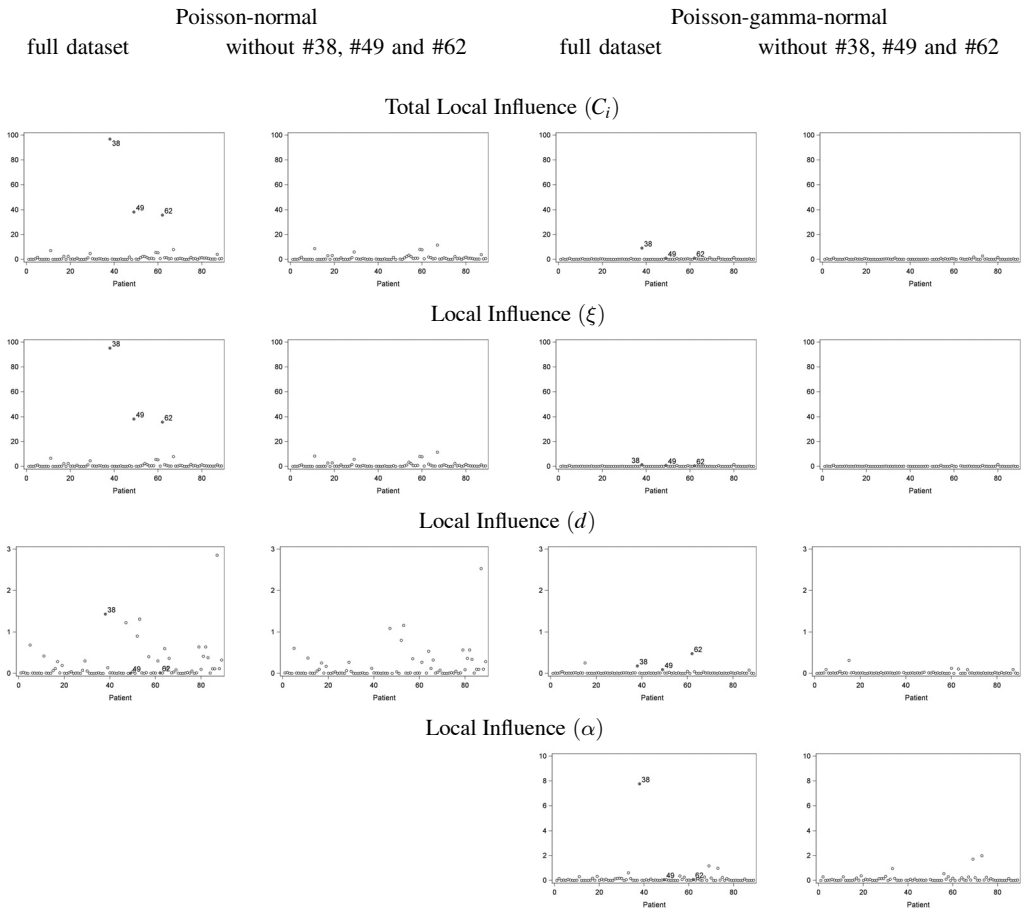

Figure 2: Epilepsy data. Local influence plots.

Table 13: *Local influence. Parameter estimates (standard errors) for the generalized linear mixed and combined models.*

Epilepsy		Poisson-normal		Poisson-gamma-normal	
Effect	Par.	Full	#(38,49,62)	Full	#(38,49,62)
Interc. plac.	ξ_{00}	0.818(0.168)	0.903(0.157)	0.911(0.176)	0.907(0.163)
Slope plac.	ξ_{01}	−0.014(0.004)	−0.031(0.005)	−0.025(0.008)	−0.031(0.008)
Interc. treat.	ξ_{10}	0.648(0.170)	0.492(0.162)	0.656(0.178)	0.510(0.169)
Slope treat.	ξ_{11}	−0.012(0.004)	−0.007(0.005)	−0.012(0.007)	−0.009(0.007)
Treat. eff.	$\xi_{11}-\xi_{10}$	0.002(0.006)	0.024(0.007)	0.013(0.011)	0.022(0.011)
Treat. eff.	ξ_{11}/ξ_{10}	0.840(0.398)	0.236(0.170)	0.475(0.335)	0.281(0.250)
Std. rand. int.	σ	1.076(0.086)	0.982(0.081)	1.063(0.087)	0.969(0.082)
Overdisp. par.	α			2.464(0.211)	3.109(0.329)
Onychomycosis		Logit-normal		Logit-beta-normal	
Effect	Par.	Full	#(6,30,53)	Full	#(6,30,53)
Interc. plac.	ξ_0	−1.630(0.435)	−1.940(0.523)	−1.604(4.026)	−2.420(3.089)
Slope plac.	ξ_1	−0.404(0.046)	−0.430(0.049)	−6.478(1.439)	−6.075(1.264)
Interc. treat.	ξ_2	−1.749(0.448)	−1.604(0.536)	−16.21(3.58)	−15.21(3.02)
Slope treat.	ξ_3	−0.563(0.060)	−0.872(0.100)	−8.075(1.600)	−8.755(1.437
Treat. eff.	$\xi_{11}-\xi_{10}$	−0.159(0.072)	−0.442(0.105)	−1.596(0.858)	−2.680(0.822)
Treat. eff.	ξ_{11}/ξ_{10}	1.394(0.206)	2.028(0.302)	1.246(0.148)	1.441(0.171)
Std. rand. int.	σ	4.015(0.381)	4.814(0.490)	60.88(14.22)	56.47(11.69)
Overdisp. par.	α/β			0.281(0.035)	0.231(0.031)

α, respectively. Patients #38, #49, and #62 stand out with large total influence C_i when compared to other patients. Importantly, influences show a major drop when switching from (P-N) to (PGN). This is most prominently seen for #38. For an explanation, turn to the right hand panel of Figure 1. Patient #38 (and to some extent also #62 on the left hand side) alternates periodically between very high numbers of episodes and periods virtually without. This implies that their mean, variance, and association structure are rather different from the majority of subjects. The impact on the mean structure, by way of the fixed effects, is evident in the second row. For the (P-N) it is less clear when turning to d, but we gain a lot of insight from the (PGN) results. Overall influence and influence on $\boldsymbol{\xi}$ reduce drastically, but there now is clear influence on d and α. What it means is that with these subjects present, the overdispersion parameter helps capturing their anomalous behaviour, which 'deflates' d. In other words, adding overdispersion protects the inferentially crucial fixed-effects parameter vector. When removing these subjects, and also #49, little or no influence is left.

Note that the (PGN) model fitted to the full dataset exhibits a smaller value for α, which corresponds to more overdispersion (no overdispersion corresponds to α approaching $+\infty$), while it does not vanish with removal of the three subjects. Thus, there appears to be genuine overdispersion in the data, further inflated by the influential subjects.

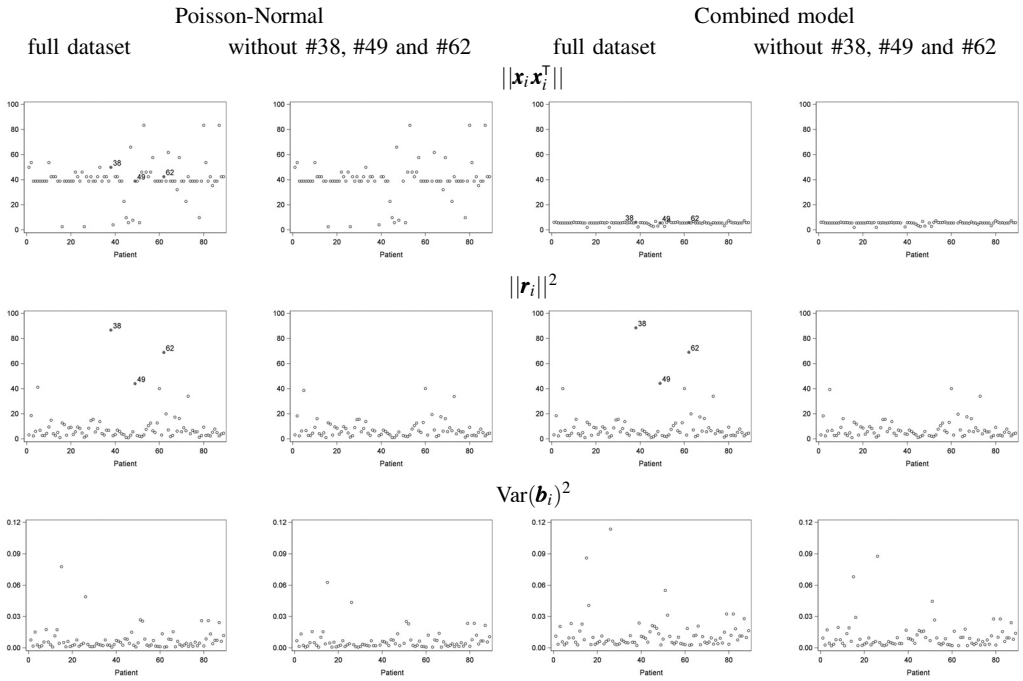

Figure 3: Epilepsy data. Plots of interpretable components of local influence.

In agreement with MVD, MVDV, and our earlier analysis, Rakhmawati et al. (2017) considered the treatment effect in additive ($\xi_{11}-\xi_{01}$) and multiplicative (ξ_{11}/ξ_{01}) form. Important differences are seen on the additive scale. (P-N) shows no significance ($p=0.7106$), which is sustained for (PGN), with $p=0.2225$. Removing the influential subjects leads to a highly significant result for (P-N), with $p=0.0009$, which changes to the still significant $p=0.0350$ for (PGN). Hence, the influential subjects mask a treatment effect. This is logical, because the influential subjects exhibit an oscillating behaviour, introducing an important source of variability. At the multiplicative level, where the null hypothesis is for the ratio to be 1, the story is nicely confirmed, with $p=0.6872$ and $p=0.1166$ for (P-N) and (PGN), respectively; the counterparts after deletion are $p<0.0001$ and $p=0.0040$, respectively.

To get further insight as to why these subject have higher influence than others, plots with interpretable components are given in Figure 3: 'squared length of the fixed effects' $||\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}||$, 'squared length of the residual' $||\boldsymbol{r}_i||^2$, and 'random-effect variability' $\mathrm{Var}(b_i)^2$. It is hardly surprising that #38 stands out in terms of $||\boldsymbol{r}_i||^2$. Influences on #49 and #62 are less pronounced.

Our analysis has provided insight not available from earlier analysis. The influential subjects exhibit a cyclic behaviour not observed in the majority of patients, but at the same time well documented. Based on these findings, a focused clinical discussion can take place, to determine the course of action. Options include removal, retention, or

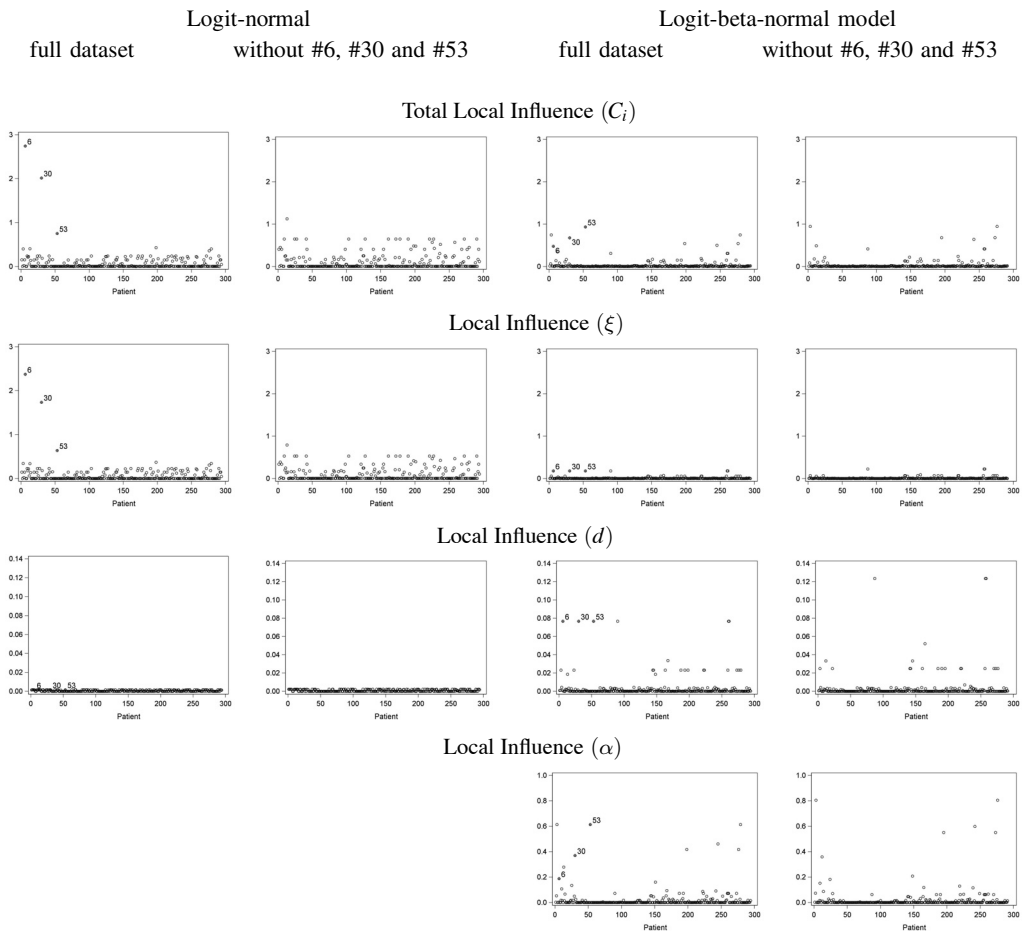

Figure 4: Onychomycosis data. Local influence plots.

even setting up a dedicated study to further scrutinize this sub-population. In this case, a small group of patients with oscillating behaviour between two poles has been identified.

12.2. A clinical trial in onychomycosis

Before, we assumed $Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij})$, where Y_{ij} is severity of infection (1 for severe, 0 for non-severe) for patient i at occasion j, T_i is the treatment indicator (1 for experimental, 0 for standard) for subject, t_j is the time point (months) at which the jth measurement has been taken, and $b_i \sim N(0,d)$. The conditional success probability is expressed as:

$$\text{logit}(\pi_{ij}) = \xi_1(1-T_i) + \xi_2(1-T_i)t_{ij} + \xi_3 T_i + \xi_4 T_i t_{ij} + b_i.$$

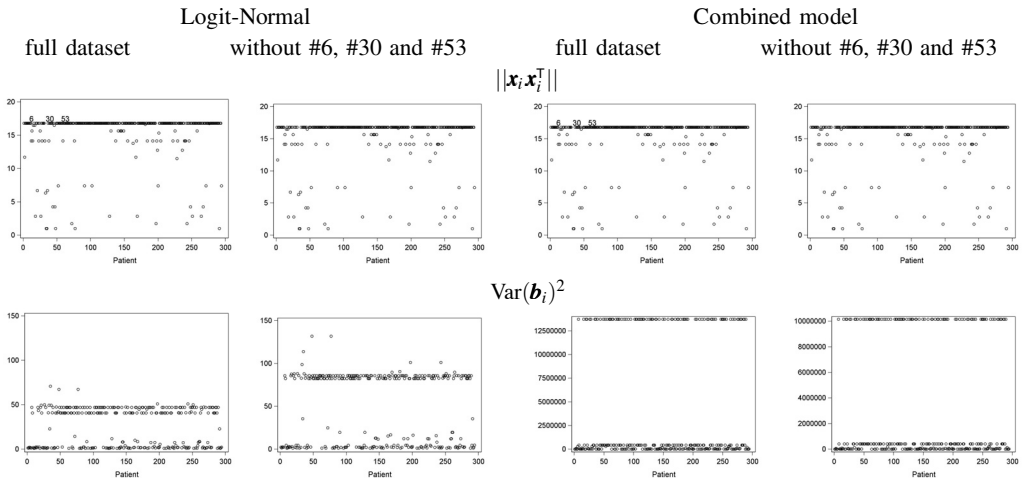

Figure 5: Onychomycosis data. Plot of interpretable components of local influence.

Both the logit-normal (L-N) and logit-beta-normal (LBN) are fitted. Parameter estimates (standard errors) are displayed in Table 13, with local influence plots in Figure 4. Subjects #6, #30, and #53 are detected as influential, overall, and with respect to the fixed effects, in the (L-N). Accommodating overdispersion, hence turning to the (LBN), deflates the magnitude of influence. Likewise, influence is drastically diminished by removing these three subjects. Thus, in case the influential subjects should remain in the analysis, the (LBN) may be the most sensible route forward. Alternatively, in case they are considered anomalous, one can remove them. To decide on which scenario is preferred in this case, we note that all three subjects are unusual: they set out with a sequence of non-severe ratings, but then switch to a severe rating ('0000111' for #6, '0000011' for #30, and '0000001' for #53). Arguably, there is no reason to remove these subjects from analysis, partly also to safeguard randomization. However, it is uncommon to switch from non-severe to severe in this particular way, so these patients must be further clinically scrutinized. Also for these data, the interpretable components do not lead to further insight (Figure 5).

The (L-N) and (LBN) lead to borderline significance when applied to the full data [$p = 0.0268$ additively and $p = 0.0560$ multiplicatively for (L-N); $p = 0.0627$ additively and $p = 0.0964$ multiplicatively for (LBN)]. When influential subjects are removed, these values all become highly significant [in the same order, $p < 0.0001$, $p = 0.0007$, $p = 0.0011$, and $p = 0.0099$]. These findings are qualitatively similar to the epilepsy cases.

13. Concluding remarks

Based on work by MVD, MVDV, and subsequent references, we have reviewed a general and flexible framework for such combinations, starting from arbitrary generalized linear models and exponential family members. Specific emphasis is placed on normally distributed, binary, binomial, count, and time-to-event outcomes. There are various reasons to do so. First, non-Gaussian hierarchical data exhibit three important features: (1) the mean structure; (2) the variance structure; and (3) the correlation structure. Our proposed framework features: (a) a mean structure; (b) overdispersion, often conjugate random-effects; (c) normal random effects. It will be clear from our case studies that model fit can be improved and hence model interpretation changed, by shifting to the extended model. Second, especially in cases where the variance and/or correlation structures are of interest (e.g., surrogate marker evaluation, psychometric evaluation, etc.) such extensions are useful. Third, even when interest remains with more conventional models, such as the GLMM, the extended model can serve as a goodness-of-fit tool. Fourth, because we can derive closed-form expressions for both standard and extended models, the accuracy of parameter estimation and resulting inferences can be improved, while obviating the need for tedious numerical integration techniques. Fifth, the analysis of the case studies corroborates this need. While the model extends the classical GLMM, it is actually easy to fit when standard non-linear mixed-model software is available, such as the SAS procedure NLMIXED.

Because for most of these combined models, and their GLMM sub-models, closed form moment expressions are available, derived quantities such as correlation are easy to obtain. Furthermore, versions with mean parameters that are directly marginally interpretable can be constructed. Also, the model lends itself naturally to the joint modelling of several hierarchical sequences simultaneously. Diagnostics based on local influence ideas have been developed as well.

While we have aimed to give an extensive overview of a modelling framework to accommodate data hierarchies and overdispersion, inevitably a number of topics have been left untouched. For example, Molenberghs and Verbeke (2011b), Pryseley et al. (2011) examined the occurrence of negative variance components in hierarchical data, which is also relevant for this context. Likewise, underdispersion has received some treatment (Oliveira et al., 2016; 2017).

Acknowledgment

Financial support from the IAP research network #P7/06 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. This work was partially supported by CNPq, a Brazilian science funding agency.

References

Abrams, S., Aerts, M., Molenberghs, G. and Hens, N. (2017). Parametric overdispersed frailty models for current status data. *Biometrics*, DOI: 10.1111/biom.12692.

Aerts, M., Geys, H., Molenberghs, G. and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: John Wiley & Sons.

Aitkin, M. (l999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55, 117–128.

Alonso, A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N.J., Shkedy, Z. and Van der Elst, W. (2017). *Applied Surrogate Endpoint Evaluation with SAS and R*. Boca Raton: Chapman & Hall/CRC.

Alfò, M. and Aitkin, M. (2000). Random coefficient models for binary longitudinal responses with attrition. *Statistics and Computing*, 10, 279–288.

Aregay, M., Shkedy, Z. and Molenberghs, G. (2013). A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: a simulation study. *Computational Statistics and Data Analysis*, 57, 233–245.

Aregay, M., Shkedy, Z. and Molenberghs, G. (2015). Comparison of additive and multiplicative Bayesian models for longitudinal count data with overdispersion parameters: a simulation study. *Communications in Statistics, Computation and Simulation,* 44, 454–473.

Ashford, J.R. and Sowden, R.R. (1970) Multivariate probit analysis. *Biometrics*, 26, 535–546.

Bennett, S. (1983). Log-logistic regression models for survival data. *Applied Statistics*, 32, 165–171.

Böhning, D. (2000) *Computer-assisted Analysis of Mixtures and Applications. Meta-analysis, Disease Mapping and Others*. London: Chapman & Hall/CRC.

Booth, J.G., Casella, G., Friedl, H. and Hobert, J.P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, 3, 179–181.

Borgermans, L., Goderis, G., Van Den Broeke, C., Verbeke, G., Carbonez, A., Ivanova, A., Mathieu, C., Aertgeerts, B., Heyrman, J. and Grol, R. (2009). Interdisciplinary diabetes care Teams operating on the interface between primary and specialty care are associated with improved outcomes of care: Findings from the Leuven Diabetes Project, *BMC Health Services Research*, 9, 179.

Breslow, N. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, 33, 38–44.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82, 81–91.

Collett, D. (2003). *Modelling Survival Data in Medical Research* (2nd ed.). Boca Raton: CRC Press.

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, 48, 133–169.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall/CRC.

Dean, C.B. (1991). Estimating equations for mixed-Poisson models. In: *Estimating Functions*, V.P. Godambe (Ed.). Oxford: Oxford University Press.

De Backer, M., De Keyser, P., De Vroey, C. and Lesaffre, E. (1996). A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250 mg/day vs. itraconazole 200 mg/day – a double-blind comparative trial. *British Journal of Dermatology*, 134, 16–17.

Del Fava, E., Shkedy, Z., Aregay, M. and Molenberghs, G. (2014). Modelling multivariate, overdispersed binomial data with additive and multiplicative random effects. *Statistical Modelling*, 14, 99–133.

Duchateau, L. and Janssen, P. (2007). *The Frailty Model*. New York: Springer.

Efendi, A. and Molenberghs, G. (2013). A multilevel model for hierarchical, repeated, and overdispersed time-to-event outcomes and its estimation strategies. *Journal of Biopharmaceutical Statistics*, 23, 1420–1434.

Efendi, A., Molenberghs, G. and Iddi, S. (2014). A Marginalized combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event outcomes. *Communications in Statistics*, 43, 4806–4828.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, 48, 1–22.

Faught, E., Wilder, B.J., Ramsay, R.E., Reife, R.A., Kramer, L.D., Pledger, G.W. and Karim, R.M. (1996). Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages, *Neurology*, 46, 1684–1690.

Gentle, J.E. (2003). *Random Number Generation and Monte Carlo Methods*. New York: Springer.

Ghebretinsae, A.H., Faes, C., Molenberghs, G., De Boeck, M. and Geys, H. (2013). A Bayesian generalized frailty model for comet assays. *Journal of Biopharmaceutical Statistics*, 11, 449–455.

Ghebretinsae, A., Faes, C., Molenberghs, G., Geys, H. and Van der Leede, B.-J. (2012). Joint modelling of hierarchically clustered and overdispersed non-Gaussian continuous outcomes for comet assay data. *Pharmaceutical Statistics*, 11, 449–455.

Gibbons, R.D. and Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, 53, 1527–1537.

Greene, W. (1994). *Accounting for Excess Zeros and Sample Selection in Poisson and Negative-Binomial Regression Models*. Working Paper EC-94–10, Department of Economics, New York University.

Griswold, M.E. and Zeger, S.L. (2004). *On Marginalized Multilevel Models and their Computation*. John Hopkins University, Dept. of Biostatistics Working Papers.

Guilkey, D.K. and Murphy, J.L. (1993). Estimation and testing in the random effects probit model. *Journal of Econometrics*, 59, 301–317.

Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data *Biometrics*, 55, 688–698.

Heagerty, P.J. and Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference, *Statistical Science*, 15, 1–26.

Hedeker, D. and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 51, 933–944.

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

Hinde, J. and Demétrio, C.G.B. (1998b). *Overdispersion: Models and Estimation*. São Paulo: XIII Sinape.

Iddi, S. and Molenberghs, G. (2012a). A combined overdispersed and marginalized multilevel model. *Computational Statistics and Data Analysis*, 56, 1944–1951.

Iddi, S. and Molenberghs, G. (2012b). A joint marginalized multilevel model for continuous and binary longitudinal outcomes. *Journal of Applied Statistics*, 56, 1944–1951.

Iddi, S. and Molenberghs, G. (2013). A marginalized model for zero-inflated, overdispersed and correlated count data. *Electronic Journal of Applied Statistical Analysis*, 6, 149–165.

Iddi, S., Molenberghs, G., Aregay, M. and Kalema, G. (2014). Empirical Bayes estimates for correlated hierarchical data with overdispersion. *Pharmaceutical Statistics*, 13, 316–326.

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

Ivanova, A., Molenberghs, G. and Verbeke, G. (2016). Mixed model approaches for joint modelling of different types of responses. *Journal of Biopharmaceutical Statistics*, 26, 601–618.

Johnson, N.L., Kemp, A., Kotz, S. (2005). *Univariate Discrete Distributions* (3rd ed.). Hoboken: John Wiley & Sons.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics, Continuous Univariate Distributions*, Vol. 2. Boston: Houghton-Mifflin.

Kalema, G., Iddi, S. and Molenberghs, G. (2016). The combined model: a tool for simulating correlated counts with overdispersion. *Communications in Statistics*, 45, 2491–2510.

Kalema, G. and Molenberghs, G. (2015). Generating correlated and/or overdispersed count data; a SAS implementation. *Journal of Statistical Software*, 00, 000–000.

Kassahun, W., Neyens, T., Faes, C., Molenberghs, G. and Verbeke, G. (2014a). A Zero-inflated overdispersed hierarchical Poisson model. *Statistical Modelling*, 14, 439–456.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C. and Verbeke, G. (2012). Modelling overdispersed longitudinal binary data from the Jimma longitudinal studies using a combined beta and normal random-effects model. *Archives of Public Health*, 70, Article 7.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C. and Verbeke, G. (2014b). Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. *Statistics in Medicine*, 33, 4402–4419.

Kassahun, W., Neyens, T., Molenberghs, G., Faes C. and Verbeke, G. (2015). A joint model for hierarchical continuous and zero-inflated overdispersed count data. *Journal of Statistical Computation and Simulation*, 85, 552–571.

Kenward, M.G. and Molenberghs, G. (2016). A taxonomy of mixing and outcome distributions based on conjugacy and bridging. *Communications in Statistics, Theory and Methods*, 45, 1953–1968.

Kleinman, J. (1973). Proportions with extraneous variance: single and independent samples. *Journal of the American Statistical Association*, 68, 46–54.

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

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, 15, 209–225.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, 58, 619–678.

Lee, Y. and Nelder, J.A. (2001a). Two ways of modelling overdispersion. *Applied Statistics*, 49, 591–598.

Lee, Y. and Nelder, J.A. (2001b). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika*, 88, 987–1006.

Lee, Y. and Nelder, J.A. (2003). Extended-REML estimators. *Journal of Applied Statistics*, 30, 845–856.

Lee, Y., Nelder, J.A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood.* Boca Raton: Chapman & Hall/CRC.

Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine*, 10, 1391–1403.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, 54, 570–582.

Liang, K.Y. and McCullagh, P. (1993). Case studies in binary dispersion. *Biometrics*, 49, 623–630.

Liu, L. and Yu, Z. (2008) A likelihood reformulation method in non-normal random-effects models. *Statistics in Medicine*, 27, 3105–3124.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall/CRC.

McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89, 330–335.

McLachlan, G. and Peel, D.A. (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Milanzi, E., Molenberghs, G., Alonso, A., Verbeke, G. and De Boeck, P. (2015). Reliability measures in item response theory: Manifest versus latent correlation functions. *British Journal of Mathematical and Statistical Psychology*, 68, 43–64.

Molenberghs, G., Kenward, M.G., Verbeke, G., Efendi, A. and Iddi, S. (2013). On the connections between bridge distributions, marginalized multilevel models, and generalized linear mixed models. *International Journal of Statistics and Probability*, 2, 1–21.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician*, 61, 1–6.

Molenberghs, G. and Verbeke, G. (2011a). On the Weibull-Gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, Cauchy, and extreme-value distributions. *Journal of Statistical Planning and Inference*, 141, 861–868.

Molenberghs, G. and Verbeke, G. (2011b). A note on a hierarchical interpretation for negative variance components. *Statistical Modelling*, 11, 389–408.

Molenberghs, G., Verbeke, G. and Demétrio, C. (2007). An extended random-effects approach to modelling repeated, overdispersed count data. *Lifetime Data Analysis*, 13, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B. and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, 25, 325–347.

Molenberghs, G., Verbeke, G., Efendi, A., Braekers, R. and Demétrio, C.G.B. (2015). A combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event data. *Statistical Methods in Medical Research*, 24, 434–452.

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

Moore, D.F. and Tsiatis, A.A. (1991). Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics*, 47, 383–401.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–65.

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

Nelson, K.P., Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J., Parzen, M. and Strawderman, R. (2006). Use of the probability integral transformation to fit nonlinear mixed-effects models with non-normal random effects. *Journal of Computational and Graphical Statistics*, 15, 39–57.

Neyens, T., Faes, C., and Molenberghs, G. (2012). A generalized Poisson-gamma model for spatially overdispersed data. *Spatial and Spatio-temporal Epidemiology*, 3, 185–194.

Njeru Njagi, E., Molenberghs, G., Rizopoulos, D., Verbeke, G., Kenward, M.G., Dendale, P. and Willekens, K. (2016). A flexible joint-modelling framework for longitudinal and time-to-event data with overdispersion. *Statistical Methods in Medical Research*, 25, 1661–1676.

Oliveira, I.R.C., Molenberghs, G., Demétrio, C.G.B., Giolo, S. and Dias, C.T.S. (2016). Quantifying intra-class correlations for nonnegative traits. *Biometrical Journal*, 58, 852–867.

Oliveira, I.R.C., Molenberghs, G., Verbeke, G., Demétrio, C.G.B. and Dias, C.T.S. (2017). Negative variance components for non-negative hierarchical data with correlation, over-, and/or underdispersion. *Journal of Applied Statistics*, 44, 1047–1063.

Ouwens, M.J.N.M., Tan, F.E.S. and Berger, M.P.F. (2001). Local influence to detect influential data structures for generalized linear mixed models. *Biometrics*, 57, 1166–1172.

Pryseley, A., Tchonlafi, C., Verbeke, G. and Molenberghs, G. (2011). Estimating negative variance components from Gaussian and non-Gaussian data: a mixed models approach. *Computational Statistics and Data Analysis*, 55, 1071–1085.

Rakhmawati, T., Molenberghs, G., Verbeke, G. and Faes, C. (2016a). Local influence diagnostics for hierarchical count data models with overdispersion and excess zeros. *Biometrical Journal*, 58, 1390–1408.

Rakhmawati, T., Molenberghs, G., Verbeke, G. and Faes, C. (2016b). Local influence diagnostics for incomplete overdispersed longitudinal counts. *Journal of Applied Statistics*, 43, 1722–1737.

Rakhmawati, T., Molenberghs, G., Verbeke, G. and Faes, C. (2017). Local influence diagnostics for generalized linear mixed models with overdispersion. *Journal of Applied Statistics*, 44, 620–641.

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, 44, 649–667.

Ridout, M., Demétrio, C.G.B. and Hinde, J. (1998). Models for count data with many zeros. In: *International Biometric Conference XIX*, Cape Town. Invited Papers, pp. 179–192.

Rinne, H. (2009). *The Weibull Distribution. A Handbook.* Boca Raton: CRC/Chapman & Hall.

Rizzato, F.B., Leandro, R.A., Demétrio, C.G.B. and Molenberghs, G. (2016). A Bayesian approach to analyse overdispersed longitudinal count data. *Journal of Applied Statistics*, 43, 2085–2109.

Roberts, D.T. (1992). Prevalence of dermatophyte onychomycosis in the United Kingdom: Results of an omnibus survey. *British Journal of Dermatology*, 126, 23–27.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719–729.

Shepard, T.H., Mackler, B. and Finch, C.A. (1980). Reproductive studies in the iron-deficient rat. *Teratology*, 22, 329–334.

Shoukri, M.M., Mian, I.U.M. and Tracy, D.S. (1988). Sampling properties of estimators of the log-logistic distribution with application to Canadian precipitation data. *Canadian Journal of Statistics*, 16, 223–236.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials, *Journal of the Royal Statistical Society, Series B*, 10, 257–261.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling*. London: Chapman & Hall/CRC.

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46, 657–671.

Vangeneugden, T., Molenberghs, G., Laenen, A., Alonso, A. and Geys, H. (2008). Generalizability in non-Gaussian longitudinal clinical trial data based on generalized linear mixed models. *Journal of Biopharmaceutical Statistics*, 18, 691–712.

Vangeneugden, T., Molenberghs, G., Laenen, A., Geys, H., Beunckens, C. and Sotto, C. (2010). Marginal correlation in longitudinal binary data based on generalized linear mixed models. *Communications in Statistics, Theory & Methods*, 39, 3540–3557.

Vangeneugden, T., Molenberghs, G., Verbeke, G. and Demétrio, C. (2011). Marginal correlation from an extended random-effects model for repeated and overdispersed counts. *Journal of Applied Statistics*, 38, 215–232.

Vangeneugden, T., Molenberghs, G., Verbeke, G. and Demétrio, C.G.B. (2014). Marginal correlation from logit- and probit-beta-normal models for hierarchical binary data. *Communications in Statistics*, 43, 4164–4178.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

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

Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049–1060.

SORT 41 (2) July-December 2017, 255-276 DOI: 10.2436/20.8080.02.59

A bivariate response model for studying the marks obtained in two jointly-dependent modules in higher education

Emilio Gómez-Déniz[1], Nancy Dávila Cárdenes[1] and María D. García Artiles[2]

Abstract

We study the factors which may affect students' marks in two modules, mathematics and statistics, taught consecutively in the first year of a Business Administration Studies degree course. For this purpose, we introduce a suitable bivariate regression model in which the dependent variables have bounded support and the marginal means are functions of explanatory variables. The marginal probability density functions have a classical beta distribution. Simulation experiments were performed to observe the behaviour of the maximum likelihood estimators. Comparisons with univariate beta regression models show the proposed bivariate regression model to be superior.

MSC: 62P25, 62E99, 97D60.

Keywords: Beta distribution, bivariate beta distribution, conditional distributions, covariate, marginal distributions, regression, mathematics, statistics, business studies.

1. Introduction

Event counts such as the number of claims for third-party liability, other claims under guarantee, medical consultations, the use of prescription drugs, and voluntary and/or involuntary job changes, among many others, are likely to be jointly dependent. In these cases, it is of interest to study how different covariates or factors may simultaneously affect the two random (dependent) variables involved. Bivariate Poisson regression models, bivariate negative binomial regression models (see Maher, 1990) and their extensions (see Gurmu and Elder, 2000), among other approaches, have been applied in these settings. Nevertheless, few such studies have been conducted when the dependent variables are continuous and bounded.

[1] Department of Quantitative Methods and TiDES Institute. University of Las Palmas de Gran Canaria, Spain. emilio.gomez-deniz@ulpgc.es, nancy.davila@ulpgc.es

[2] Department of Quantitative Methods. University of Las Palmas de Gran Canaria, Spain. mdgartiles@ulpgc.es

Received: March 2016
Accepted: March 2017

In the univariate case, Papke and Wooldridge (1996) examined potential econometric alternatives when the dependent variable is fractional, in a study of employee participation rates in 401(k) pension plans. More recently, Papke and Wooldridge (2008) analysed test pass rates and the portfolio choices of Australian households. Other research work related to the beta regression model includes Cepeda-Cuervo (2001), Paolino (2001), Ferrari and Cribari-Neto (2004) and Huang and Oosterlee (2011). The model proposed by Gómez-Déniz, Sordo and Calderín-Ojeda (2013) provides an alternative to the beta regression model, and affords a better fit, at least in an actuarial setting. The model proposed by Pérez-Rodríguez and Gómez-Déniz (2015) also appears to be comparable to the beta regression approach in financial econometrics. Using Bayesian methodology, Bayes, Bazán and García (2012) presented a variation of the beta regression model, while Cepeda-Cuervo and Núñez-Antón (2013), used spatial regression in an analysis of the quality of education. In the bivariate case, Cepeda-Cuervo, Achcarb and Garrido (2014) proposed a bivariate beta regression model with joint modelling of the mean and dispersion parameters.

As an extension of the works related above, we propose a flexible bivariate fractional response model in which the dependent variables are bounded and the marginal means are functions of explanatory variables.

Although a bivariate regression model could be built by using, for instance, copulas from the Sarmanov family of distributions (see Lee (1996)), we chose the bivariate beta regression proposed by Olkin and Liu (2003) for this study because it is a simple model with which to compute marginal distributions, means and variances. In this model, the beta distribution has a straightforward formulation in which the Euler Gamma function is the only one considered. In this respect, Cepeda-Cuervo et al. (2014) used copulas to obtain a bivariate beta regression model in which, as in our own model, the marginal distributions are beta. These authors assumed weak dependence between the variables of interest and modelled the dependence using a Farlie-Gumbel-Morgenstern copula function.

The model we propose is less complex than that presented in Cepeda-Cuervo et al. (2014) and therefore, by the Ockham's razor principle, it might be preferable (Jaynes, 1994).

In this paper, we study how some covariates may simultaneously affect the marks (ranging from 0 to 10) obtained by students in two first-degree subjects – Mathematics for Business and Basic Statistics in Business Administration Studies – taught at the University of Las Palmas de Gran Canaria (Spain) during two consecutive terms (first mathematics and then statistics). We assume that a good knowledge of mathematics will significantly influence the student's understanding of statistics and therefore that there exists a positive correlation between these two variables. Accordingly, the model proposed would be suitable for studying this relationship.

The importance of mathematical skills in other quantitative disciplines has been widely examined. In the fields of business and economics, many studies have analysed basic mathematical abilities as determinant factors of academic performance among first year university students: see, for example, Johnson and Kuennen (2006), Dolado and Morales (2009), Lunsford and Poplin (2011) and Arnold and Straten (2012).

Study plans in business and economics courses are organised in different ways, depending on the institution, but all have in common a requirement of basic mathematics to favour a better understanding of subjects that require this skill as a tool to develop more complex theories. In the present study, we focus on the above-mentioned mathematics and statistics modules, to determine whether certain common factors might explain the students' marks obtained in each subject.

The rest of this paper is structured as follows. Section 2 describes the bivariate model proposed by Olkin and Liu (2003), from which we derive the proposed bivariate regression model. This model and its parameters are studied in Section 3. The data are described in detail in Section 4. In Section 5 we fit the marginal beta regression models and the bivariate beta regression models, comparing the univariate and the bivariate models. Finally the results obtained and the main conclusions drawn are reported in Section 6.

2. Modelling bivariate marks

Although mathematics and statistics are known to be logical and effective means of solving certain problems, most Business Administration students, especially those in the first and second years of their degree courses, are not interested in these course subjects. Indeed, numerous students in this area of study present some form of rejection of mathematics and statistics. Nevertheless, our empirical evidence shows that the marks obtained by students in statistics are positively related to those achieved in mathematics. We assume this is because mathematics is an instrumental subject that influences the results achieved in statistics.

Let Y_1 and Y_2 be two random variables which represent the marks achieved in mathematics and statistics, respectively. To address the study goal presented in the introduction, and taking into account the above comments, we need a bivariate distribution that meets the following conditions:

a) The support of the distribution should be bounded, since the marks are usually restricted to a given interval.
b) The bivariate distribution should provide a dependence structure.
c) The correlation between the two random variables should be positive. That is, $\rho(Y_1, Y_2) > 0$.
d) Preferably, $\Pr(Y_2 > y_2 | Y_1 > y_1)$ should be a nondecreasing function in y_1 for all y_2. Thus, the higher the mathematics mark, the greater the probability of obtaining higher marks in statistics.
e) Because we wish to study the factors which may affect the marks obtained in the two courses, using a regression analysis, the marginal mean (the response variable) should be expressed as a function of the explanatory variables through a simple expression.

In this case, the standard beta distribution may be extended to the bivariate case. Many bivariate beta distributions have been derived from an application or as extensions to or generalisations of other well-known bivariate beta distributions. Since the latter are used in a wide variety of applications, the development and derivation of new bivariate beta distributions has been extensively studied. Nevertheless, few such distributions present these five features simultaneously. One, however, was proposed by Olkin and Liu (2003), with the following probability density function (pdf):

$$f(y_1,y_2)=\frac{y_1^{a_1-1}y_2^{a_2-1}(1-y_1)^{a_2+a_3-1}(1-y_2)^{a_1+a_3-1}}{B(a_1,a_2,a_3)(1-y_1y_2)^{a_1+a_2+a_3}}, \tag{1}$$

where $0<y_i<1$ $(i=1,2)$, $a_i>0$ $(i=1,2,3)$ and where $B(a_1,a_2,a_3)$ is given by $B(a_1,a_2,a_3)=\prod_{i=1}^{3}\Gamma(a_i)/\Gamma(\sum_{i=1}^{3}a_i)$, where $\Gamma(\cdot)$ is the Euler Gamma function. Henceforth, we use the expression $(Y_1,Y_2)\sim\mathscr{BB}(a_1,a_2,a_3)$ when the two random variables (Y_1,Y_2) fit the pdf (1).

The marginal distributions of Y_1 and Y_2 are beta distributions with parameters (a_1,a_3) and (a_2,a_3), respectively. Thus, the marginal means, the variances and the cross moment are given by

$$E(Y_i)=\frac{a_i}{a_i+a_3},\quad i=1,2, \tag{2}$$

$$var(Y_i)=\frac{a_ia_3}{(a_i+a_3)^2(a_i+a_3+1)},\quad i=1,2.$$

$$E(Y_1Y_2)=\frac{a_1a_2\Gamma(a_1+a_3)\Gamma(a_2+a_3)}{m\Gamma(a_3)\Gamma(m+1)}\,{}_3F_2(\{m_1,m_2,m\};\{m+1,m+1\};1), \tag{3}$$

where $m_i=a_i+1$ $(i=1,2)$, $m=a_1+a_2+a_3$ and ${}_3F_2$ is the generalised hypergeometric function. For details about this special function see, for instance, Gottschalk and Maslen (1988). This can be computed using the Mathematica package (see Wolfram (2003)). Using (2) and (3) we can obtain the covariance, $cov(Y_1,Y_2)$, and the correlation between Y_1 and Y_2, $\rho(Y_1,Y_2)$. For reasons of space, these large expressions are not shown here. Olkin and Liu (2003) showed that the correlation is always positive, with values in the interval $(0,1)$. The following result is obtained for the conditional distribution:

$$f(y_1|y_2)=\frac{y_1^{a_1-1}(1-y_1)^{a_2+a_3-1}(1-y_2)^{a_1}}{B(a_1,a_2+a_3)(1-y_1y_2)^{a_1+a_2+a_3}}, \tag{4}$$

$$f(y_2|y_1)=\frac{y_2^{a_2-1}(1-y_2)^{a_1+a_3-1}(1-y_1)^{a_2}}{B(a_2,a_1+a_3)(1-y_1y_2)^{a_1+a_2+a_3}}. \tag{5}$$

After some algebra, we derive the conditional mean obtained from (4) and (5). Thus

$$E(Y_1|Y_2=y_2)=\frac{a_1}{a_1+a_2+a_3}\,{}_2F_1(1,a_2+a_3;a_1+a_2+a_3;y_2),$$
$$E(Y_2|Y_1=y_1)=\frac{a_2}{a_1+a_2+a_3}\,{}_2F_1(1,a_1+a_3;a_1+a_2+a_3;y_1), \tag{6}$$

where ${}_2F_1$ represents the hypergeometric function (see Gradshteyn and Ryzhik, 1994).

One of the advantages of using the pdf given in (1) is that for this distribution we have

$$\begin{aligned}\Pr(Y_2>y_2|Y_1>y_1^0) &\leq \Pr(Y_2>y_2|Y_1>y_1^1), \quad y_1^0<y_1^1,\\ \Pr(Y_2\leq y_2|Y_1\leq y_1^0) &\geq \Pr(Y_2\leq y_2|Y_1\leq y_1^1), \quad y_1^0<y_1^1,\end{aligned}$$

for all y_2. In other words, $\Pr(Y_2>y_2|Y_1>y_1)$ is a nondecreasing function in y_1 for all y_2 and $\Pr(Y_2\leq y_2|Y_1\leq y_1)$ is a nonincreasing function in y_1 for all y_2, because the pdf (1) is positively likelihood ratio dependent (see Tong (1980) and Olkin and Liu (2003) for details). This is corroborated by the fact that in our case the random variables Y_1 and Y_2 are positively quadrant dependent, a concept introduced by Lehmann (1996). Thus, we have

$$\begin{aligned}\Pr(Y_2>y_2|Y_1>y_1) &\geq \Pr(Y_2>y_2)\Pr(Y_1>y_1),\\ \Pr(Y_2\leq y_2|Y_1\leq y_1) &\geq \Pr(Y_2\leq y_2)\Pr(Y_1\leq y_1).\end{aligned}$$

A possible interpretation of the parameters of the distribution in (1) is this. Let W be a random variable measuring a student's lack of mathematics skills for use in subjects such as mathematics, statistics and physics. Empirical evidence shows that when Business Administration students are asked about their skills in mathematics and statistics, most of them acknowledge inadequacy in this field. Let U_i $(i=1,2)$ be the random variable representing the student's willingness to study these subjects i $(i=1,2)$. Assuming that W and U_i can take values in $(0,\infty)$, then the marks obtained in subject i can be represented by the random variables

$$Y_i=\frac{1}{1+W/U_i}=\frac{U_i}{U_i+W}, \quad i=1,2.$$

The gamma distribution provides a flexible representation of a variety of distribution shapes, by varying the shape parameter. Let us now assume that the random variables U_1, U_2 and W are independent and follow a standard gamma distribution with shape parameters a_1, a_2 and a_3, respectively. Then, the random variable (Y_1,Y_2) follows the distribution given in (1).

In conclusion, the pdf given in (1) seems to be a suitable distribution to model the joint random variables corresponding to mathematics and statistics marks when the latter are influenced by the former.

3. Regression model and estimation

Let us now consider a more realistic model, in which covariates are included. The linear regression model, which makes no distributional assumptions, is likely to be unsatisfactory because certain combinations of parameters and regressors could violate the nonnegative restriction and the upper limit on the mean. To avoid this situation we propose a parametric model based on using the distributional assumptions presented in the previous section.

When a regression analysis is to be performed, it is often useful to model the mean of the response. By equating the mean given in (2) to μ_i $(i=1,2)$, solving for a_i $(i=1,2)$, taking $a_3=\theta$ and replacing the resulting expression in the pdf of the bivariate beta distribution in (1), we obtain the following reparametrisation.

$$f(y_1,y_2)=\frac{y_1^{\phi_1\mu_1-1}y_2^{\phi_2\mu_2-1}(1-y_1)^{\phi_2-1}(1-y_2)^{\phi_1-1}}{B(\phi_1\mu_1,\phi_2\mu_2,\theta)\,(1-y_1y_2)^{(1-\mu_1\mu_2)\phi_1\phi_2/\theta}}, \tag{7}$$

where $\phi_i=\theta/(1-\mu_i)$, $0<\mu_i<1$, $i=1,2$; with $0<y_1<1$, $0<y_2<1$ and $\theta>0$. Under this reparametrisation of the bivariate beta distribution, the marginal mean is $E(Y_i)=\mu_i$, for $i=1,2$.

Now, let $\boldsymbol{x}_{\kappa i}^{\mathsf{T}}=(x_{1i},x_{2i},\ldots,x_{pi})$ be a vector of the p covariates associated with the ith observation. This is a vector of linearly independent regressors that are thought to determine (y_1,y_2). For the ith observation, the model takes the form

$$(Y_{1i},Y_{2i})\sim\mathscr{BB}(\mu_{1i},\mu_{2i},\theta),$$

$$\mu_{\kappa i}(\boldsymbol{x}_{\kappa i},\boldsymbol{\beta}_\kappa)\equiv\mu_{\kappa i}=\frac{\exp(\boldsymbol{x}_{\kappa i}^{\mathsf{T}}\boldsymbol{\beta}_\kappa)}{1+\exp(\boldsymbol{x}_{\kappa i}^{\mathsf{T}}\boldsymbol{\beta}_\kappa)},\quad \kappa=1,2.$$

Here, $i=1,\ldots,n$ denotes the number of observations, $\boldsymbol{x}_{\kappa i}$ denotes a vector of p explanatory variables for the ith observation and $\boldsymbol{\beta}_\kappa=(\beta_{\kappa 1},\ldots,\beta_{\kappa p})^{\mathsf{T}}$, $\kappa=1,2$, denotes the corresponding vectors of regression coefficients. It is clear that each variable Y_1 and Y_2 may be influenced by different characteristics and variables. For this reason, the explanatory variables that are used to model each mean $\mu_{\kappa i}$, may not be the same. Furthermore, observe that the logit link assumed ensures that $\mu_{\kappa i}$ falls within the interval $(0,1)$.

Under this model the log-likelihood function takes the form given in the Appendix, which shows the equations used to provide the estimates of the parameters. The above model presents the advantage of simplicity; on the other hand, the normal equations require the use of the digamma function, $\psi(z) = \frac{d}{dz}\log(\Gamma(z))$, $z > 0$, in order to estimate all the model parameters. However, this problem is overcome by means of Mathematica routines (see Wolfram, 2003) and RATS (see Brooks, 2009), which work well with this special function.

Because the equations which provide the estimates of the parameters cannot be solved explicitly, they must be addressed either by numerical methods or by directly maximising the log-likelihood function; in this study, the latter approach is adopted. Since the global maximum of the log-likelihood surface is not guaranteed, different initial values of the parametric space can be considered as seed points. In this sense, we have used the `FindMaximum` function of the Mathematica software package v.11.0 (Wolfram, 2003). Moreover, other methods provided by Mathematica, such as Newton, PrincipalAxis and QuasiNewton (all of which are available in Mathematica) obtain the same result. Finally, the standard errors of the estimated parameter are approximated by inverting the Hessian matrix. This can also be done by approximating the Hessian matrix and recovering it from the Cholesky factors. These parameters were also computed by the RATS package, and the same values were obtained.

3.1. Marginal effects

The marginal effect reflects the variation of the conditional mean produced by a one-unit change in the jth covariate ($j = 1,\ldots,p$). The marginal effect can be calculated as $\delta_j = \frac{\partial \mu_{\kappa i}}{\partial x_{ji}} = \beta_{\kappa j}\mu_{\kappa i}(1-\mu_{\kappa i})$, $\kappa = 1,2$; $i = 1,\ldots,n$; $j = 1,\ldots,p$. Thus, the marginal effect indicates that a one-unit change in the jth regressor increases or decreases the expectation of marks for the jth covariate by δ_j units, $j = 1,\ldots,p$. This expression is the one normally obtained under the logit marginal effect. For indicator variables which takes only the values 0 or 1 the marginal effect is $\delta_j = E(y_\kappa|x_{ji} = 1)/E(y_\kappa|x_{ji} = 0) \approx \exp(\beta_{\kappa j})$, $\kappa = 1,2$; $i = 1,\ldots,n$; $j = 1,\ldots,p$. Therefore, the conditional mean is $\exp(\beta_{\kappa j})$ times larger if the indicator variable is one rather than zero.

3.2. Simulation study

We now present some simulation results, obtained by a bootstrap experiment, to study the behaviour of the maximum likelihood estimators. The Mathematica package was used to create random variables from the pdf (7). In this process, the first component of the vector was generated from a marginal, and then a second one from a conditional distribution. The estimated values of the parameters were then computed directly using the `FindMaximum` function of Mathematica v.11.0 (Wolfram (2003)). The following sets of model parameters were considered:

Table 1: *Average estimates (first row), the square root of the mean squared errors (second row in parenthesis) and the correlation (ρ) between estimated parameters based on 1000 replications.*

n	$\mu_1 = 0.15$	$\mu_2 = 0.25$	$\theta = 0.85$	$\rho(\mu_1, \mu_2)$	$\rho(\mu_1, \theta)$	$\rho(\mu_2, \theta)$
25	0.1402	0.2213	0.9827	0.7524	–0.7481	–0.6804
	(0.0367)	(0.0585)	(0.4490)			
50	0.1932	0.3004	0.7256	0.6425	–0.8068	–0.7392
	(0.0303)	(0.0438)	(0.1364)			
75	0.1795	0.2537	0.7326	0.6009	–0.7757	–0.7566
	(0.0246)	(0.0310)	(0.1116)			
100	0.1526	0.2820	0.7292	0.4851	–0.7637	–0.5173
	(0.0167)	(0.0269)	(0.0842)			
	$\mu_1 = 0.25$	$\mu_2 = 0.75$	$\theta = 0.5$	$\rho(\mu_1, \mu_2)$	$\rho(\mu_1, \theta)$	$\rho(\mu_2, \theta)$
25	0.2308	0.7273	0.6525	0.4976	–0.7454	–0.5647
	(0.0558)	(0.0347)	(0.1829)			
50	0.2542	0.7529	0.4906	0.5719	–0.7883	–0.5647
	(0.0375)	(0.0323)	(0.0998)			
75	0.2395	0.7190	0.5886	0.4495	–0.7168	–0.3814
	(0.0326)	(0.0274)	(0.0715)			
100	0.2992	0.7943	0.4170	0.4116	–0.7040	–0.3814
	(0.0298)	(0.0180)	(0.0438)			
	$\mu_1 = 0.50$	$\mu_2 = 0.15$	$\theta = 1.50$	$\rho(\mu_1, \mu_2)$	$\rho(\mu_1, \theta)$	$\rho(\mu_2, \theta)$
25	0.5210	0.1273	1.9765	0.1815	0.0217	–0.7145
	(0.0484)	(0.0268)	(0.4869)			
50	0.5713	0.1603	1.4338	0.4822	–0.4643	–0.5710
	(0.0332)	(0.0249)	(0.1888)			
75	0.5289	0.1600	1.3432	0.1993	–0.4655	–0.5616
	(0.0293)	(0.0199)	(0.1618)			
100	0.5063	0.1851	1.5894	0.5823	–0.6167	–0.6597
	(0.0240)	(0.0198)	(0.2168)			

$$(\mu_1, \mu_2, \theta) = (0.15, 0.25, 0.85),$$
$$(\mu_1, \mu_2, \theta) = (0.25, 0.75, 0.50),$$
$$(\mu_1, \mu_2, \theta) = (0.50, 0.15, 1.50).$$

In all three cases, we have simulated observations with a sample size given by $n = 25, 50, 75$ and 100. We report the average estimates and the square root of the mean squared errors based on 1000 replications, i.e. the bootstrap sample is taken from the original by using sampling with replacement 1000 times. Additional replications are considered unnecessary, as the computational time needed would be prohibitive; nevertheless, we acknowledge that the use of fewer replications might reduce the statistical accuracy obtained. The results are shown in Table 1. In general, as the sample size increases the estimates approach the true values and the biases and the mean squared er-

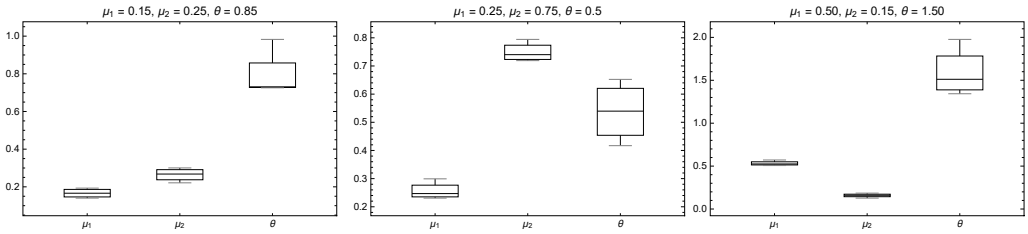

Figure 1: *Box-and-whisker charts showing the differences between the true parameter values and the estimates based on the data in Table 1.*

rors decrease. These outcomes corroborate the consistency of the maximum likelihood estimates. From the standard errors obtained, it is evident that the errors are smaller as the sample size increases. Furthermore, the correlation between the parameters is always positive for μ_1 and μ_2 and negative for μ_2 and θ. Hence, the correlation between these two sets of parameters is not very high. Finally, Figure 1 shows that the parameters estimated have a slight negative bias, which is more apparent in the θ parameter.

4. Factors affecting the mathematics and statistics marks obtained

In order to make use of the bivariate regression model, we examined the relation between the marks achieved by the students in two course subjects: Mathematics and Statistics in Business Administration. Most of these students, before entering the university had studied subjects focused on statistics, more so than basic mathematics. In fact, many of them believed they did not need mathematics and did not consider the two courses to be related. During the first term, difficulties were encountered in mathematics, but with the start of the statistics class, in the second term, the students believed their performance would be better. Therefore, at the beginning of the mathematics course, the students were informed of the analysis that would be conducted, and were asked to complete a questionnaire on this subject. The following section describes how the data were compiled and how many students comprised the final study sample.

4.1. The sample

The data for this study were collected from eight student groups in the Mathematics for Business and Statistics modules taught during the first year of the Business Administration degree course at the University of Las Palmas de Gran Canaria (Spain). The study population was initially composed of 725 students enrolled in these groups. On the first day of classes in 2013, a questionnaire was handed out to 456 students.

The final sample was composed of the 213 students who completed both modules (mathematics and statistics) and answered the survey. The questionnaire was divided

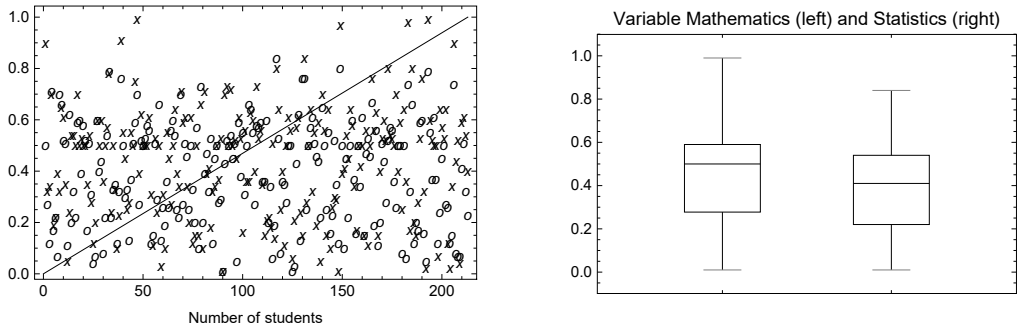

Figure 2: *Scatter plot of the marks in mathematics (×) and statistics (○) on the left and box-and-whisker charts for the two dependent variables on the right.*

into two parts; the first contained questions dealing with personal and academic information, and the second presented four short mathematics exercises.

During the academic year, the students are given three opportunities to take the class exams. In total, 114 students (54%) of the students who completed the initial questionnaire passed the Mathematics for Business exam, and 92 (43%) passed in Basic Statistics. The final marks for the students in the sample, for each of these two subjects, are shown in the scatter plot in Figure 2. Few students obtained high marks in mathematics, and there was a large concentration of values below 0.6. Figure 2 also shows box-and-whisker charts for the two dependent variables. Since the support of (1) does not include the values zero or one, and taking into account that the data contained very few such marks, instead of removing them, these marks were replaced by 0.001 and 0.999, respectively.

The range of possible exam marks was from 0 to 10. In Spain, a pass mark is 5 or more. In the study sample, most of the marks obtained were between 4.5 and 7.0. Many were under 4.0 and very few were over 7.0.

4.2. Personal and academic factors

The survey data collected concerning personal and academic factors are shown in Table 2. Among the personal and academic information sought in the questionnaire, the variable AGE was obtained by dividing the students into those born in 1995 and those born earlier. The year 1995 was taken because in 2013, when the study data were compiled, these students would be aged 18 years, which is the usual age for university entrance.

To accompany the continuous variable AGE, a box-and-whisker chart is included in Figure 3. This chart shows that most of the students in the study sample were aged 18-21 years. The outliers in the sample were aged 27-32 years. These students had entered the university at age 25 years or older, and had had to pass specific examinations to do so.

Table 2: *Descriptive data: personal and academic information.*

Personal and academic information		% Respondents N = 213	% Mathematics Passes N = 114	% Statistics Passes N = 92
AGE	Born in 1995	47	43	40
	Born before 1995	53	57	60
SCHOOL	Public	82	80	84
	Private	18	20	16
TRACK	Technical-Science	12	14	12
	Other	88	86	88
ADMSCORE	[5, 8]	54	45	49
	(8, 14]	46	55	51
PREF	Yes	90	94	92
Business	No	10	6	8
	Yes	71	68	64
NEWCOMER	No	29	32	36
GRANT	Yes	65	63	62
	No	35	37	38
WORKING	Yes	4	4	8
	No	96	96	92
GENDER	Male	48	51	51
	Female	52	49	49

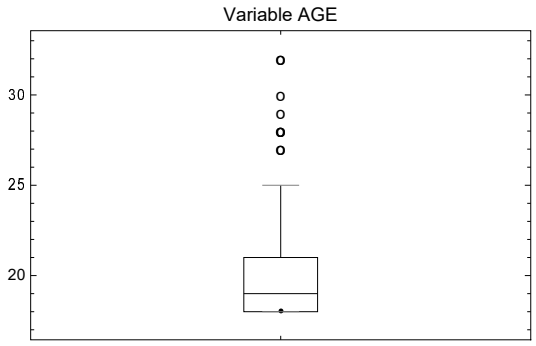

Figure 3: *Box-and-whisker chart of the continuous variable AGE.*

Other information requested concerned the type of school (public or private) attended before university entrance. This variable was termed SCHOOL, and the academic specialisation chosen by the student during the last two years of high school was termed TRACK. Different types of track are available, but for the purposes of this study, they were divided into Technical-Science and Others.

In Spain, university entrance requires a specific examination, known as PAU (Prueba de Acceso a la Universidad) to be taken, in addition to the final high-school exams.

The weighted average of the latter mark and the PAU result determines the final admission score obtained; we term this variable ADMSCORE. The threshold for university entrance ranges from 5 to 14, depending on demand for the course and on the places available. The students in our sample were divided into those who obtained an admission score of 8 or less, and those who obtained 9 or more. After taking the PAU exam, a period is allowed during which students may choose the degree course they wish to study. For the purposes of this research, to identify the strength of vocation in the students' choice, the questionnaire asked whether the degree in Business was their first preference. The study variable in this respect was termed PREF.

Every year, 400 new students enrol in the Mathematics for Business module. However, the total number of students enrolled each year is almost double this figure due to the high number of students who failed to pass or did not sit the previous year's exam, and who had to retake the course. To distinguish new students from those retaking the course, this information was requested, and the corresponding variable was termed NEWCOMER. Other variables included were the students' gender (GENDER), whether they were receiving a study grant (GRANT) and whether they were working (WORKING).

4.3. Factors related to mathematics skills

The variables concerning the students' mathematics skills are shown in Table 3. As part of the questionnaire, the students were asked to solve four exercises and to describe how they had done so. Different steps were involved in each exercise. The score awarded for

Table 3: *Descriptive data for mathematics skills.*

BASIC MATHEMATICS SKILLS	**Variables**	**Exercices**	**% Respondents**	**% Pass rate**
1. LINEAR EQUATION				
Handle rational coefficients	FRACTIONS	$\frac{1}{2}x+\frac{3}{4}x=0$	46	55
Solving the equation	LINEAREQ		33	38
2. EQUATIONS SYSTEM				
Resolution system method	SYSTEM		64	73
Raising the quadratic equation	EQ2	$\left.\begin{array}{r} -x^2+2x-3y=0, \\ 5x+3y=0. \end{array}\right\}$	48	52
Solving the quadratic equation	SOLVINGEQ2		22	21
Discuss the solutions	DISCUSSEQ2		10	12
3. ALGEBRAIC EXPRESSIONS				
Clear the unknown	CLEANUNKN	$2x^2y^3-ax^3y=0$	24	31
Simplify exponents	SIMPLIFYEXP		17	25
4. BASIC DERIVATIVES				
Polynomial with integer exponent	DERINTEXP		47	53
Polynomial with rational exponent	DERRATEXP	$f(x)=\frac{1}{3}x^3-x^{1/2}+2$	23	34
Simplify the final expression	SIMPLYFYDER		11	18

each step in the exercise was 1 when it was performed correctly, or 0 otherwise. The following variables were associated with each step in the procedure.

The first exercise was a linear equation with rational coefficients. The question was evaluated according to whether the student was able to handle the basic algebra involved in manipulating the coefficients. The study variable in this respect was termed FRACTIONS. Solving the equation and obtaining the value of the unknown parameter were represented by the term LINEAREQ.

The second question concerned a very basic non-linear equation system. The marking criteria were defined using the following variables: the application of a method for solving linear systems, SYSTEM; giving the resulting incomplete quadratic equation, EQ2; solving the equation and discussing the solution obtained, SOLVINGEQ2 and DISCUSSEQ2, respectively.

The third exercise consisted in giving the value of the parameter "a" after simplifying the algebraic expression (see Table 3). To achieve a positive score, the students had to clear the unknown, CLEANUNKN, and simplify the exponents, SIMPLIFYEXP. In the final exercise, the students were asked to calculate the derivative function of an elementary polynomial expression including an integer and rational exponents. The exercise was evaluated according to whether the integer exponent was correctly derived, DERINTEXP, whether the derivative of the rational expression was correctly given, DERRATEXP, and whether the last expression was correctly simplified, SIMPLYFYDER.

Observation of the final column in Table 3, the percentage of students who passed the final subject exams, clearly shows that although the percentage of passing students is higher among those who correctly responded to the questions in the initial survey, the pass rates are still unacceptably low. The results obtained reflect a lack of basic skills in some areas of mathematics.

5. Testing the models

The descriptive values obtained for the dependent variables are given by $E(Y_1) = 0.44169$, $var(Y_1) = 0.04728$, $E(Y_2) = 0.39145$ and $var(Y_2) = 0.04000$. Thus, the mean value of Y_1 is larger than that of Y_2 while the variance is similar in each case. The correlation is positive, with a value of 0.67243 indicating that these values increase or decrease together.

5.1. Model without covariates

The model was initially implemented without covariates, which produced the parameter values shown in Table 4. The standard errors are shown between parentheses, and the estimates obtained when the univariate beta distribution is assumed are also shown. In addition, we show the value obtained for the Akaike information criterion (AIC).

Table 4: *Univariate and bivariate models without covariates.*

	Y_1 : Maths	Y_2 : Statistics	(Y_1, Y_2)
$\widehat{\mu}_1$	0.441821		0.438451
	(0.015631)		(0.014354)
$\widehat{\mu}_2$		0.384893	0.388744
		(0.013748)	(0.014500)
$\widehat{\theta}$	3.561590	4.745440	2.497550
	(0.307957)	(0.422396)	(0.179619)
AIC	–40.832	–93.8506	–203.260

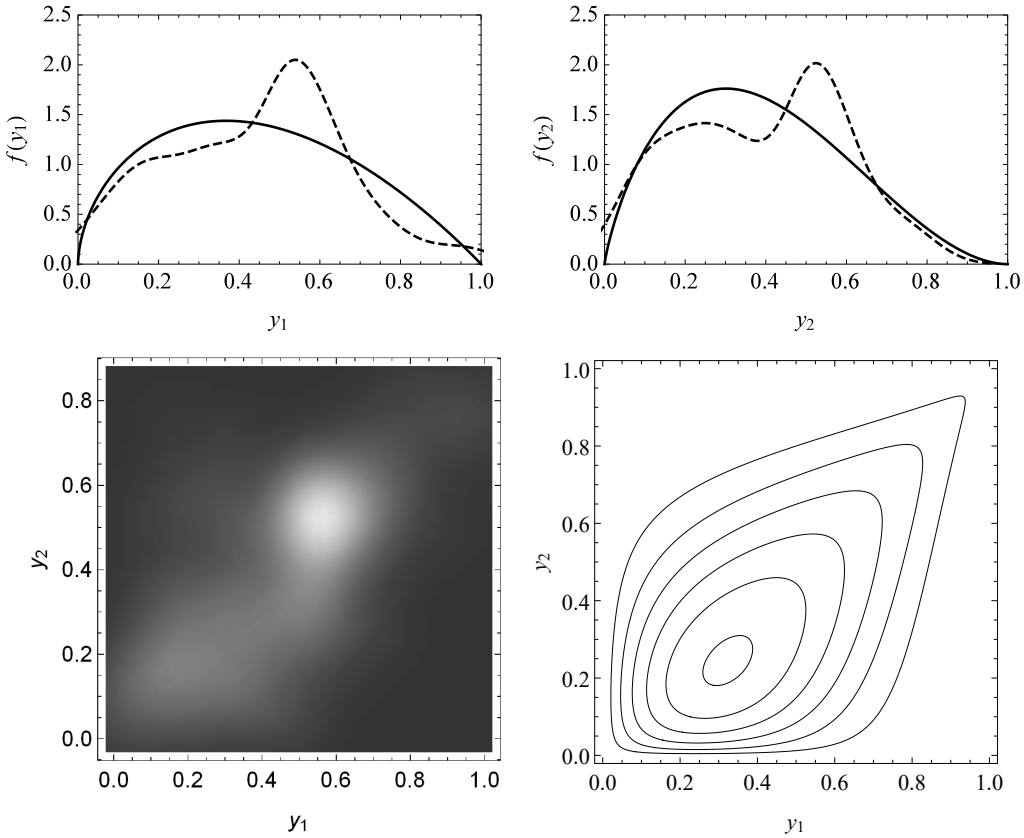

Figure 4: *Top: the smooth kernel densities (dashed curves) and the pdfs (solid curves) of the estimated univariate beta distribution. Bottom: the smooth contour plot obtained from the data (left) and the estimated contour plots of the bivariate beta pdf.*

(AIC $= 2(k - \ell_{\max})$, where k is the number of model parameters and $\ell_{\max}$ is the maximum value of the log-likelihood function; see Akaike (1974) for details.) These values should be compared only with the fitted models obtained when covariates are included, as described below.

From the parameter values obtained by the model without covariates, we obtain the estimated descriptive statistics $\widehat{E(Y_1)} = 0.4384$, $\widehat{var(Y_1)} = 0.0452$, $\widehat{E(Y_2)} = 0.389$ and $\widehat{var(Y_2)} = 0.047$. The estimated correlation result is 0.417. Thus, except for the correlation, the estimated values are close to the true empirical values.

Figure 4 shows (top) the smooth kernel density and the pdf of the estimated univariate beta distribution. The lower part of the figure shows the smooth contour plot obtained from the data (left) and the estimated contour plots of the bivariate beta pdf. Clearly, the univariate beta distribution provides a better fit to the sample values of the mathematics marks than to those for statistics. The contour plot has a similar shape to the smooth contour plot, from which we conclude that the bivariate distribution is a better model than the univariate one.

5.2. Including covariates

The two models, univariate and bivariate, were then evaluated, making use of all the covariates described in Tables 2 and 3. The normal equations and Fisher's information matrix for the univariate beta regression model, given by

$$f(y_i) = \frac{\Gamma(\theta)}{\Gamma(\theta\mu_i)\Gamma((1-\mu_i)\theta)} y_i^{\mu_i\phi_i - 1}(1-y_i)^{(1-\mu_i)\theta-1}, \quad i = 1,2$$

are discussed in detail in Ferrari and Cribari-Neto (2004). The regression results are shown in Tables 5 and 6. Better results are obtained with the bivariate regression model than by the separate estimation of two univariate beta regressions. When the covariance of the joint model is close to zero, the two models are nested. A likelihood-ratio test, comparing the bivariate value to the sum of the log-likelihood values of the separate estimation, provided further evidence of the advantages of the bivariate beta regression model.

The univariate model was then analysed for each module (see Table 5). With respect to personal and academic information, the following significant variables were obtained: AGE and ADMSCORE. In addition, the AGE-SQUARED variable was introduced to determine whether increased age was associated with poorer performance in this academic area.

In the univariate model, the marks for statistics did not seem to be influenced by the students' skills in mathematics, as the significant variables for statistics did not differ from those found for the Mathematics for Business class. This fact might be related to the students' background and/or to the class content. A good command of systems of equations, together with an understanding of derivatives and of the simplification process can have a positive effect on the marks obtained for mathematics, because a large amount of basic calculus is included in the topics addressed in this subject.

The bivariate model obtains better results because new significant variables are present. On the one hand, for personal and academic information, the significant variables

are the same as in the univariate model, AGE and ADMSCORE (see Table 6). In both cases, the marginal effect is positive. Thus, the older the students and the better their admission score, the higher the marks obtained in the Mathematics for Business course. However, for the AGE-SQUARED variable, the marginal effect is negative. The age factor may have a positive effect on the students who are retaking the course, due to the knowledge acquired from the previous year, in the case of those whose age is close to that of the non-retakers (i.e. 18 years). On the other hand, when the AGE-SQUARED variable is considered, the students' additional age has a negative effect. We believe this is because older students have much greater difficulty in understanding the course contents. The same effects of the covariates were observed with respect to the statistics course. In the latter case, however, a further variable, NEWCOMER, was significantly present in the bivariate model, with a negative marginal effect. It may be relevant that the new students, before starting university studies, took a course focused on statistics, although not on calculus; however, this background does not seem to have any positive impact on their later performance.

Table 5: *Details for univariate fitted models including covariates.*

MATHEMATICS				
Personal and academic information				
Variable	Coeff	Std Error	$\|t\|$-Stat	p-value
AGE	0.47428300	0.24773300	1.91450000	0.05695610
AGE-SQUARED	–0.00859712	0.00535907	1.60422000	0.11021300
ADMSCORE	0.47627700	0.12479800	3.81639000	0.00017952
Skill in Mathematics				
Variable	Coeff	Std Error	$\|t\|$-Stat	p-value
SYSTEM	0.35490300	0.12840300	2.76398000	0.00623282
DERRATEXP	0.46429600	0.18433700	2.51874000	0.01254520
SIMPLIFYDER	0.58471300	0.24651500	2.37192000	0.01862630
STATISTICS				
Personal and academic information				
Variable	Coeff	Std Error	$\|t\|$-Stat	p-value
AGE	0.31052900	0.23665300	1.31217000	0.19091700
AGE–SQUARED	–0.00528830	0.00513981	1.02889000	0.30473200
ADMSCORE	0.34814600	0.12009600	2.89889000	0.00414850
Other parameters				
Constant for Mathematics	–6.80259000	2.81041000	2.42050000	0.01637510
Constant for Statistics	–4.65696000	2.66416000	1.74800000	0.08194620
θ for Mathematics	4.88422000	0.43697500	11.17730000	0.00000000
θ for Statistics	5.06500000	0.45354600	11.16760000	0.00000000
Value of the AIC for Mathematics: –94.917				
Value of the AIC for Statistics: –101.721				

Table 6: *Details for bivariate fitted model including covariates.*

MATHEMATICS					
Personal and academic information					
Variable	Coeff	Std Error	$\|t\|$-Stat	p-value	δ_j
AGE	0.228201842	0.183687046	1.24234	0.21411099	1.256
AGE-SQUARED	–0.003363921	0.004001199	0.84073	0.40050008	0.996
ADMSCORE	0.434494686	0.090404187	4.80613	0.00000154	1.544
Skill in Mathematics					
Variable	Coeff	Std Error	$\|t\|$-Stat	p-value	δ_j
FRACTIONS	0.254534936	0.089643529	2.83941	0.00451967	1.289
SYSTEM	0.412448241	0.124542939	3.31170	0.00092733	1.510
EQ2	–0.365876393	0.124355071	2.94219	0.00325899	0.693
SIMPLIFYEXP	0.421013978	0.110824320	3.79893	0.00014532	1.523
DERRATEXP	0.299258577	0.130434786	2.29432	0.02177237	1.348
SIMPLIFYDER	0.434442994	0.159227590	2.72844	0.00636346	1.544
STATISTICS					
Personal and academic information					
Variable	Coeff	Std Error	$\|t\|$-Stat	p-value	δ_j
AGE	0.152238309	0.210965922	0.72163	0.47052499	1.164
AGE-SQUARED	–0.001916838	0.004500367	–0.42593	0.67015944	0.998
ADMSCORE	0.375708496	0.099697582	3.76848	0.00016424	1.456
FRESHMEN	–0.357923647	0.108415456	3.30141	0.00096201	0.700
Skill in Mathematics					
Variable	Coeff	Std Error	$\|t\|$-Stat	p-value	δ_j
DERINTEXP	0.205826943	0.089275677	2.30552	0.02113741	1.228
Other parameters					
θ	3.164211826	0.199656264	15.84830	0.00000000	
Constant for Mathematics	–3.989088242	2.067120538	1.929780	0.05363408	
Constant for Statistics	–2.720631590	2.433508707	1.117990	0.26357246	
Value of the AIC: –285.820					

In the bivariate model, with respect to mathematics skills, some of the variables observed in the univariate model were again found to be relevant; in addition, the folowing new ones appeared: FRACTIONS, SYSTEM, EQ2, SIMPLIFYEXP, DERRATEXP and SIMPLIFYDER. In every case, the marginal effects were positive. Thus, when students are competent with the basic algebra of rational expressions, they are more likely to obtain higher marks in mathematics. The same is true when they can correctly apply a method for resolving a linear equations system to generate a quadratic equation to be solved. Another factor that appears to be significant is the ability to simplify algebraic expressions, to derive polynomial functions with rational exponents and to sim-

plify the expression of the derivative function obtained. For these covariates, the positive marginal effects mean that the students' marks increase when they are able to correctly complete the exercises in question. However, the corresponding results for the statistics course show that the only significant factor was the covariate defining whether the students were capable of determining the derivative of a polynomial expression with integer exponents. Success in this task was also associated with higher marks in the subject, possibly because this type of expression appears in some elements of the statistics course.

6. Results and conclusions

As part of the Business Administration degree offered by the University of Las Palmas de Gran Canaria (Spain), a Mathematics for Business course is taught in the first term of the first year; this is followed by a course focusing on applied statistics in social sciences. In view of the obvious connection between these two courses, we decided to analyse the relationship between the marks obtained in each course and to determine which covariates might affect these marks.

Accordingly, we considered a flexible bivariate regression model to be applied when the dependent variables are bounded and the marginal means are functions of the explanatory variables. This model was applied to study the personal and academic factors relevant to the students in our study sample and the basic mathematical skills that may affect the marks obtained in the above-mentioned courses (mathematics and statistics). In our opinion, the model proposed is competitive with that presented by Cepeda-Cuervo et al. (2014), who generated a bivariate beta regression model from copulas evaluating it using a Bayesian methodology. As in our own case, the marginal distributions of the latter model were beta, but these authors assumed a weak dependence between the variables of interest, which was modelled by a Farlie-Gumbel-Morgenstern copula function. The model we propose has fewer parameters and therefore is simpler.

The results obtained in the present analysis show that the mean value of the marks obtained increases with the age of the students, in both courses. Specifically, the students who were born before 1995 had higher marks both for mathematics and for statistics. We interpret this finding as follows: some of the students in the final sample had been enrolled in the same mathematics course the previous year, and so they were not newcomers to the subject. Indeed, some had taken remedial courses, or had transferred from other undergraduate studies. Thus, following an initial lack of success, these students subsequently acquired mathematics skills enabling them to achieve better marks in the subject.

With regard to the admission score variable, this too was significant for both subjects, with a positive marginal effect. Thus, the higher the admission mark the better the marks obtained for mathematics and statistics. In this respect, obviously, the best students were most likely to achieve the best marks in mathematics and statistics.

Among the other variables related to personal information, another relevant factor was whether the students were newcomers, i.e. studying these subjects for the first time. Nevertheless, this variable was only significant for the statistics subject, which probably reflects the background acquired in this respect in the Social Sciences track studied at high school.

Finally, with regard to the influence of mathematical skills on the marks obtained for statistics, only the variable related to obtaining the derivative of polynomial expressions with integer exponents was found to be significant. It is striking that no other mathematical ability affected the marks for statistics. This might be because the basic statistics course in question is mainly descriptive, merely introducing the main concepts; consequently, most of the students were already acquainted with these concepts having opted for the Social Science track at high school. Despite these considerations, however, the marks obtained for statistics and the success rate in this course were even worse than for the business mathematics course.

In the light of the results obtained, we conclude that the bivariate beta regression model is more suitable than the univariate model for the analysis described in this paper.

Acknowledgements

The authors would like to express their gratitude to the Associate Editor and to an anonymous reviewer for their valuable comments and suggestions. EGD was partially funded by grant ECO2013-47092 (Ministry of Economy and Competitiveness, Spain).

Appendix

We present the equations needed to perform the estimation using the maximum likelihood method when covariates are introduced into the model. Consider a sample consisting of n observations $(\tilde{y}_1, \tilde{y}_2) = \{(y_{11}, y_{21}), \ldots, (y_{1n}, y_{2n})\}$, taken from the probability function (7). The log-likelihood is given by

$$\begin{aligned}\ell \equiv \ell(\theta, \beta_1, \beta_2; (\tilde{y}_1, \tilde{y}_2)) = \sum_{i=1}^{n} [&(\phi_{2i} - 1)\log(1 - y_{1i}) + (\phi_{1i} - 1)\log(1 - y_{2i}) \\ &+ (\phi_{1i}\mu_{1i} - 1)\log y_{1i} + (\phi_{2i}\mu_{2i} - 1)\log y_{2i} \\ &- \frac{\phi_{1i}\phi_{2i}}{\theta}(1 - \mu_{1i}\mu_{2i})\log(1 - y_{1i}y_{2i}) \\ &- \log B(\phi_{1i}\mu_{1i}, \phi_{2i}\mu_{2i}, \theta)], \end{aligned} \tag{8}$$

where $\phi_{\kappa i} = \theta/(1 - \mu_{\kappa i})$, $\kappa = 1, 2$.

From straightforward computation, we have

$$\frac{\partial \mu_{\kappa i}}{\partial \beta_{\kappa j}} = \mu_{\kappa i} x_{\kappa j}, \;\; \frac{\partial \phi_{\kappa i}}{\partial \beta_{\kappa j}} = \frac{1}{\theta}(\phi_{\kappa i}\mu_{\kappa i} x_{\kappa j})^2,$$

from which we obtain the first partial derivatives of the log-likelihood function (8) with respect to θ and $\beta_{\kappa j}$ $(\kappa = 1,2,\ j = 1,\ldots,p)$, given by

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta} &= \sum_{i=1}^{n}\left[\frac{\log(1-y_{1i})}{1-\mu_{2i}} + \frac{\log(1-y_{2i})}{1-\mu_{2i}} + \frac{\mu_{1i}\log y_{1i}}{1-\mu_{1i}} + \frac{\mu_{2i}\log y_{2i}}{1-\mu_{2i}}\right. \\
&\quad \left. + \frac{\phi_{1i}\phi_{2i}}{\theta^2}(1-\mu_{1i}\mu_{2i})\log(1-y_{1i}y_{2i}) - \frac{\psi(\theta)-\psi(\theta+\sum_{\kappa=1}^{2}\phi_{\kappa i}\mu_{\kappa i})}{B(\phi_{1i}\mu_{1i},\phi_{2i}\mu_{2i},\theta)}\right], \\
\frac{\partial \ell}{\partial \beta_{1j}} &= \sum_{i=1}^{n}\mu_{1i}x_{1j}\left[\left(\mu_{2i} - \frac{\phi_{1i}\mu_{1i}x_{1j}}{\theta}(1-\mu_{1i}\mu_{2i})\right)\frac{\phi_{1i}\phi_{2i}}{\theta}\log(1-y_{1i}y_{2i})\right. \\
&\quad + \left(1+\frac{\phi_{1i}}{\theta}\mu_{1i}^2 x_{1j}\right)\phi_{1i}\log y_{1i} + \frac{\phi_{1i}^2\mu_{1i}x_{1j}}{\theta}\log(1-y_{2i}) \\
&\quad \left. + \frac{\mu_{1i}\phi_{1i}x_{1j}}{B(\phi_{1i}\mu_{1i},\phi_{2i}\mu_{2i},\theta)}\left(\psi(\mu_{1i}\phi_{1i}) - \psi(\theta+\sum_{\kappa=1}^{2}\phi_{\kappa i}\mu_{\kappa i})\right)\right], \\
\frac{\partial \ell}{\partial \beta_{2j}} &= \sum_{i=1}^{n}\mu_{2i}x_{2j}\left[\left(\mu_{1i} - \frac{\phi_{2i}\mu_{2i}x_{2j}}{\theta}(1-\mu_{1i}\mu_{2i})\right)\frac{\phi_{1i}\phi_{2i}}{\theta}\log(1-y_{1i}y_{2i})\right. \\
&\quad + \left(1+\frac{\phi_{2i}}{\theta}\mu_{2i}^2 x_{2j}\right)\phi_{2i}\log y_{2i} + \frac{\phi_{2i}^2\mu_{2i}x_{2j}}{\theta}\log(1-y_{1i}) \\
&\quad \left. + \frac{\mu_{2i}\phi_{2i}x_{2j}}{B(\phi_{1i}\mu_{1i},\phi_{2i}\mu_{2i},\theta)}\left(\psi(\mu_{2i}\phi_{2i}) - \psi(\theta+\sum_{\kappa=1}^{2}\phi_{\kappa i}\mu_{\kappa i})\right)\right],
\end{aligned}$$

where $j = 1,\ldots,p$. By equating these $2p+1$ equations to zero and then solving, we obtain the maximum likelihood estimates of the model parameters.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Arnold, J.M. and Straten, J.T. (2012). Motivation and skills as determinants of first-year performance in economics. *The Journal of Economic Education*, 43, 33–47.

Bayes, C.L., Bazán, J.L. and García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, 4, 841–866.

Brooks, C. (2009). *RATS Handbook to Accompany Introductory Econometrics for Finance*. Cambridge University Press.

Cepeda-Cuervo, E. (2001). Modelagem de variabilidade em modelos lineares generalizados. *PhD thesis, Mathematics Institute, Federal University of Rio de Janeiro. Unpublished Thesis*.

Cepeda-Cuervo, C., Achcarb, J.A. and Garrido, L. (2014). Bivariate beta regression models: joint modeling of the mean, dispersion and association parameters. *Journal of Applied Statistics*, 41, 677–687.

Cepeda-Cuervo, E., Núñez-Antón, V. (2013). Spatial double generalized beta regression models: extensions and applications to study quality of education in Colombia. *Journal of Educational and Behavioral Statistics*, 38, 604–628.

Dolado, J.J. and Morales, E. (2009). Which factors determine academic performance of economics freshers? some spanish evidence. *Investigaciones Económicas*, 33, 179–210.

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

Gómez-Déniz, E., Sordo, M. and Calderín-Ojeda, E. (2013). The log-Lindley distribution as an alternative to the beta regression model with applications in insurance *Insurance: Mathematics and Economics*, 54, 49–57.

Gottschalk, J. and Maslen, E. (1988). Reduction formulae for generalised hypergeometric functions of one variable. *Journal of Physics A: Mathematical and Theoretical*, 21, 1983–1998.

Gradshteyn, I. and Ryzhik, I. (1994). *Table of Integrals, Series, and Products*. Boston: Academic Press (5^{th} ed. A. Jeffrey).

Gurmu, S. and Elder, J. (2000). Generalized bivariate count data regression. models. *Economics Letters*, 68, 31–36.

Huang, X. and Oosterlee, C. (2011). Generalized beta regression models for random loss given default. *The Journal of Credit Risk*, 7, 1–26.

Jaynes, E. (1994). *Probability Theory: The Logic of Science*. `http://omega.math.albany.edu:8008/JaynesBook.html`.

Johnson, M. and Kuennen, E. (2006). Basic math skills and performance in an introductory statistics course. *Journal of Statistics Education*, 14, 20–40.

Lee, M.-L.T. (1996). Properties and applications of the Sarmanov family of bivariate distributions. *Communications in Statistics-Theory and Methods*, 25, 1207–1222.

Lehmann, E. (1996). Some concepts of dependence. *Annals of Mathematical Statistics*, 37, 1137–115.

Lunsford, M.L. and Poplin, P. (2011). From research to practice: basic math skills and success in introductory statistics. *Journal of Statistics Education*, 19, 1–22.

Maher, M. (1990). A bivariate negative binomial model to explain traffic accident migration. *Accident Analysis and Prevention*, 22, 487–498.

Olkin, I. and Liu, R. (2003). A bivariate beta distribution. *Statistics & Probability Letters*, 62, 407–412.

Paolino, P. (2001). Maximum likelihood estimation of models with beta distributed dependent variables. *Political Analysis*, 9, 325–346.

Papke, L. and Wooldridge, J. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, 11, 619–632.

Papke, L. and Wooldridge, J. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, 145, 121–133.

Pérez-Rodríguez, J.V. and Gómez-Déniz, E. (2015). Spread component costs and stock trading characteristics in the spanish stock exchange. *Quantitative Finance*, 15, 1943–1962.

Tong, Y. (1980). *Probability Inequalities in Multivariate Distributions*. Academic Press, New York.

Wolfram, S. (2003). *The Mathematica Book*. Wolfram Media, Inc.

SORT 41 (2) July-December 2017, 277-296 DOI: 10.2436/20.8080.02.60

Bayesian hierarchical models for analysing the spatial distribution of bioclimatic indices

Xavier Barber*,[1], David Conesa[2], Antonio López-Quílez[2], Asunción Mayoral[1], Javier Morales[1] and Antoni Barber[3]

Abstract

A methodological approach for modelling the spatial distribution of bioclimatic indices is proposed in this paper. The value of the bioclimatic index is modelled with a hierarchical Bayesian model that incorporates both structured and unstructured random effects. Selection of prior distributions is also discussed in order to better incorporate any possible prior knowledge about the parameters that could refer to the particular characteristics of bioclimatic indices. MCMC methods and distributed programming are used to obtain an approximation of the posterior distribution of the parameters and also the posterior predictive distribution of the indices. One main outcome of the proposal is the spatial bioclimatic probability distribution of each bioclimatic index, which allows researchers to obtain the probability of each location belonging to different bioclimates. The methodology is evaluated on two indices in the Island of Cyprus.

MSC: 62F15, 62M30, 62P10, 62P12, 86A32.

Keywords: Bioclimatology, geostatistics, parallel computation, spatial prediction.

1. Introduction

Bioclimatology is an ecological science that studies the relationship between climate and the distribution of the living species on Earth, particularly the distribution of vegetation. It aims to determine the relationship between certain numerical values of temperature and precipitation and the areas in which single plant species and plant communities are geographically distributed. The spatial distribution of the species and the relationship between climate and vegetation allows us to better manage plant resources and landscape, as well as to forecast the production of agricultural and forestry resources to combat hunger and determine future vegetation scenarios in certain geographic areas through the study of vegetation borders.

[1] Centro de Investigación Operativa. Universidad Miguel Hernández de Elche. xbarber@umh.es

[2] Dpt. Estadística i Investigació Operativa. Universitat de València.

[3] IDENTIA Institute.

Received: June 2016
Accepted: March 2017

As an ecological science, the distribution of the spatial structure of species and its relationship with environmental factors having high spatial dependence has been an important subject of study for several years. Osborne et al. (2000); Britton et al. (2001); Cheddadi, Guiot and Jolly (2001); Tasser and Tappeiner (2002); Legendre, Borcard and Peres-Neto (2005); Dostálek, Frantík and Šilarová (2014); Baltensperger and Huettmann (2015) are examples of studies applying these ideas to analyse land-use changes and distribution of terrestrial vegetation.

Bioclimatic Classification Systems have been introduced to assign bioclimates to a region under study by means of what are known as bioclimatic indices. But more importantly, these bioclimates allow us to identify the geographical limits of the main types of vegetation in the region under study. As a result, having a good spatial representation of the bioclimatic indices is key to describing the relationship between climate and the distribution of vegetation.

Information about bioclimatic indices is usually available only in meteorological stations, not in the whole region of study. It is therefore important to be able to construct maps from these data. Until now, many studies have used only standard geographical information system (GIS) techniques. Geostatistics has also been proposed as a way to explain bioclimatic indices (Robertson, 1987; Rossi et al., 1992; Burrough, 2001; Garzón-Machado, Otto and del Arco Aguilar, 2014), although this approach can present certain obstacles such as spatial scale problems (Atkinson and Tate, 2000).

Our main interest in this research is twofold. Firstly, we present another way to model the spatial distribution of bioclimatic indices. Specifically, we propose a hierarchical Bayesian model to predict (in non-sampled locations) the bioclimatic index values by incorporating the altitude and spatial features of each sampled location. As usual in Bayesian approaches, we also explain how to select prior distributions in this context. But more importantly, we secondly describe the two main outcomes of the modelling, i.e., the posterior predictive distribution of bioclimatic indices and the probability maps for the bioclimates, which provide more realistic geographical limits. As the resulting hierarchical model has no closed expression for the posterior distribution of all the parameters, we also present how to perform inference by MCMC methods, and how to predict on non-observed locations by means of distributed programming, reducing the computation time by more than 80% in comparison to standard R packages.

The remainder of this article is organised as follows. After this introduction, Section 2 presents a general Bayesian hierarchical spatial model of the bioclimatic indices. In Section 3, we describe how to select prior distributions, while Section 4 explains how to perform inference and prediction for these indices. In Section 5, we apply this methodology in a real setting, we obtain the predictive distributions of two bioclimatic indices on the island of Cyprus, using the altitude and the climate information (temperatures and rainfall) from 59 meteorological stations. Finally, Section 6 concludes and presents some future lines of research.

2. Modelling bioclimatic indices

In what follows, we first introduce three bioclimatic indices of the Worldwide Bioclimatic Classification System by Rivas-Martínez (Rivas-Martínez, 1994; Rivas-Martínez et al., 2002; Rivas-Martínez and Rivas-Saenz, 2016), one of the most popular Bioclimatic Classification Systems available. This classification encompasses five macrobioclimates (Tropical, Mediterranean, Temperate, Boreal and Polar), which are in turn subdivided into twenty-seven bioclimates and five bioclimatic variants. It is worth noting that all the results presented here could also be applied to any other bioclimatic index from any classification selected. After defining the bioclimatic indices, Section 2.2 describes the Bayesian hierarchical spatial model for each one of them.

2.1. Bioclimatic indices

As previously mentioned, the procedure for constructing bioclimatic maps is based on the bioclimatic indices. In general, these indices are values obtained by simple mathematical expressions that combine certain climatic parameters and factors such as altitude or latitude, and which are commonly used to characterise the climate of a region. This makes it possible to recognise climatically homogeneous areas that may have similar vegetation types (species, communities, series).

One of the most important bioclimatic indices is the *Ombrothermic Index* (OI), which relates the rainfall and the temperature in an area using an average of the last n years (usually at least 25 years), and it is defined by

$$OI = \frac{10}{n} \sum_{j=1}^{n} \left(\frac{P_{p,j}}{T_{p,j}} \right), \tag{1}$$

where P_p is the sum of the average rainfall (in mm.) of the months whose average temperature is above zero degrees Celsius, and T_p is the sum of monthly average temperatures above zero degrees Celsius, expressed in tenths of a degree.

The variation of temperature (thermicity) over the seasons in an area is one of the most influential factors in the characterisation of climate, since the vegetation distribution is greatly affected by the area's thermicity. Hence, another important bioclimatic index is the *Thermicity Index* (TI) of the last n years, defined as

$$TI = \frac{10}{n} \sum_{j=1}^{n} (T_j + m_j + M_j), \tag{2}$$

where T is the sum of the annual mean temperature in decimal degrees, m is the average of the minimum temperature of the coldest month and M is the average of the maximum temperature of the coldest month.

This Thermicity Index has some problems of definition in extratropical regions (North and South of latitude 23 N and S respectively). The Compensated Thermicity Index (TIc) avoids these problems by weighting the Thermicity Index value (TI) by adding or subtracting the Compensation Value, C_i, in those places where the Continentality Index (CI), defined as the annual oscillation variation of temperature $CI = T_{\max} - T_{\min}$) takes extreme values:

$$TIc = \begin{cases} TI & \text{if } 8 \leq CI \leq 18, \\ TI + C_i & \text{if } CI < 8 \text{ or } CI > 18, i = 0, \ldots, 4 \end{cases} \tag{3}$$

Note that all the temperatures are in Celsius, and periods are 25 years, the minimum recommended period.

2.2. Bayesian hierarchical model for bioclimatic indices

After presenting the bioclimatic indices, we now introduce a way of modelling them by means of a Bayesian hierarchical spatial model. If $\boldsymbol{Y} = [Y(s_i)]_{i=1}^n$ represents the vector of values of the bioclimatic index in a subset of locations $\boldsymbol{s} = (s_1, \ldots, s_n)$ in the region D, then the usual geostatistical assumption is that $\boldsymbol{Y}$ is multivariate normal:

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{\mu}$ denotes the mean vector of the process, and $\boldsymbol{\Sigma}$ represents the covariance matrix between locations. This matrix can be re-written separately as spatial and non-spatial covariances matrices

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_r, \tag{5}$$

which, assuming that the observations are conditionally independent given the spatial process, can also be expressed as

$$\boldsymbol{\Sigma}_w = \sigma^2 \boldsymbol{H}(\boldsymbol{\theta}); \text{ and } \boldsymbol{\Sigma}_r = \tau^2 \boldsymbol{I}, \tag{6}$$

where $\boldsymbol{H}(\boldsymbol{\theta})$ is the Matérn correlation matrix between locations (Matérn, 1986), which depends on two parameters $\boldsymbol{\theta} = (\phi, \nu)$; the scale parameter $\phi > 0$ and the shape parameter $\nu > 0$. It is worth noting that the Matérn is a really flexible and general family of correlation generalising many of the most-used covariance models in spatial statistics (exponential and Gaussian among them).

The mean vector of the process can be related with covariates (in our case, altitude), and so the bioclimatic index is expressed as

$$\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{W},\tau^2 \sim \mathcal{N}\left(X\boldsymbol{\beta}+\boldsymbol{W},\tau^2\boldsymbol{I}\right), \tag{7}$$

where $X\boldsymbol{\beta}$ represents the linear predictor associated with the covariates at the locations $\boldsymbol{s}=(s_1,\ldots,s_n)$.

Hence, the Bayesian hierarchical model corresponding to geostatistical homogeneous Gaussian process data for a bioclimatic index is expressed in three levels of information as

$$\begin{aligned} &(I) \quad \boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{W},\tau^2 \sim \mathcal{N}\left(X\boldsymbol{\beta}+\boldsymbol{W},\tau^2\boldsymbol{I}\right) \\ &(II) \quad \boldsymbol{W}|\sigma^2,\boldsymbol{\theta} \sim \mathcal{N}\left(0,\sigma^2\boldsymbol{H}(\boldsymbol{\theta})\right) \\ &(III) \quad p(\boldsymbol{\beta},\sigma^2,\tau^2,\boldsymbol{\theta}), \end{aligned} \tag{8}$$

where the first level is the Gaussian process, the second level shows the information on the spatial effect and the third level specifies the prior distribution parameters and hyper-parameters.

Following Yan et al. (2007), and in order to avoid the identifiability problem of spatial and non-spatial variability, we reparametrise (8) as

$$\begin{aligned} &(I)\,\boldsymbol{Y} \sim \mathcal{N}\left(X\boldsymbol{\beta},\xi^2\left[(1-\kappa)\boldsymbol{H}(\boldsymbol{\theta})+\kappa\boldsymbol{I}\right]\right) \\ &(II)\,p(\boldsymbol{\beta},\xi^2,\kappa,\boldsymbol{\theta}), \end{aligned} \tag{9}$$

where $\xi^2=\sigma^2+\tau^2$ now represents the total variability of the random effects, and $\kappa=\tau^2/\xi^2$ stands for the proportion of the non-spatial variability with respect to the total variability.

Once the model is determined, the next step is to estimate its parameters. As we are using the Bayesian paradigm, we have to select the prior distribution for the vector of parameters involved in the model.

3. Selection of prior distributions

Making use of previous information is considered one of the most useful characteristics of Bayesian statistics. A subjective approach involves defining prior distributions for unknown parameters according to personal experience and impression, recognising that the expert opinion is better than no knowledge. In contrast, objective Bayesians defend the idea that no other information should be considered apart from that introduced during model specification, although finding that prior distribution which contains only that knowledge can sometimes be tricky. In the context of spatial geostatistics models, the case in hand, it must be taken into account that using non-informative priors can lead to improper posterior distributions (De Oliveira, 2007).

A usual assumption when expressing prior knowledge is to consider prior independence of the parameters, that is,

$$p(\boldsymbol{\beta}, \xi^2, \kappa, \boldsymbol{\theta}) = p(\boldsymbol{\beta}) p(\xi^2) p(\kappa) p(\boldsymbol{\theta}) .$$

In order to express our knowledge for each of these parameters, we must elicit both their distributions and the values of their hyperparameters. As mentioned above, for the latter the choice of their values can come under the "complete ignorance" premise, although we can also include the information available about them in order to improve the final posterior distribution (Dongen, 2006).

In particular, the distribution for $\boldsymbol{\beta}$ is again based on the assumption of prior independence of its components, the usual choice being either Gaussian distributions or non-informative improper distributions. As the resulting posterior is in both cases proper, we use the improper one, that is,

$$p(\boldsymbol{\beta}) = p(\beta_0, \beta_1) = p(\beta_0) p(\beta_1) \propto 1 .$$

With respect to the proportion κ, the natural choice is a uniform distribution between 0 and 1, $\kappa \sim \mathcal{U}(0,1)$.

For the Matérn function parameters, $\boldsymbol{\theta} = (\phi, \nu)$, and taking into account that we are using the parameterisation proposed by Handcock and Wallis (1994) in which the parameter ϕ is largely independent of ν, we propose using a product of two independent distributions. In particular, our choice for the prior distribution of ϕ is

$$p(\phi) = \mathcal{U}\left(\frac{1}{d_1}, \frac{1}{d_2}\right), \tag{10}$$

where d_1 is the furthest distance between two locations, and d_2 is the minimum distance between the two nearest locations. Following recommendations by Stein (1999) and Finley, Banerjee and Gelfand (2015), our choice of smoothing parameter ν is $\nu \sim \mathcal{U}(0.05, 1.95)$.

The last parameter to be elicited is the total variability ξ^2 of the bioclimatic index. In this case, note that information is available which can be included in the prior. Indeed, as explained in the previous section, indices depend on temperature and precipitation by definition, and therefore, they only take values within a defined range (the highest and lowest value of the index in the region of study, according the Rivas-Martínez classification), denoted by $(Y_{\min}, Y_{\max})$, with $Y_{\min} > 0$.

This information about ξ^2 can be incorporated in the scale parameters of different distributions. The underlying idea is to consider that the observed values of the index on a set of locations is a priori uniformly distributed between $(Y_{\min}, Y_{\max})$. Note that this uniform distribution is the most disadvantageous option as this would imply that all the

regions have the same orographic features. The corresponding variability of this uniform distribution is

$$\mathrm{Var}(Y) = \frac{(Y_{\max} - Y_{\min})^2}{12}, \tag{11}$$

the maximum value of which (denoted as $V_{\max}$) would be an upper bound of the variability index. A prior distribution could then be constructed by matching the range of variability $(a, V_{\max})$ with the quantile 0.95 of any chosen distribution (Chambers and Dunstan, 1986; Strupczewski et al., 2007). In other words,

$$0.95 = \int_a^{V_{\max}} f(y|\boldsymbol{\alpha})dy, \tag{12}$$

where f is the chosen prior distribution and $\boldsymbol{\alpha}$ its corresponding parameters. Since variability is always positive, a can be chosen to be as small as possible (e.g. $a = 0.001$). Table 1 shows the resulting scale parameters for the usual priors: uniform over the variance, uniform over the standard deviation, inverse gamma or half-Cauchy.

Table 1: *Upper bound for the variability index and prior distribution for a specific bioclimatic index range.*

$p(\xi^2) \sim \mathcal{U}(0.001, b)$	$b = \frac{V_{\max} - 0.00005}{0.95}$
$p(\xi) \sim \mathcal{U}(0.001, \sqrt{b})$	$b = \frac{V_{\max} - 0.00005}{0.95}$
$p(\xi^2) \sim \mathcal{IG}(2, \beta)$	$0.95 = \int_{0.001}^{V_{\max}} \frac{\beta^2}{\Gamma(2)} x^{-3} e^{-\beta x} dx$
$p(\xi^2) \sim \mathcal{HC}(\delta)$	$\delta = \frac{V_{\max}}{tan\left(\frac{1}{2} \cdot \pi \cdot 0.95\right)}$

To summarise, the final model for any bioclimatic index $\boldsymbol{Y}$ using the second option of Table 1 (uniform over the standard deviation) is

$$\begin{aligned} &(I) \quad \boldsymbol{Y} \sim \mathcal{N}\left(X\boldsymbol{\beta}, \xi^2\left[(1-\kappa)\boldsymbol{H}(\boldsymbol{\theta}) + \kappa \boldsymbol{I}\right]\right);\ \boldsymbol{\theta} = (\phi, \nu) \\ &(II) \quad p(\boldsymbol{\beta}, \xi, \kappa, \phi, \nu) \propto 1 \times \mathcal{U}(0.001, \sqrt{b}) \times \mathcal{U}(0,1) \times \mathcal{U}(1/d_1, 1/d_2) \times \mathcal{U}(0.05, 1.95) \end{aligned} \tag{13}$$

Note that the advantage of this final model is that we only have to assign a prior distribution on ξ, since the remaining parameters are obtained as $\sigma^2 = (1-\kappa)\xi^2$ and $\tau^2 = \kappa\xi^2$.

4. Inference and prediction

The model in (13) contains all our knowledge about the index, but it does not yield closed analytic expressions for the posterior distribution of the parameters, $p(\boldsymbol{\beta},\xi,\kappa,\phi,\nu|Y,X)$. Therefore, numerical approximations are needed in order to make inference about them. Among others, one feasible (indeed one of the most popular) possibility is to use Markov chain Monte Carlo (MCMC) methods (Gamerman and Lopes, 2006) that draw samples from any intractable posterior by running a cleverly constructed Markov chain over a long period, the stationary distribution of which is the one we want to simulate from. Among the different ways of building these chains, the most popular are Gibbs sampling and the Metropolis-Hastings algorithm (Gilks, Richardson and Spiegelhalter, 1996).

In our case, we use WinBUGS (Lunn et al., 2000), a flexible software for performing the Bayesian analysis of complex statistical models (see Banerjee, Carlin and Gelfand, 2014 for examples of how to implement spatial hierarchical Bayesian models with WinBUGS). The reason for this choice is that it gives us more flexibility when specifying the matrix variance-covariance of the first hierarchy level. Moreover, it allows us to easily set prior distributions over the standard deviation.

As usual in MCMC, we run three chains for a long period discarding the first hundreds or thousands (depending on the convergence, the burn-in period can be extended) and then take samples from the three chains. Regarding convergence (to the correct stationary distribution) assessment, the Brooks-Gelman-Rubin statistic and the effective sample size (see Gelman et al., 2013 for more information about these statistics) can be calculated for every parameter in the model. The Brooks-Gelman-Rubin statistic must have a value under 1.1, while the effective number of iterations must be above 100 for every mentioned parameter.

Once the inference has been carried out, the next step is to predict the values of the bioclimatic indices in the rest of the area of interest, especially in unsampled locations. In our case, as we are using the Bayesian approach, prediction is reduced to obtain the posterior predictive distribution of the indices in a set of new locations.

In particular, if $\boldsymbol{Y}_p$ represents the values of a bioclimatic index in a new set of locations with observed covariates X_p, then the posterior predictive distribution of the new values $\boldsymbol{Y}_p$ (conditional to the observed ones, henceforth, $\boldsymbol{Y}_o$) is

$$p(\boldsymbol{Y}_p|\boldsymbol{Y}_o,X_o,X_p) = \int p(\boldsymbol{Y}_p|\boldsymbol{Y}_o,X_p,\boldsymbol{\beta},\xi,\kappa,\phi,\nu)p(\boldsymbol{\beta},\xi,\kappa,\phi,\nu|Y,X)d(\boldsymbol{\beta},\xi,\kappa,\phi,\nu)\,, \tag{14}$$

where the extended data vector $p(\boldsymbol{Y}_p|\boldsymbol{Y}_o,X_p,\boldsymbol{\beta},\xi,\kappa,\phi,\nu)$ has a conditional multivariate normal distribution arising from the joint multivariate distribution of $\boldsymbol{Y}_p$ and $\boldsymbol{Y}_o$ in (7).

As with the posterior distribution of the parameters, expression (14) has no closed form, and again numerical approximations are needed. One way to obtain a simulated sample from this posterior predictive distribution is via the composition method. In particular, if $\{\boldsymbol{\beta}_i,\xi_i,\kappa_i,\phi_i,\nu_i\}_{i=1}^{M}$, represents a simulated sample from the posterior distribu-

tion of the parameters, then a simulated sample from the posterior predictive distribution is obtained by simulating from the conditional multivariate distribution of the observed values $\boldsymbol{Y}_p$, that is, $\{p(\boldsymbol{Y}_p|\boldsymbol{Y}_o, X_p, \boldsymbol{\beta}_i, \xi_i, \kappa_i, \phi_i, \nu_i)\}_{i=1}^{M}$.

Note that the conditional multivariate distribution $p(\boldsymbol{Y}_p|\boldsymbol{Y}_o, X_p, \boldsymbol{\beta}, \xi, \kappa, \phi, \nu)$ is a multivariate normal distribution with mean

$$E\left[\boldsymbol{Y}_p|\boldsymbol{Y}_o\right] = \boldsymbol{\mu}_p + \boldsymbol{\Sigma}_{po}\boldsymbol{\Sigma}_{oo}^{-1}(\boldsymbol{Y}_o - \boldsymbol{\mu}_o) \tag{15}$$

and variance-covariance matrix

$$V\left[\boldsymbol{Y}_p|\boldsymbol{Y}_o\right] = \boldsymbol{\Sigma}_{pp} - \boldsymbol{\Sigma}_{po}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{op}\,, \tag{16}$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{pp} & \boldsymbol{\Sigma}_{po} \\ \boldsymbol{\Sigma}_{op} & \boldsymbol{\Sigma}_{oo} \end{pmatrix}$$

is the covariance matrix of the joint multivariate normal distribution of the extended data vector $(\boldsymbol{Y}_p, \boldsymbol{Y}_o)$.

As we are following the reparametrisation by Yan et al. (2007) in (9), the conditional multivariate distribution $p(\boldsymbol{Y}_p|\boldsymbol{Y}_o, X_p, \boldsymbol{\beta}, \xi, \kappa, \phi, \nu)$ is a multivariate normal distribution but with mean

$$E(\boldsymbol{Y}_p|\boldsymbol{Y}_o) = X_p\boldsymbol{\beta} + \left((1-\kappa)\boldsymbol{H}_{po}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}\right)\left((1-\kappa)\boldsymbol{H}_{oo}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}\right)^{-1}(\boldsymbol{Y}_o - X_o\boldsymbol{\beta}) \tag{17}$$

and variance-covariance matrix

$$V(\boldsymbol{Y}_p|\boldsymbol{Y}_o) = \tag{18}$$
$$\xi^2\left[\left((1-\kappa)\boldsymbol{H}_{pp}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}\right) - \left((1-\kappa)\boldsymbol{H}_{po}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}\right)\left((1-\kappa)\boldsymbol{H}_{oo}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}\right)^{-1}\left((1-\kappa)\boldsymbol{H}_{op}(\boldsymbol{\theta}) + \kappa\boldsymbol{I}\right)\right]$$

where

$$\boldsymbol{H}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{H}_{pp}(\boldsymbol{\theta}) & \boldsymbol{H}_{po}(\boldsymbol{\theta}) \\ \boldsymbol{H}_{op}(\boldsymbol{\theta}) & \boldsymbol{H}_{oo}(\boldsymbol{\theta}) \end{pmatrix}$$

is the Matérn correlation matrix between predicted and observed locations.

Implementing the above composition method implies evaluating this mean vector and variance-covariance matrix for each of the simulations. But note that this evaluation can be computationally expensive. Dealing with 15000 simulations (5000 per chain) from the posterior distribution and about 1000 new locations (to predict) would involve evaluating 15000 times expressions (17) and (18). This is the reason why we do not use WinBUGS, because although feasible, it is really slow.

An obvious (but naive) option would be to consider fewer points over the surface to predict, and a small random sample from the posterior distribution. However, this option would produce posterior predictive distributions with lower resolution and, therefore, the resulting predictive maps would have no practical interest. Other options would be to use the spatial-temporal modelling R library `spBayes` (Finley et al., 2015), or to directly implement equations (15) and (16) using programming languages such as the R matrix computation language (Bates and Maechler, 2015); C++ via the interface package `Rcpp` to connect with R (Eddelbuettel et al., 2011); or directly C++ (Sanderson, 2010).

Our approach is to use intensive computation techniques such as parallel computation (Adams et al., 1996; Blackford et al., 1997; Rosenthal, 2000; Rossini, Tierney and Li, 2007; Whiley and Wilson, 2004), that allow us to increase the performance when doing matrix calculations, and therefore, work with a large number of new locations to predict with all the samples previously obtained by simulation from the posterior distribution using WinBUGS. Nevertheless, as stated by Golub and Van Loan (1996) and Cuenca, Giménez and González (2004), the use of parallel computation is convenient only if computational times are substantially reduced.

In this study we use C language to program the prediction equations, and then the ScaLAPACK and PLAPACK libraries to perform the linear algebra calculations needed to obtain the mean vector and variance-covariance matrix. Interestingly, with this parallelisation of the algorithm for generating a multivariate normal sample, we reduce the computation time by close to 80% compared to other options such as `spBayes` and similar R packages.

Graphical representation of the posterior predictive distributions of Bioclimatic indices

Having obtained the posterior predictive distribution of the indices, our final task is to represent these distributions throughout the area of interest in order to obtain a good visualisation of their behaviour in the area. We present two different representations of these predictive distributions, the first one being the mean and the standard deviation of the posterior predictive distribution, and the second one, the probability distribution of each bioclimatic index belonging to different bioclimates.

To obtain the map of the mean (similarly the map of standard deviation), we use multilevel B-splines Approximation (Lee, 1997) to interpolate the values of the mean (the standard deviation) of the bioclimatic indices over the whole area using the obtained values of the posterior mean (standard deviation) predictive distribution on the predicted locations.

Although the mean and the standard deviation reflect most of the information about the posterior predictive distributions, the most valuable information we can get from these distributions comes from the way that they can show us the probability of each location belonging to the different bioclimates. Indeed Rivas-Martínez's bioclimatic classification system uses different ranges of the bioclimatic indices to classify the different bioclimates. For example, the Continentality Index ranks the climate in three

types, namely, Hyperoceanic ($CI \in [0,11[$), Oceanic ($CI \in [11,21[$), and Continental ($CI \in [22,65]$). Note that representing the probability of the predictive distribution of belonging to each of these ranges can be very relevant for studying changes in vegetation zones, climate change advances, and many other climatic issues that could provide valuable information for the management and use of land in the area under study.

Obtaining this probability is straightforward using the simulated values of the predictive distribution. If a bioclimatic index Y is defined in l disjoint intervals $R_1, R_2, \ldots, R_l$ that describe l bioclimates, and $\{r_{ik}\}_{i=1}^{n}$ represents a sample from the posterior predictive distribution for each location in $\{s_k\}_{k=1}^{m}$, then the posterior probability that each location belongs to each interval constituting the index is given by:

$$P(Y(s_k) \in R_j) = \int_{Rj} \int p(Y(s_k)|\boldsymbol{Y}_o, X_p, \boldsymbol{\beta}, \xi, \kappa, \phi, \nu) p(\boldsymbol{\beta}, \xi, \kappa, \phi, \nu | Y, X) d(\boldsymbol{\beta}, \xi, \kappa, \phi, \nu) dY(s_k)$$

$$\approx \frac{\#\{r_{ik} \in R_j\}}{n}, \quad j = 1, \cdot, l, k = 1, \cdot, m. \tag{19}$$

The result is a discrete probability distribution for each location that we call the spatial bioclimatic probability distribution. Note that the best way to represent this distribution is by presenting a single figure made up of different graphs, each one showing the probability of belonging to each bioclimate (see Figure 4 for an example).

The representation of each probability can be seen as a puzzle of pieces that fit by overlapping and provide the distribution boundaries between the types of bioclimates for each index. These boundaries are highly relevant because they determine the areas that could be about to change in the near future (caused for example by a slight change in climatic parameters). This representation is therefore critical in studies about climate change and its effects on the vegetation of a region.

5. Bioclimatic classification of the island of Cyprus

We illustrate the usefulness of the approach presented here through an application to analyse two bioclimatic indices (Ombrothermic Index and Thermicity Index) on the island of Cyprus with the final aim of showing its bioclimatic classification.

Cyprus is an island country in the Eastern Mediterranean. It is the third largest and the third most populous island in all the Mediterranean. Some of its geographical characteristics are as follows: it measures 240 kilometres (149 miles) long and 100 kilometres (62 miles) wide at its widest point; it lies between latitudes 34° and 36° N, and longitudes 32° and 35° E. Cyprus is dominated by two mountain systems, the Troodos and the Kyrenia Mountains, between which lies a central plateau, the Mesaoria.

The information gathered to create the bioclimatic classification of the island consisted of the geographical location, the altitude and the values of the two bioclimatic indices from 59 meteorological stations across the island, together with the geographi-

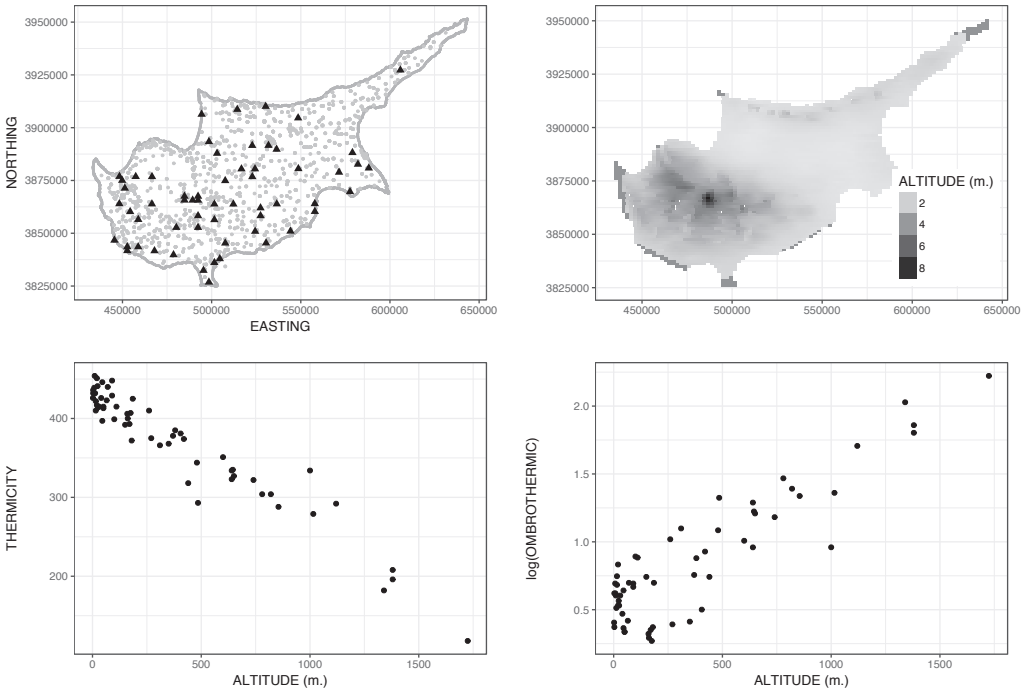

Figure 1: *Upper left: geographical location of observed and predicted sites in the Cyprus island. Black triangles represent the 59 meteorological stations (observed locations), while red points represent the 755 locations where prediction had to be performed. Upper right: contour map of the island. Lower left: thermicity and altitude relationship. Lower right: Log(Ombrothermic index) and altitude relationship.*

cal location and altitude of other 775 locations (used to predict the indices), in particular, the ones that the geographical map of the island provides. Figure 1 shows the geographical location of observed and predicted sites, jointly with the contour map of the island and the relationship between both indices and the altitude. It is worth mentioning that these two indices are not related, as can be seen in the left side of Figure 2. This allows us to analyse both indices independently. If the indices were related, a joint modelling would be necessary (see the right side of Figure 2 for an example of two related indices, namely the ombrothermic and continentality indices).

5.1. Ombrothermic Index

We first present the results obtained when analysing the Ombrothermic Index (using the logarithm transformation to improve its linear relationship with altitude). Table 2 presents the median of the posterior distribution of the parameters of the model in equation (13) along with their corresponding 95% credible intervals. These posterior distributions were obtained by simulation using WinBUGS (Lunn et al., 2000). Each posterior distribution was approximated from 15000 (5000 from each of three simulation chains)

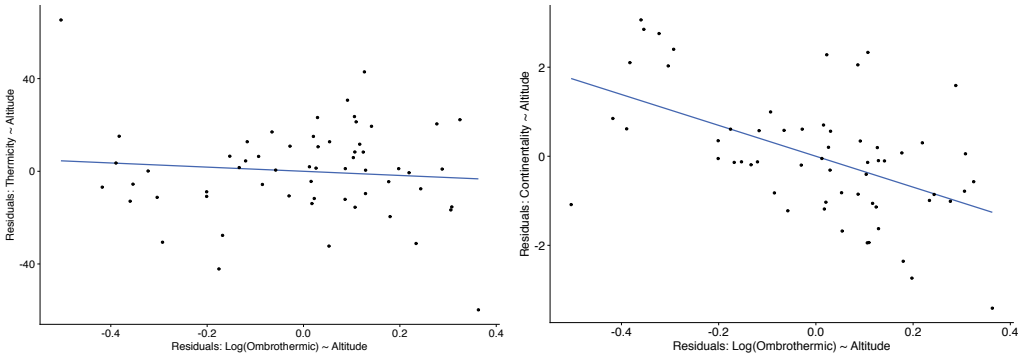

Figure 2: *Relationship between indices after adjusting linear regression of each index by altitude. Left side, relationship between residuals of ombrothermic and thermicity indices. Right side, relationship between residuals of ombrothermic and continentality indices.*

Table 2: *Median of the posterior distribution and* 95% *credible intervals of the parameters for the Ombrothermic Index model.*

Parameters	Median	$p_{2.5}$	$p_{97.5}$
β_0	5.28×10^{-1}	3.63×10^{-1}	6.84×10^{-1}
β_1	7.61×10^{-4}	6.02×10^{-4}	9.12×10^{-4}
ξ^2	4.80×10^{-2}	2.37×10^{-2}	1.70×10^{-1}
κ	1.03×10^{-1}	1.91×10^{-2}	3.83×10^{-1}
ϕ	5.54×10^{-5}	2.92×10^{-5}	9.55×10^{-5}
ν	1.48	1.02	1.93

simulated values (obtained after discarding ten thousand simulations from a burn-in period that guaranteed convergence). As commented above, these posterior distributions were obtained using the uniform distribution over the standard deviation.

As expected, results for β_1 in Table 2 show a positive effect on the altitude. Note also that the spatial effect is necessary to describe the behaviour of the index, as expressed by the small value of κ (which indicates the small proportion of non-spatial variability with respect to the total variance). It is also worth noting that the maximum variance (used in expression (11) to obtain the prior distributions) in Mediterranean bioclimate, 0.515, does not affect our results. Indeed, this shows that our prior construction methodology can really be considered as uninformative.

It is worth noting that a sensitivity analysis about the prior selection was performed for both indices. In particular, we fitted different models using all the different priors introduced in Table 1. Results indicate that both estimations and credible intervals obtained were similar independently of the priors used.

Figure 3 shows the maps of the mean and standard deviation (as a prediction error measure) of the posterior predictive distribution of the Ombrothermic Index. As men-

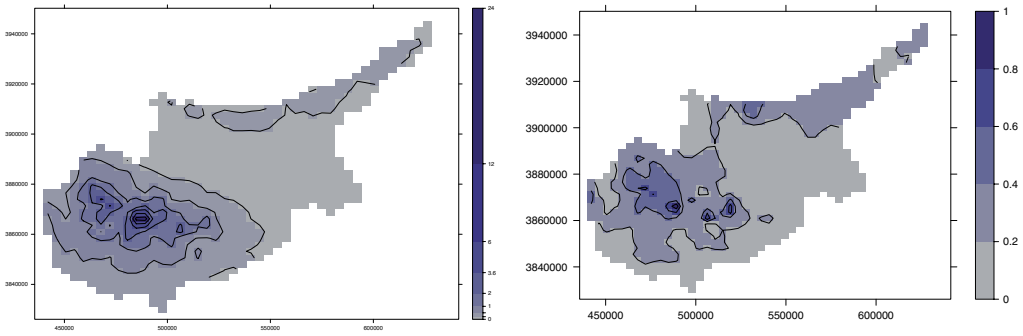

Figure 3: *Mean (left) and standard deviation (right) of the posterior predictive distribution of the Ombrothermic Index.*

tioned above, this predictive distribution was approximated by means of intensive computation techniques that allow us to predict the values of the bioclimatic indices in the 775 unsampled locations.

The mean map clearly reflects the topography of the island, while the standard deviation map shows the uncertainty in areas with no data but, more importantly, it also reflects the areas where the terrain is changing on the island. Note also that the scale of the observed prediction error is very small compared to the scale of measurement of the index considered throughout the island. The proposed method is therefore a very powerful tool for creating the bioclimatic rating of Cyprus based on the Ombrothermic Index. Note also that the map of the mean is similar that the one we could obtain using multiple linear regression followed by ordinary kriging of the regression residuals as in Garzón-Machado et al. (2014), although with our approach we can explore further the behaviour of the indices.

From a biological point of view, also note that the mean map in Figure 3 also properly reflects zones with higher altitude (corresponding to larger values of the index), and those areas with the highest rainfall. Indeed, the predicted map obtained shows the landscape changes that can be observed in any orthophoto of the island.

Once we have the posterior predictive distribution of the index we can use it to obtain the maps of the spatial bioclimatic probability distribution introduced in the previous section. As mentioned above, these maps show the posterior probability of an index belonging to each subtype.

Figure 4 shows the posterior probability of the nine possible ombrotypes (categories of the Ombrothermic Index) that can be observed in the Mediterranean bioclimate. The figure represents the probability of one location on the island belonging to each ombrotype. Note that in Cyprus the subtypes Hyperhumid, Ultrahyperhumid, Arid, Hyperarid and Ultrahyperarid are not possible, while probabilities greater than zero indicate that Humid and Subhumid are possible at the highest altitudes, and the Dry subtype is possible on the coast and Semiarid in the north and east.

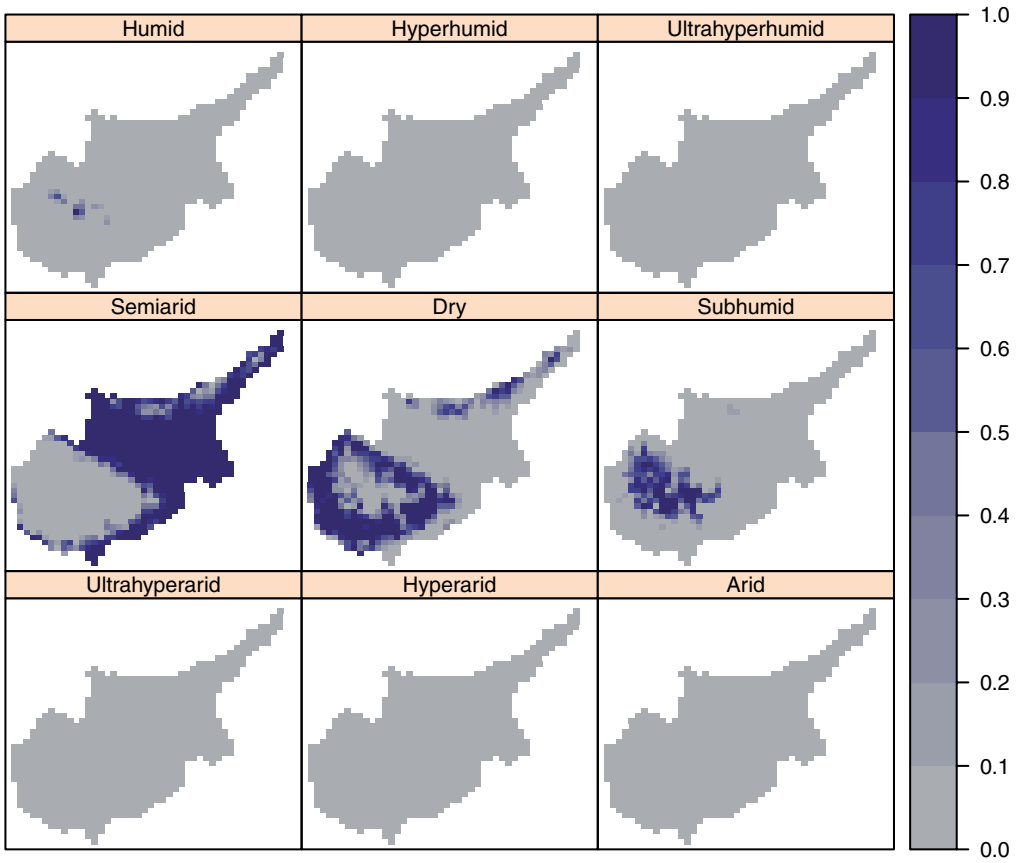

Figure 4: *Spatial bioclimatic probability distribution of the Ombrothermic Index.*

As it can be seen from the figure, there is a high probability of the Humid subtype being found in the two mountain peaks of the Central mountains and of the Subhumid subtype occurring in the mountainous area of the central mountains. The Dry subtype has a high probability of occurring on the hillsides of those peaks and the northern ridge of the island and finally the Semiarid subtype is likely to be found in the central plateau. It is worth noting how important these probability distribution maps are from a biological point of view, as they provide more accurate information on the subtype boundaries, by using a gradient map showing the border from one subtype to another.

5.2. Thermicity Index

We now show the results for the Thermicity Index in Cyprus. This index presents a peculiar relationship with the orography, and obviously with the temperature-altitude pair, i.e., higher altitude is associated with lower temperature.

Table 3: *Median of the posterior distribution and 95% credible intervals of the parameters for the Thermicity Index model.*

Parameters	Median	$p_{2.5}$	$p_{97.5}$
β_0	6.10	6.06	6.13
β_1	-5.48×10^{-4}	-6.04×10^{-4}	-4.92×10^{-4}
ξ^2	7.90×10^{-3}	5.53×10^{-3}	1.21×10^{-2}
κ	3.95×10^{-1}	1.94×10^{-2}	9.11×10^{-1}
ϕ	3.12×10^{-4}	2.52×10^{-5}	5.40×10^{-4}
ν	0.594	0.0763	1.75

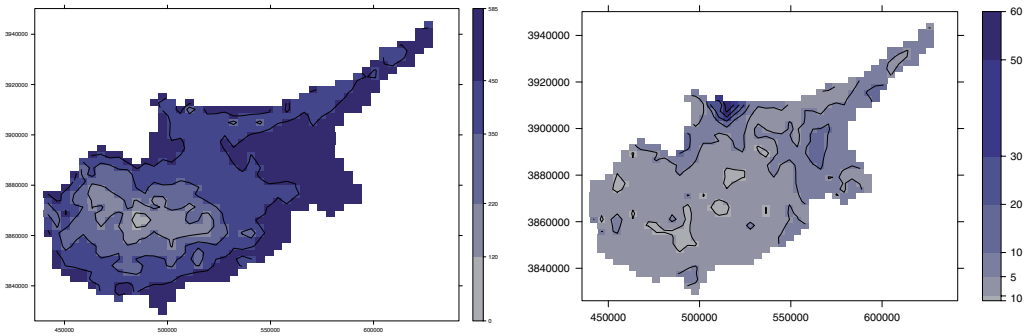

Figure 5: *Mean (left) and standard deviation (right) of the posterior predictive distribution of the Thermicity Index.*

Table 3 shows the median of the posterior distribution of the parameters along with their corresponding 95% credible intervals for this index. As above, these posterior distributions were obtained by simulation using WinBUGS, although in this case neither efficiency (in terms of computational time) nor convergence were as good as for the Ombrothermic Index (indeed the number of discard simulations needed in the burn-in was 20000 for this index).

Results for β_1 in Table 3 now show a negative effect on the altitude, which corresponds to the effect in climatology known as the mountain-valley wind effect. The value for κ is around 0.395 with a credible interval that nearly covers the whole $[0,1]$ interval. This clearly indicates that the model can not distinguish between the spatial and non-spatial variabilities. The fact that some weather stations present different values even though they are close to one another, clearly indicates that this index probably does not have a major spatial effect.

Figure 5 shows the mean and the standard deviation of the posterior predictive distribution of the parameters for the Thermicity Index model. The mean map clearly shows the island's mountain system, which is a real factor in explaining the variability for the

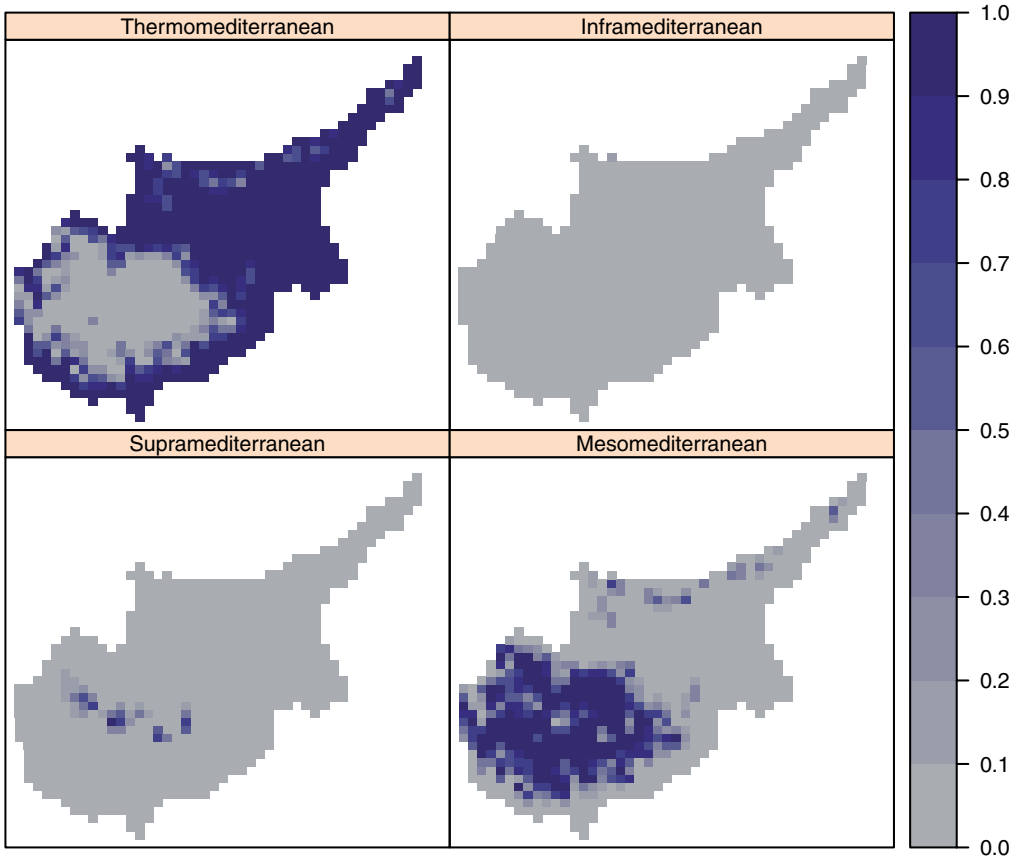

Figure 6: *Spatial bioclimatic probability distribution of the Thermicity Index.*

Thermicity Index, as mentioned previously. Figure 5 also shows the differences between the south and the north of the island and the two principal mountains.

Figure 6 shows the spatial bioclimatic probability distribution of the four possible thermotypes that can be observed in the Mediterranean bioclimate. As can be appreciated from the figure, there is a strong relationship between altitude and thermicity. The Supramediterranean subtype is very likely to be found at the highest locations, while on the hillsides there is a high probability the Mesomediterranean thermotype. Finally, there is a high probability of the ombrotype for the rest of the island beinf Thermomediteranean. Again, this map could be very helpful for landscape management, as it illustrates the vegetation frontiers, due to the close relationship between thermicity and vegetation.

6. Conclusions

In this study, we have introduced a hierarchical Bayesian model that allows us to obtain the spatial distribution of bioclimatic indices by incorporating the altitude and spatial features of each sampled location. Two of the most important advantages of the Bayesian model formulation are that it incorporates parameter uncertainty (both in the inferential and prediction processes), and also prior information can be easily handled. In this context, we have shown how to incorporate our prior knowledge about the parameters via their prior distributions taking into account the particular characteristics of bioclimatic indices. Interestingly, this approach could be easily extended in other contexts. Moreover, sensitivity analysis have shown that there is no dependence on the prior selected.

Also interest is the usefulness of the two main outcomes of the modelling. Posterior predictive distributions reflect most of the information about the bioclimates, but the most valuable information they provide comes from the fact that they inform us of the probability of each location belonging to the different bioclimates. This is done using what we have called the spatial bioclimatic probability distributions. These distributions could be a powerful tool in studies about climate change and its effects on the vegetation of a region, but also in landscape management, in particular to establish future policies or future resource management.

This study also explains how to use MCMC methods, in particular WinBUGS, for the inference in this context, and also how to perform distributed programming for the prediction, which allows us to reduce the computation time.

Another important issue to be mentioned is that in the case that the two analysed indices were related, a joint modelling should be used. In our case, as the Thermicity and Ombrothermic indices are not related there is no need for it, but with other indices the opposite applies and a joint modelling would be needed.

Finally, it should be noted that all the analytical approaches we used here to document the spatial distribution of bioclimatic indices can be applied in any other part of the world.

Acknowledgments

Xavier Barber, David Conesa and Antonio López-Quílez would like to thank the Ministerio de Economía y Competitividad (the Spanish Ministry of Economy and Finance) for its support in the form of the research grant MTM2016-77501-P (jointly financed with the European Regional Development Fund –FEDER–).

References

Adams, N., Kirby, S., Harris, P. and Clegg, D. (1996). A review of parallel processing for statistical computation. *Statistics and Computing*, 6, 37–49.

Atkinson, P. and Tate, N. (2000). Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer*, 52, 607–623.

Baltensperger A. and Huettmann, F. (2015). Predicted shifts in small mammal distributions and biodiversity in the altered future environment of alaska: an open access data and machine learning perspective. *PloS one*, 10, e0132054.

Banerjee, S., Carlin, B. and Gelfand, A. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, Second Edition. Chapman and Hall/CRC, Boca Raton.

Bates, D. and Maechler, M. (2015). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-0.

Blackford, L., Choi, J., Cleary, A., D'Azevedo, E., E., Demmel, J., Dhillon, I., Dongarra, J., Hammarling, S., Henry, G., Petitet, A., Stanley, K., Walker, D. and Whaley, R. (1997). *ScaLAPACK Users' Guide*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia.

Britton, A., Pakeman, R., Carey, P., and Marrs, R. (2001). Impacts of climate, management and nitrogen deposition on the dynamics of lowland heathland. *Journal of Vegetation Science*, 12, 797–806.

Burrough, P. (2001). GIS and geostatistics: essential partners for spatial analysis. *Environmental and Ecological Statistics*, 8, 361–377.

Chambers, R. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597–604.

Cheddadi, R., Guiot, J. and Jolly, D. (2001). The Mediterranean vegetation: what if the atmospheric CO_2 increased? *Landscape Ecology*, 16, 667–675.

Cuenca, J., Giménez, D. and González, J. (2004). Architecture of an automatically tuned linear algebra library. *Parallel Computing*, 30, 187–210.

De Oliveira, V. (2007). Objective Bayesian analysis of spatial data with measurement error. *The Canadian Journal of Statistics*, 35, 1–19.

Dongen, V. (2006). Prior specification in Bayesian statistics: three cautionary tales. *Journal of Theoretical Biology*, 242, 90–100.

Dostálek, J., Frantík, T. and Šilarová, V. (2014). Changes in the distribution of alien plants along roadsides in relation to adjacent land use over the course of 40 years. *Plant Biosystems*, (published online: 20 Dec 2014): 1–17.

Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D. and Ushey, K. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40, 1–18.

Finley, A., Banerjee, S. and Gelfand, A. (2015). spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models. *Journal of Statistical Software*, 63, 1–28.

Gamerman, D. and Lopes, H. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. CRC Press, Boca Raton.

Garzón-Machado, V., Otto, R. and del Arco Aguilar, M.J. (2014). Bioclimatic and vegetation mapping of a topographically complex oceanic island applying different interpolation techniques. *International Journal of Biometeorology*, 58, 887–899.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, Third Edition. Chapman and Hall-CRC, Boca Raton.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*, Third Edition. Johns Hopkins University Press, Baltimore.

Handcock, M. and Wallis, J. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, 89, 368–390.

Lee, P. (1997). *Bayesian Statistics: an Introduction*, Second Edition. Arnold, London.

Legendre, P., Borcard, D. and Peres-Neto, P. (2005). Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*, 75, 435–450.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Matérn, B. (1986). *Spatial Variation*, Second Edition. Springer-Verlag, Berlin.

Osborne, C., Mitchell, P., Sheehy, J. and Woodward, F. (2000). Modelling the recent historical mpacts of atmospheric CO_2 and climate change on Mediterranean vegetation. *Global Change Biology*, 6, 445–458.

Rivas-Martínez, S. (1994). Clasificación bioclimática de la tierra (bioclimatic classification system of the earth). *Folia Botanica Matritensis*, 13, 1–25.

Rivas-Martínez, S. and Rivas-Saenz, S. (2016). Worldwide Bioclimatic Classification System. Phytosociological Research Center. Spain, http://www.globalbioclimatics.org.

Rivas-Martínez, S., Rivas-Sáenz, S., Penas, A. et al. (2002). *Worldwide Bioclimatic Classification System*. Backhuys Pub.

Robertson, G. (1987). Geostatistics in ecology: interpolating with known variance. *Ecology*, 68, 744–748.

Rosenthal, J. (2000). Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics*, 4, 207–236.

Rossi, R., Mulla, D., Journel, A. and Franz, E. (1992). Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, 62, 277–314.

Rossini, A.J., Tierney, L. and Li, N. (2007). Simple parallel statistical computing in R. *Journal of Computational and Graphical Statistics*, 16, 399–20.

Sanderson, C. (2010). *Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments*. NICTA.

Stein, M. (1999). *Interpolation of Spatial Statistics Data: Some Theory for Kriging*. Springer-Verlag, New York.

Strupczewski, W., Kochanek, K., Weglarczyk, S. and Singh, V. (2007). On robustness of large quantile estimates to largest elements of the observation series. *Hydrological Processes*, 21, 1328–1344.

Tasser, E. and Tappeiner, U. (2002). Impact of land use changes on mountain vegetation. *Applied Vegetation Science*, 5, 173–84.

Whiley, M. and Wilson, S. (2004). Parallel algorithms for Markov chain Monte Carlo methods in latent spatial gaussian models. *Statistics and Computing*, 14, 171–179.

Yan, J., Cowles, M., Wang, S. and Armstrong, M. (2007). Parallelizing MCMC for Bayesian spatiotemporal geostatistical models. *Statistics and Computing*, 17, 323–335.

SORT 41 (2) July-December 2017, 297-318 DOI: 10.2436/20.8080.02.61

The Pareto IV power series cure rate model with applications

Diego I. Gallardo[1], Yolanda M. Gómez[2], Barry C. Arnold[3] and Héctor W. Gómez[4]

Abstract

Cutaneous melanoma is thought to be triggered by intense, occasional exposure to ultraviolet radiation, either from the sun or tanning beds, especially in people who are genetically predisposed to the disease. When skin cells are damaged by ultraviolet light in this way, often showing up as a sunburn, they are more prone to genetic defects that cause them to rapidly multiply and form potentially fatal (malignant) tumors. Melanoma originates in a type of skin cell called a melanocyte, such cells help produce the pigments of our skin, hair, and eyes. We propose a new cure rate survival regression model for predicting cutaneous melanoma. We assume that the unknown number of competing causes that can influence the survival time is governed by a power series distribution and that the time until the tumor cells are activated follows the Pareto IV distribution. The parameter estimation is based on the EM algorithm which for this model can be implemented in a simple way in computational terms. Simulation studies are presented, showing the good performance of the proposed estimation procedure. Finally, two real applications related to a cutaneous melanoma and melanoma data sets are presented.

MSC: 62N01, 62N02, 62P10.

Keywords: Competing risks, cure rate models, EM algorithm, Pareto IV distribution, power series distribution.

1. Introduction

Cancer is a process of uncontrolled growth and dissemination of cells. It can occur in practically any location in the body. The tumor can invade the neighbouring region of the body and can also provoke metastasis in parts of the body remote from the original site. Many types of cancer can be prevented by avoiding exposure to common risk factors

[1] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile. diego.gallardo@uda.cl (Corresponding author).

[2] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile. yolanda.gomez@uda.cl

[3] Statistics Department, University of California, Riverside, CA, USA. barry.arnold@ucr.edu

[4] Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile. hector.gomez@uantof.cl

Received: February 2016

Accepted: April 2017

such as, for example, tobacco smoke. Moreover, a major proportion of cancers can be cured by surgery, chemotherapy or radiation, especially if they are detected at an early stage. Melanoma that occurs on the skin, called cutaneous melanoma, is the most common type of melanoma. This type of melanoma occurs in all parts of the skin, including the soles of feet, on the palms of the hand, in between toes and fingers, and underneath the finger and toe nails.

Skin melanoma occurs most frequently in people with a light complexion, since they are least protected against UV radiation. Also, people with more than 50 moles, a family history of melanoma, a weakened immune system, or those who sunbathe or use tanning beds, are at increased risk. Melanoma is the fastest growing cancer in men and the second fastest growing cancer in women (after lung cancer).

Regression models for survival data with a surviving fraction (also known as cure rate models or long-term survival models) play an important role in reliability and survival analysis. These models typically assume that all units under study are susceptible to an event of interest and will eventually experience it if follow-up is sufficiently long. However, there are situations in which a fraction of individuals are not expected to experience the event of interest, that is, those individuals are cured or not susceptible. For example, researchers may be interested in analysing the recurrence of a disease. Many individuals may never experience a recurrence; therefore, a cured fraction of the population exists. Cure rate models have been applied to investigate the possible existence of a cured fraction. An approach for those models is the following.

Let M be a random variable denoting the initial number of carcinogenic cells of an individual. Several different assumptions about the probability mass function of M have appeared in the literature: Bernoulli (Berkson and Gage, 1952), Poisson (Yakolev and Tsodikov, 1996), Negative Binomial (Rodrigues et al., 2009a), among others. A generalization that includes all these models is the power series distribution (Noack, 1950) used by Cancho, Louzada and Ortega (2013a) in the cure rate context. Evidently this model doesn't include all distributions that can be used in this context (see for instance, Rodrigues et al., 2009b and Rodrigues et al., 2015).

On the other hand, let W_a be a random variable expressing the time at which the a-th cell produces a detectable cancer. In their proposal, Cancho et al. (2013a) used the Weibull distribution. Other approaches include the generalized gamma (Ortega et al., 2014), the Beta-Weibull (Ortega et al., 2015) and the Birnbaum-Saunders distribution (Cordeiro et al., 2016). Our proposal is one in which we assume for each W_a a Pareto IV distribution (Arnold 1983, 2015). This is a very flexible model which includes some interesting distributions as particular cases and which has the characteristic that both, the survival and density functions, have analytic tractable forms.

The sections of this paper are organized in the following manner. In Section 2, we explain the model formulation and give some of its main properties. In Section 3, we develop parameter estimation for the model based on the EM algorithm. In Section 4, two real data applications are discussed. In Section 5, a simulation study is presented. Finally, some conclusions are given in Section 6.

2. The Pareto IV power series cure rate model

The model proposed in Cancho et al. (2013a) can be defined as follows. Let M be a random variable denoting the initial number of carcinogenic cells of an individual with probability mass function given as in Noack (1950) by

$$P(M=m;\theta)=\frac{a_m\theta^m}{A(\theta)},\quad m=0,1,2,\ldots, \tag{1}$$

where $a_m>0$ and $A(\theta)=\sum_{m=0}^{\infty}a_m\theta^m$. θ is the so-called power parameter of the distribution and $A(\theta)$ is the series function. We denote the distribution in (1) by $\mathrm{PS}(\theta,A(\theta))$. Table 1 shows some particular cases of this distribution. Θ denotes the parameter space for θ in each model.

Table 1: *Some particular cases of* $PS(\theta,A(\theta))$.

Distribution	a_m	$A(\theta)$	$E_\theta[M^d]$, $d=1,2$	Θ
Poisson	$(m!)^{-1}$	e^θ	$\theta+(d-1)\theta^2$	$(0,\infty)$
Logarithmic	$(m+1)^{-1}$	$-\frac{\log(1-\theta)}{\theta}$	$1-\frac{\theta}{(1-\theta)\log(1-\theta)}\left(\frac{3+2\theta}{1-\theta}\right)^{d-1}$	$(0,1)$
Negative Binomial	$\binom{m+q-1}{m}$	$(1-\theta)^{-q}$	$\left(\frac{\theta}{1-\theta}\right)\left(\frac{1+q\theta}{1-\theta}\right)^{d-1}$	$(0,1)$
Binomial	$\binom{q}{m}$	$(1+\theta)^q$	$q\left(\frac{\theta}{1+\theta}\right)\left(\frac{q\theta^2+(q+1)\theta+1}{q(1+\theta)}\right)^{d-1}$	$(0,\infty)$

Note: We denote those distributions as $\mathrm{Po}(\theta)$, $\mathrm{Lo}(\theta)$, $\mathrm{NB}(q,\theta)$ and $\mathrm{Bin}(q,\theta)$ respectively. In both, $\mathrm{NB}(q,\theta)$ and $\mathrm{Bin}(q,\theta)$, q is considered known.

Denote by W_a the random variable representing the time at which the a-th cell produces a detectable cancer. For non-cured patients, $M>0$ and $W_a,\ a=1,2,\ldots,M$, are conditionally independent given M and identically distributed with common cumulative distribution and survival functions $F(t;\lambda)$ and $S(t;\lambda)=1-F(t;\lambda)$, where λ is a vector of unknown parameters. For cured patients, $M=0$ and it is assumed that $P(W_0=\infty)=1$. The distribution F is assumed to be a proper distribution function. The time until the event of interest depends upon the count variable (M) and the survival time variables $(W_1,\ldots,W_M)$ and can be expressed by $T=\min\{W_a,\ 0\le a\le M\}$. As mentioned by Cancho et al. (2013a), it can be verified that the survival function for T (also known as population survival function) is given by

$$S_{pop}(t;\theta,\lambda)=\frac{A(\theta S(t;\lambda))}{A(\theta)}. \tag{2}$$

From (2), it is possible to verify that the cure fraction of the model is $p_0 = A(0)/A(\theta) = a_0/A(\theta)$ and the corresponding density function for (2) is given by

$$f_{pop}(t;\theta,\lambda) = \frac{A'(\theta S(t;\lambda))}{A(\theta)}\theta f(t;\lambda),$$

where $A'(\eta) = \frac{\partial A(\eta)}{\partial \eta}$ and $f(t;\lambda)$ is the density function corresponding to time till the event of interest for each of the carcinogenic cells W_a.

The Weibull distribution is extensively used in survival analysis because it explains biological processes relatively well and because it is a distribution that is easy to work with. For these reasons, Cancho et al. (2013a) considered this distribution.

However, the Pareto IV distribution is more flexible than the Weibull distribution and is not markedly more difficult to work with in the cure rate models context. For this reason, we propose to use the Pareto IV distribution for modeling the time until the activation of the carcinogenic cells.

The Pareto IV distribution (Arnold 1983, 2015) is very flexible and has the convenient feature that its survival function is available in a simple analytic form. Let W be a random variable with a Pareto IV distribution and corresponding vector of parameters $(\mu,\sigma,\gamma,\alpha)$. (We denote this by $W \sim \text{P4}(\mu,\sigma,\gamma,\alpha)$). The survival function of W is

$$S(w;\mu,\sigma,\gamma,\alpha) = \left[1+\left(\frac{w-\mu}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}, \quad w>\mu, \mu\in R, \sigma,\gamma,\alpha>0,$$

with the corresponding density function

$$f(w;\mu,\sigma,\gamma,\alpha) = \frac{\alpha}{\gamma\sigma}\left[1+\left(\frac{w-\mu}{\sigma}\right)^{1/\gamma}\right]^{-\alpha-1}\left(\frac{w-\mu}{\sigma}\right)^{1/\gamma-1}, w>\mu, \mu\in R, \sigma,\gamma,\alpha>0.$$

The s-th moment of this distribution is given by

$$E(W^s) = \frac{\sigma^s\Gamma(\alpha-\gamma s)\Gamma(1+\gamma s)}{\Gamma(\alpha)}, \quad \text{if } -1<\gamma s<\alpha, \tag{3}$$

and the pth quantile, say w_p, is given by

$$w_p = \sigma(p^{-1/\alpha}-1)^{\gamma}, \quad 0<p<1. \tag{4}$$

Since we are working in a context of positive variables which are not bounded away from 0, we fix $\mu = 0$. Thus, the parameter vector related to the initial concurrent causes are defined by $\boldsymbol{\lambda} = (\sigma,\gamma,\alpha)$.

Some particular cases of this distribution are the following

- $\gamma = 1$: The Pareto II distribution (P2) also known as Lomax distribution.
- $\alpha = 1$: The Pareto III distribution (P3).

Since these models are particular cases of the P4 distribution, it is possible to use, for instance, likelihood ratio tests to decide between the hypothesis $H_0 : \gamma = 1$ ($\alpha = 1$) and $H_1 : \gamma \neq 1$ ($\alpha \neq 1$).

The model in (2) in which we assume that $S(\cdot;\boldsymbol{\lambda})$ is the survival function of a P4 distribution will be called the Pareto IV Power series cure rate model (henceforth, P4PS). Below we describe some particular cases of this model.

- **The Binomial Pareto IV (BP4) model.** If $A(\theta) = (1+\theta)^q$, then $M \sim \text{Bin}(q,\theta)$. Note that q is a positive integer that can be interpreted as the maximum number of carcinogenic cells for each individual. The cure rate is $p_0 = (1+\theta)^{-1}$. The case $q = 1$ ($M \sim \text{Bernoulli}(\theta)$) corresponds to the first survival model with cure rate in the literature (*the mixture model*) proposed in Berkson and Gage (1952). The population survival function of the BP4 model is

$$S_{pop}(t;\theta,\lambda) = \left(\frac{1+\theta\left[1+\left(\frac{t}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}}{1+\theta} \right)^q .$$

- **The Poisson Pareto IV (PP4) model.** If $A(\theta) = e^{\theta}$, then $M \sim \text{Po}(\theta)$. This is the same assumption used in Yakolev and Tsodikov (1996), the so-called *promotion time cure rate model* and it is the only cure rate model with proportional hazard structure (see Theorem 5 in Rodrigues et al., 2009a). The cure rate of the model is $p_0 = e^{-\theta}$. The population survival function is

$$S_{pop}(t;\theta,\lambda) = \exp\left\{ -\theta\left(1 - \left[1+\left(\frac{t}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}\right)\right\}.$$

- **The Negative Binomial Pareto IV model.** If $A(\theta) = (1-\theta)^{-q}$, then $M \sim \text{NB}(q,\theta)$. Here, typically, q is a positive integer although the definition remains valid if q is any positive real number. The Negative Binomial distribution includes the Poisson distribution as a limiting case. Moreover an extended definition of the Negative Binomial distribution (introduced by Piegorsch, 1990) allowing q to be negative permits one to view the binomial and Bernoulli distributions as particular cases. This observation was used in Rodrigues et al. (2009a) in unifying the *mixture model* and the *promotion time cure rate model* (the most popular cure rate model

until then). The cure rate is given by $p_0 = (1-\theta)^q$. The particular case $q = 1$, i.e., when M has a Geometric distribution, it is usually used in literature. (For instance, Cancho, Louzada and Barriga, 2013b and Gómez and Bolfarine, 2016). The population survival function is

$$S_{pop}(t;\theta,\boldsymbol{\lambda}) = \left(\frac{1-\theta}{1-\theta\left[1+\left(\frac{t}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}} \right)^q.$$

- **The Logarithmic Pareto IV model.** If $A(\theta) = -\theta^{-1}\log(1-\theta)$, then $M \sim \text{Lo}(\theta)$ in contrast to the other models, the mode of M in this case is zero implying that the probabilities for M are decreasing. The cure rate is given by $p_0 = -\theta/\log(1-\theta)$. This is not a very common model in literature. The population survival function is

$$S_{pop}(t;\theta,\boldsymbol{\lambda}) = \frac{\log\left(1-\theta\left[1+\left(\frac{t}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}\right)}{\left[1+\left(\frac{t}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}\log(1-\theta)}.$$

3. Estimation

In this section, we discuss the estimation for the P4PS cure rate model using a classical approach. Assume that the data are obtained with right censoring. Thus, the observed data for the i-th individual can be represented by $T_i = \min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leq C_i)$, $1,\ldots,n$, where T_i^* and C_i denote failure and censoring times respectively. Denote the observed data by $D_{obs} = (\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{z})$, with $\boldsymbol{t} = (t_1,\ldots,t_n)^\intercal$, $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_n)^\intercal$ and $\boldsymbol{z} = (z_i,\ldots,z_n)^\intercal$, where $\boldsymbol{z}_i$ is a vector of covariates (of dimension $r \times 1$) related to the cure of the i-th individual. For each individual, those covariates can be introduced into the model by allowing the parameter θ to depend on the covariates in the following manner,

$$\theta_i = \begin{cases} \exp(\boldsymbol{z}_i^\intercal\boldsymbol{\beta}) & \text{for the Poisson and Binomial models} \\ \dfrac{\exp(\boldsymbol{z}_i^\intercal\boldsymbol{\beta})}{1+\exp\{\boldsymbol{z}_i^\intercal\boldsymbol{\beta}\}} & \text{for the Logarithmic and Negative Binomial models} \end{cases} \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1,\beta_2,\ldots,\beta_r)^\intercal$ is a vector of parameters of dimension r. Note that this specification guarantees the identifiability of the model in the sense of Li et al. (2001) and Hanin and Li-Shang (2014).

On the other hand, note that the vector $\boldsymbol{M} = (M_1,\ldots,M_n)$ is non-observable and thus the complete data are $D_{comp} = (\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{z},\boldsymbol{M})$. In Cancho et al. (2013a), the estimation procedure for $\boldsymbol{\psi} = (\boldsymbol{\beta},\boldsymbol{\lambda})$ was performed maximizing the observed likelihood, i.e., maximizing the following expression

$$\ell(\boldsymbol{\psi} \mid D_{obs}) = \sum_{i=1}^{n} \left[\delta_i \log f_{pop}(t_i;\boldsymbol{\psi}) + (1-\delta_i)\log S_{pop}(t_i;\boldsymbol{\psi})\right] = \sum_{i=1}^{n} \Big[\log A(\theta_i S(t_i;\boldsymbol{\lambda})) + \delta_i \left(\log\theta_i + \log f(t_i;\boldsymbol{\lambda}) + \log A'(\theta_i S(t_i;\boldsymbol{\lambda})) - \log A(\theta_i S(t_i;\boldsymbol{\lambda}))\right) - \log A(\theta_i)\Big]. \tag{6}$$

However, the maximization of $\ell(\cdot)$ can be difficult because there are many parameters, especially when the number of covariates that are used is high. For this reason, in a cure rate model context there are many proposals based on the EM algorithm (see for instance, Gallardo, Bolfarine and Pedroso-de-Lima, 2016a; Gallardo and Bolfarine, 2016b; Gallardo, Romeo and Meyer, 2016c and Pal and Balakrishnan, 2016). Particu-

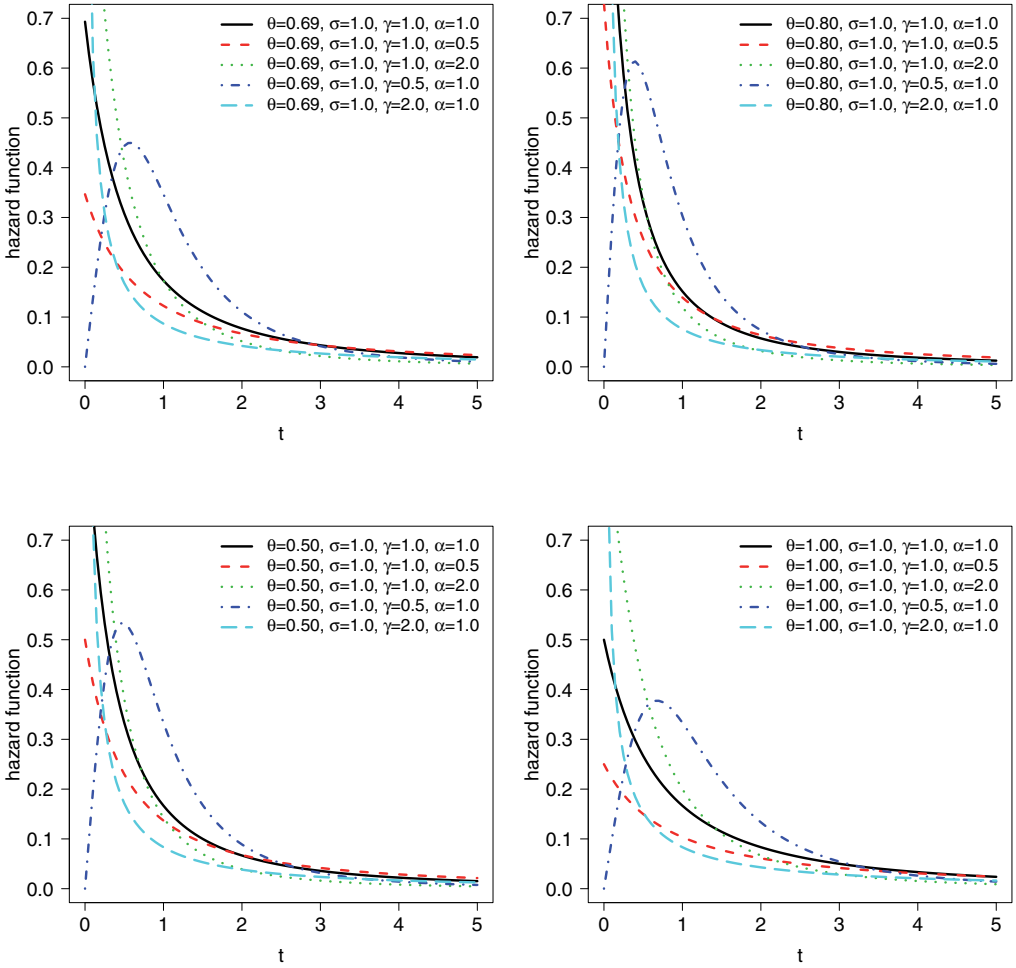

Figure 1: *Population hazard function for P4PS model with different parameters and cure rate fixed at 50%. Left upper: Poisson. Right upper: Logarithmic. Left lower: Negative Binomial ($q = 1$). Right lower: Binomial ($q = 1$).*

larly, we follows a similar scheme that Gallardo et al. (2016c) and we omit technical details about the method.

The k-th iteration of the algorithm (assuming q is known in the Binomial and Negative Binomial cases) takes the form:

- **E-step**: Define $\mu_i^{(k)} = \theta_i^{(k)} S(t_i;\boldsymbol{\lambda}^{(k)})$ and $\kappa_i^{(k)} = \left(1 - \frac{\mu_i^{(k)}}{\left(1-\mu_i^{(k)}\right)\log\left(1-\mu_i^{(k)}\right)}\right)$ and compute for $i = 1,\ldots,n$,

$$\widetilde{M}_i^{(k)} = \begin{cases} \delta_i + \mu_i^{(k-1)} & \text{for Poisson model} \\ (1-\delta_i)\kappa_i^{(k-1)} + \delta_i \frac{\left(1-\mu_i^{(k-1)}\right)^2 \log\left(1-\mu_i^{(k-1)}\right) - \mu_i^{(k-1)}\left(3+2\mu_i^{(k-1)}\right)}{\left(1-\mu_i^{(k-1)}\right)\log\mu_i^{(k-1)} - \mu_i^{(k-1)}} & \text{for Logarithmic model} \\ \frac{\delta_i + \mu_i^{(k-1)} + (q-1)\delta_i\mu_i^{(k-1)}}{1-\mu_i^{(k-1)}} & \text{for NB model} \\ (1-\delta_i)q\left(\frac{\mu_i^{(k-1)}}{1+\mu_i^{(k-1)}}\right) + \delta_i\left(\frac{q\left[\mu_i^{(k-1)}\right]^2 + (q+1)\mu_i^{(k-1)} + 1}{q\left(1+\mu_i^{(k-1)}\right)}\right) & \text{for Binomial model} \end{cases}$$

- **CM-step I**: Using $\boldsymbol{M}^{(k)} = (M_1^{(k)},\ldots,M_n^{(k)})$ obtained previously in the E-step, update $\widehat{\beta}^{(k)}$ maximizing

$$Q_1(\beta \mid \boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n}\left[\widetilde{M}_i^{(k)}\log\theta_i - \log A(\theta_i)\right]$$

  with respect to β.

- **CM-step II**: Update $\widehat{\alpha}^{(k)}$ as follows

$$\widehat{\alpha}^{(k)} = \frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n} M_i^{(k)}\log\left(1+\left(\frac{t_i}{\widehat{\sigma}^{(k-1)}}\right)^{\frac{1}{\widehat{\gamma}^{(k-1)}}}\right)}$$

- **CM-step III**: With $\boldsymbol{M}^{(k)}, \alpha^{(k)}$ and $\gamma^{(k-1)}$, update $\widehat{\sigma}^{(k)}$ solving the following nonlinear equation for σ

$$\sum_{i=1}^{n}\left[\frac{\left(\alpha^{(k)}M_i^{(k)} + 2\delta_i\right)\left(\frac{t_i}{\sigma}\right)^{1/\gamma^{(k-1)}} + \delta_i}{\left(1+\left(\frac{t_i}{\sigma}\right)^{1/\gamma^{(k-1)}}\right)}\right] = 0$$

- **CM-step IV**: With $\boldsymbol{M}^{(k)}, \alpha^{(k)}$ and $\sigma^{(k)}$, update $\widehat{\gamma}^{(k)}$ solving the following non-linear equation for γ

$$\sum_{i=1}^{n} \frac{\left(\frac{t_i}{\sigma^{(k)}}\right)^{1/\gamma}\left(2\delta_i \log\left(\sigma^{(k)}\right) - \alpha^{(k)} M_i^{(k)} \log\left(\frac{t_i}{\sigma^{(k)}}\right) + \gamma\delta_i\right) + \delta_i\left(\log\left(\sigma^{(k)} t_i\right) + \gamma^{(k)}\right)}{\left(1+\left(\frac{t_i}{\sigma^{(k)}}\right)^{1/\gamma^{(k)}}\right)} = 0.$$

The E and CM-I/CM-IV steps are alternated repeatedly until a suitable convergence rule is satisfied, e.g., the difference in successive values of the estimates is less than a tolerance value. The variance of $(\boldsymbol{\beta}, \alpha, \sigma, \gamma)$ can be estimated based on the inverse of minus the hessian matrix of the model. Details about this matrix can be seen in the additional material.

Finally, for the binomial and the negative binomial distributions for which q will typically be unknown, we can consider a grid of values for q, say $\mathfrak{Q} = q_1, q_2, \ldots, q_B$ and we apply the EM algorithm for each value in $\mathfrak{Q}$, obtaining for each q_j, $j = 1, \ldots, B$, a set of estimates parameters, say $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots, \widehat{\boldsymbol{\psi}}_B$. Then, we choose $q = q_b$ as the value in $\mathfrak{Q}$ such that

$$\max_{j=1,\ldots,B} \ell(\widehat{\boldsymbol{\psi}}_j \mid D_{obs}) = \ell(\widehat{\boldsymbol{\psi}}_b \mid D_{obs}),$$

where $\ell(\cdot)$ is the observed likelihood function defined in (6).

3.1. Interpreting the parameters

We highlight that, up till now, we have been unable to find in the literature any work where the regression coefficients are interpreted in a cure rate model context, except in the case in which $M_i \sim \text{Bin}(1, \theta)$ corresponding to the *mixture model*. In that setting the coefficients can be interpreted in terms of the log-odds ratio, similar to the case of logistic regression for dichotomic responses.

In general, efforts to interpret the coefficients are limited to illustrating the behaviour in the cure rate when varying a continuous covariate and fixing the others (as we shall illustrate this issue in the application Section). To this end, we propose the following methodology. Note that, based on a Taylor expansion of the first order around the intercept (or another convenient point) of the logarithm of the cure rate, we can write $q_{0i} \approx \exp\{a_0 + b_0 \boldsymbol{z}_i^{\top}\boldsymbol{\beta}\}$, where a_0 and b_0 depends on the respective model and the value for the intercept. If $\boldsymbol{z}_{i(j)}$ represents the $\boldsymbol{z}_i$ vector with the j-th element increased in 1 unit, then the ratio between $q_{0i(j)}$ and q_{0i} is

$$\frac{q_{0i(j)}}{q_{0i}} \approx \frac{\exp\{a_0 + b_0 \boldsymbol{z}_{i(j)}^{\top}\boldsymbol{\beta}\}}{\exp\{a_0 + b_0 \boldsymbol{z}_i^{\top}\boldsymbol{\beta}\}} = \exp\{b_0 \beta_j\},$$

providing an approximate way to interpret the β_j's in terms of the percentage increment (or decrease) in the cure rate, maintaining the rest of the covariates fixed.

Finally, in relation to the vector $\boldsymbol{\lambda}$, rather than interpreting each component it may be of more interest to evaluate descriptive measure related to the distribution of W_a's. For instance, mean, variance and quantiles can be obtained using (3) or (4). Confidence intervals can also be constructed for those quantities using the delta method Sen, Singer and Pedroso-de-Lima (2010).

4. Applications

In this section we consider two applications of the PSP4 model to real data sets.

4.1. Cutaneous melanoma data set

This data set refers to patients involved in a Phase III cutaneous melanoma clinical trial presented in Ibrahim, Chen and Sinha (2001) and is available at `http://merlot.stat.uconn.edu/~mhchen/survbook/`, labeled as E1690 data. The data set comes from a clinical trial for the evaluation of postoperative treatment performance with a high dose of the drug interferon alpha-2b in order to prevent recurrence. Patients were included in the study from 1991 to 1995, and follow-up was conducted until 1998. The response is considered to be the relapse-free survival time (in years). The data set includes information on 408 patients, for each of which the following covariates were measured: treatment (0: placebo, 198 patients; 1: interferon alpha-2b, 210 patients); tumor thickness (in mm, mean = 3.98 and standard deviation = 3.22) and nodal category (1: 110 patients; 2: 131 patients; 3: 86 patients; 4: 81 patients). Figure 2 shows the Kaplan-Meier (KM)

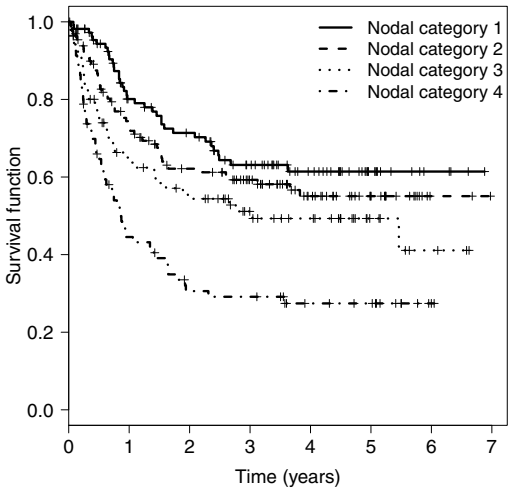

Figure 2: *KM estimator by nodal category for Phase III cutaneous melanoma clinical trial.*

estimator of the survival function by nodal category. As expected, the survival function decrease faster in more advanced categories. However, in all cases the survival function is stabilized at a certain value, suggesting that there is a proportion of patients for whom the malignant melanoma will never recur (in all nodal categories).

We fit the P2PS, P3PS and P4PS model for four particular cases. Model selection was performed based on the AIC and SBC criteria (Akaike, 1974 and Schwarz, 1978, respectively). Those criteria are presented in Table 2. We also fit the gamma and Birnbaum-Saunders (BS) PS model for the concurrent causes. The Birnbaum-Saunders model has been the subject of intense research in cure rate models in recent years. For instance, Cancho et al. (2013b) and Cordeiro et al. (2016).

Table 2: *AIC and SBC criterion for power series cure rate model with Pareto IV and BS distribution for concurrent causes.*

	P2	P3	P4	Gamma	BS
Poisson	841.42/873.51	835.66/867.75	837.23/873.33	837.96/874.06	888.27/924.37
Logarithmic	849.52/881.61	**826.26/858.35**	827.78/863.89	828.10/864.20	927.71/963.81
Geometric	841.70/873.79	830.14/862.23	831.64/867.74	831.76/867.87	907.44/943.54
Binomial	844.87/876.96	840.33/872.42	842.38/878.48	845.14/881.24	875.33/911.43

Both criterion suggest that the Logarithmic cure rate model with a Pareto III distribution for the concurrent causes is the best model. For this model, we also tested the hypotheses $H_0 : \alpha = 1$ versus $H_1 : \alpha \neq 1$ using the log-likelihood ratio (LR) test and the Wald test. In both cases, we failed to reject the null hypothesis at the 5% of significance and consequently we prefer the P3 instead of the P4 distribution for the time-to-event in the concurrent causes.

Estimates of the parameters of the selected model, i.e., the Logarithmic P3 cure rate model, are presented in Table 3. Based on the Taylor expansion of first order (around zero in this case) discussed in Section 3.1 for the logarithmic model, we obtain $b_0 \approx -0.1596685$. For this reason, we present the following approximated interpretations for the regression coefficients:

- $\exp\left(b_0 \times (\widehat{\beta}_{nodule1} - \widehat{\beta}_{nodule2})\right) = 1.193$, i.e., the cure rate for patients with nodule in stage one is 19.3% greater than the cure rate for patients in stage two.
- $\exp\left(b_0 \times (\widehat{\beta}_{nodule1} - \widehat{\beta}_{nodule3})\right) = 1.342$, i.e., the cure rate for patients with nodule in stage one is 34.2% greater than the cure rate for patients in stage three.
- $\exp\left(b_0 \times (\widehat{\beta}_{nodule1} - \widehat{\beta}_{nodule3})\right) = 1.624$, i.e., the cure rate for patients with nodule in stage one is 62.4% greater than the cure rate for patients in stage four.
- $\exp\left(-b_0 \times \widehat{\beta}_{thickness}\right) = 1.019$, i.e., for each mm that is increased the tumor thickness the cure rate is decreased in 1.9%.

- $\exp\left(b_0 \times \widehat{\beta}_{treatment}\right) = 1.079$, i.e., the cure rate for patients receiving treatment is 7.9% greater than the cure rate for patients witouth treatment.

On the other hand, the mean and median of the time-to-event of carcinogenic cells are 3.97 and 1.82 years respectively with their respective 95% confidence intervals (1.74,6.21) and (1.18,2.46).

Table 3: *Estimates, standard errors (s.e.) and 95% confidence interval for logarithmic P3 cure rate model for Phase III cutaneous melanoma clinical trial.*

Parameter	estimate	s.e.	95% Conf. Interval	
$\beta_{nodule1}$	0.2471	0.2584	−0.2594	0.7536
$\beta_{nodule2}$	1.3547	0.1658	1.0296	1.6797
$\beta_{nodule3}$	2.0878	0.2186	1.6593	2.5163
$\beta_{nodule4}$	3.2853	0.2736	2.7491	3.8216
$\beta_{thickness}$	0.1178	0.0034	0.1111	0.1245
$\beta_{treatment}$	−0.4738	0.0973	−0.6645	−0.2832
σ	1.8368	0.1126	1.6161	2.0575
γ	0.6415	0.0023	0.6370	0.6460

We also show in Figure 3 some plots showing the cure rate in terms of tumor thickness for combinations of nodule and treatment.

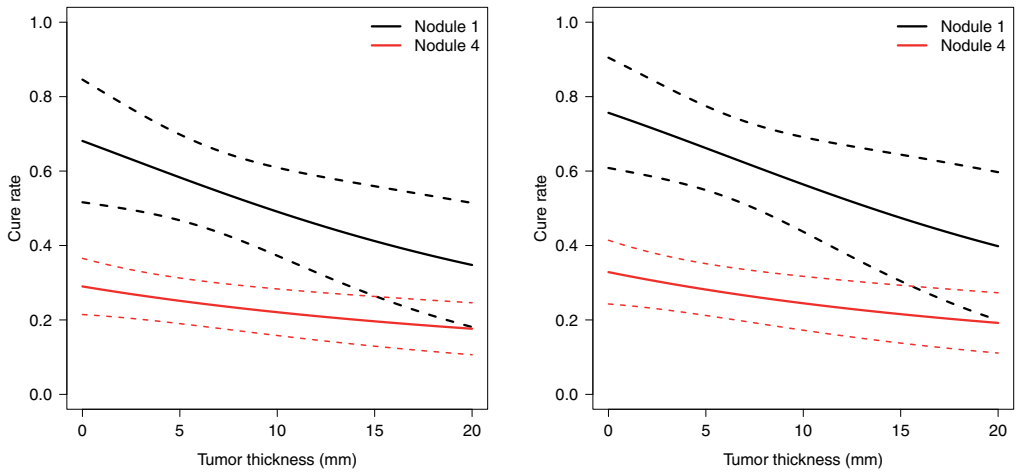

Figure 3: *Estimated cure rate for patients that received and not received treatment (left and right panel respectively) and nodule in stage 1 and 4. The continuous line represent the point estimation and the dashed line represent the respective 95% confidence interval.*

Additionally, in order to analyse possible influential observations, we compute the jackknife residuals defined by

$$J_i = \left(\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}_{(i)}\right)^{\mathsf{T}} \widehat{\boldsymbol{H}}^{-1} \left(\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}_{(i)}\right), \quad i = 1, \ldots, n.$$

where $\boldsymbol{\psi}_{(i)}$ represents the estimator of $\boldsymbol{\psi}$ without the i-th observation. Figure 4 show these residuals. Note that observation 11 is a potentially influential observation. This observation corresponded to an individual with a nodule in stage 1 who received treatment. Table 4 show a descriptive comparison of this observation with the others in same nodule stage and with treatment. Observation 11 was a patient who died in a short time when compared with others patients in similar conditions. Also his tumor thickness was very big in relation to other patients in similar conditions.

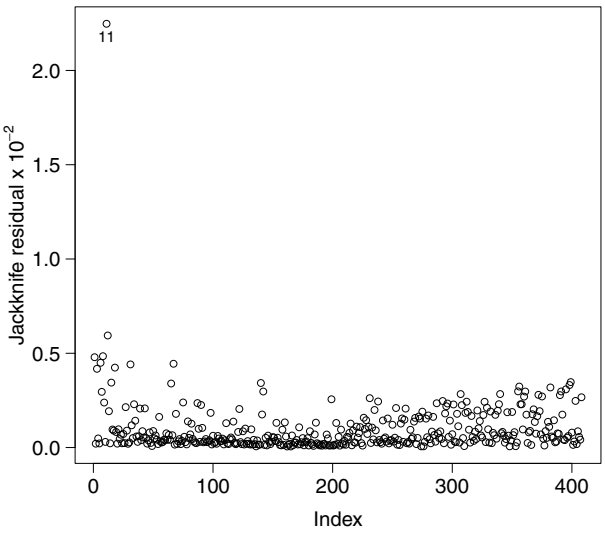

Figure 4: *Jackknife residuals for cutaneous melanoma data set.*

Table 4: *Descriptive analysis for observation 11.*

Observation	t_i	δ_i	thickness
11	0.0767	1	14.000
Mean*	5.900	0.34	5.900
Median*	2.437	0.00	6.611

*Considering the 56 observations in stage 1 that received treatment.

Finally, Table 5 shows the estimates for all parameters with observation 11 deleted from the data set. Note that the magnitudes of the estimates are different from thecorresponding values in Table 3. However, the significance and the sense of all parameter estimates is maintained.

Table 5: *Estimates, standard errors (s.e.) and 95% confidence interval for logarithmic P3 cure rate model for Phase III cutaneous melanoma clinical trial without observation 11.*

Parameter	estimate	s.e.	95% Conf. Interval	
$\beta_{nodule1}$	0.3335	0.2516	−0.1597	0.8267
$\beta_{nodule2}$	1.4401	0.1674	1.1120	1.7681
$\beta_{nodule3}$	2.1708	0.2216	1.7365	2.6050
$\beta_{nodule4}$	3.3909	0.2787	2.8446	3.9372
$\beta_{thickness}$	0.0989	0.0032	0.0926	0.1053
$\beta_{treatment}$	−0.5190	0.0980	−0.7110	−0.3269
σ	1.8228	0.1071	1.613	2.0327
γ	0.6336	0.0022	0.6292	0.638

4.2. Melanoma data set

This data set is available at `timereg` package in R Scheike (2015). The data set refers to 205 patients with malignant melanoma, followed up after removing the lesions. The following covariates were measured: ulceration (absent: 115 patients; present: 90 patients); tumor thickness (in mm, mean = 2.92 and standard deviation = 2.96). Figure 5 shows the KM estimator by ulceration status. Note that the survival function is lower for patients with ulceration. On the other hand, the survival function is stabilized at a certain value, suggesting in this study also the existence of a proportion of patients for whom the malignant melanoma will never recur.

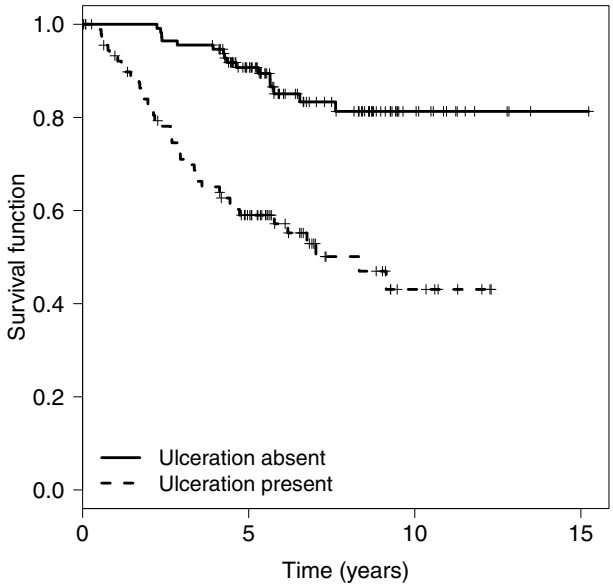

Figure 5: *KM estimator by nodal category ulceration status for melanoma data set.*

Table 6: *AIC/SBC criteria for power series cure rate model with Pareto IV and BS distribution for concurrent causes in the melanoma data set.*

	P2	P3	P4	Gamma	BS
Poisson	438.55/458.49	427.77/447.70	427.59/447.53	427.42/447.35	430.16/450.10
Logarithmic	447.47/467.41	**418.31/438.25**	418.39/438.33	**418.31/438.25**	425.88/445.82
Geometric	438.13/458.07	423.04/442.98	422.76/442.70	422.83/442.77	428.63/448.56
Binomial	441.42/461.36	432.79/452.72	432.79/452.73	432.84/452.78	432.84/452.78

In this case, we also fit the P2PS, P3PS and P4PS model for four particular cases, together with the gamma and BS models. The AIC and SBC criteria are presented in Table 6.

Both criterion suggest that the Logarithmic cure rate model with a Pareto III and gamma distributions for the concurrent causes are the best models, both yielding similar results. We also tested the hypotheses $H_0 : \alpha = 1$ versus $H_1 : \alpha \neq 1$ using the log-likelihood ratio (LR) test and the Wald test. In both cases, we failed to reject the null hypothesis at the 5% significance level and consequently, we prefer the P3 instead of the P4 distribution for the time-to-event in the concurrent causes. Parameter estimates of both selected model are presented in Table 7.

Table 7: *Estimates and standard errors (s.e.) for logarithmic P3 and gamma cure rate models for melanoma data set.*

	Estimate	s.e.		Estimate	s.e.
$\beta_{intercept}$	−0.8874	0.5714	$\beta_{intercept}$	−0.9761	0.5967
$\beta_{ulceration}$	1.9991	0.5864	$\beta_{ulceration}$	1.9619	0.5846
$\beta_{thickness}$	0.3753	0.1304	$\beta_{thickness}$	0.3774	0.1339
σ	7.3228	2.3068	α	2.6801	0.5092
γ	0.4325	0.0594	ν	0.3399	0.1726

Note that all parameters related to the regression are significantly different from zero in both models. Once more, based on a Taylor expansion of first order (around the intercept in this case) for the logarithmic model, we obtain $b_0 \approx -0.1162651$. In this manner, we present the following approximate interpretations of the regression coefficients:

- $\exp\left(-b_0 \times \widehat{\beta}_{thickness}\right) = 1.045$, i.e., for each mm that is increased the tumor thickness the cure rate is decreased in 4.5%.
- $\exp\left(-b_0 \times \widehat{\beta}_{ulceration}\right) = 1.262$, i.e., patients without ulceration have a cure rate 26.2% greater than patients with ulceration.

On the other hand, the mean and median of the time-to-event of carcinogenic cells are 10.18 and 7.32 years respectively with their respective 95% confidence intervals

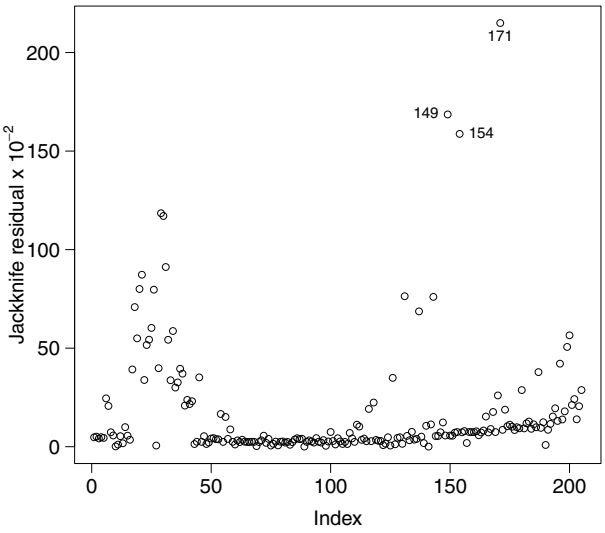

Figure 6: *Jackknife residuals for melanoma data set.*

Table 8: *Estimates, standard errors (s.e.) and 95% confidence interval for logarithmic P3 cure rate model for Phase III cutaneous melanoma clinical trial without possible influence observations.*

	deleted observations							
	149		154		171		149, 154 and 171	
	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
$\beta_{intercept}$	−1.0667	0.5242	−1.0099	0.5208	−1.0247	0.5151	−1.2670	0.4673
$\beta_{ulceration}$	2.0037	0.5691	1.9633	0.5728	1.9602	0.5722	1.9473	0.5536
$\beta_{thickness}$	0.3656	0.1269	0.3708	0.1296	0.3712	0.1297	0.3614	0.1271
σ	6.6195	1.8311	6.6301	1.8278	6.5522	1.7669	5.5503	1.2191
γ	0.4262	0.0587	0.4229	0.0583	0.4205	0.0579	0.4026	0.0558

(2.54,17.82) and (2.80,11.84). It can be verified that both models provide similar results in terms of estimated cure rates and survival functions. For this reason, henceforth we will continue the analysis based only on the logarithmic P3 model. Figure 6 shows the Jackknife residuals for this data set, suggesting that observations 149, 154 and 171 are possible influential observations. Based on a simple descriptive analysis, we note that those observations present large observed times even though the respective tumor thickness also are large.

Table 8 shows the estimates for the logarithmic P3 model deleting the possible influence observations separately and jointly. Note that in all cases the significance of parameters is unchanged and the estimates are very close to the estimations using the complete data set. Finally, Figure 7 presents the estimated cure rate and the respective

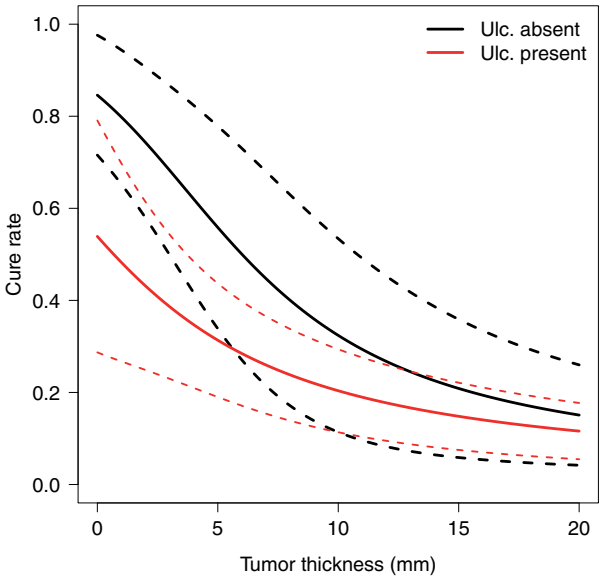

Figure 7: *Estimated cure rate for patients with ulceration status absent and present. The continuous line represent the point estimation and the dashed line represent the respective 95% confidence interval.*

95% confidence intervals, suggesting that ulceration is a risk factor. On the other hand, tumor thickness influences in the cure rate of patients subject to this intervention mainly for small tumors.

5. Simulation study

In this section we report a simulation study to assess the recovery of known parameters by the proposed estimation procedure. The data were drawn in conformity with the P4PS model. We assume the Pareto IV distribution with parameters $\alpha = 0.4$, $\sigma = 1$ and $\gamma = 0.6$ for the concurrent causes, i.e., a similar scheme to that fitted in the applications. We assume that observations belong to two groups, say $z_1 = 0$ or $z_1 = 1$. In addition, we assume a second continuous covariate, say z_2. For $i = 1, \ldots, n$, we drew z_{1i} and z_{2i} from a Bernoulli distribution with success probability equal to 0.5 and a Uniform distribution in the interval $(0, 20)$ respectively. For each model, the parameters related to the cure were computing by fixing cure rates (say q_0 and q_1) at determined values for each group, without considering the effect of covariate z_{2i}. We consider three kinds of cure rates: high ($q_0 = 0.8$ and $q_1 = 0.65$), medium ($q_0 = 0.6$ and $q_1 = 0.45$) and low ($q_0 = 0.4$ and $q_1 = 0.25$). To achieve this, the values for β_0 and β_1 for each distribution assumed for M are given in Table 9.

On the other hand, the value for β_2 was fixed as 0.1 in all cases. Using this setup, for each $i = 1, \ldots, n$ the value of θ_i was computed according to (5) and M_i was simulated

Table 9: *Values for β_0 and β_1 assumed in the simulation study.*

Distribution	High cure rate		Medium cure rate		Lower cure rate	
assumed for M	β_0	β_1	β_0	β_1	β_0	β_1
Poisson	−1.4999	0.6578	−0.6717	0.4467	−0.0874	0.4141
Logarithmic	−0.5264	0.9607	0.7343	0.9857	2.1180	1.7826
NB ($q=1$)	−1.3863	0.7673	−0.4055	0.6061	0.4055	0.6931
Binomial ($q=1$)	−1.3863	0.7673	−0.4055	0.6061	0.4055	0.6931

depending on each of the four power series distributions. We define $W_{0i}=\infty$ and for $M_i>0$, we drew $W_{1i},\ldots,W_{M_i i}$ from a Pareto IV distribution (if $U\sim U(0,1)$, so $\sigma(U^{-\frac{1}{\alpha}}-1)^{\gamma}\sim P4(\alpha,\sigma,\gamma)$). Then, we define $T_i^*=\min(W_{0i},W_{1i},\ldots,W_{M_i i})$. The failure time was defined as $T_i=\min(T_i^*,10)$ and $\delta_i=I(T_i^*\leq 10)$. We consider three sample sizes: $n=50$, $n=100$ and $n=200$. Each case was replicated 10,000 times and we report the average bias (AB) and the average of mean square error (AMSE) of the estimates. Results are presented in Table 10.

Table 10: *Simulation study for PSP4 model with cure rate.*

Distribution		$n=50$		$n=100$		$n=200$	
for M		bias	MSE	bias	MSE	bias	MSE
				High cure rate			
Poisson	β_0	−0.050	0.423	−0.024	0.186	−0.015	0.085
	β_1	0.043	0.203	0.016	0.092	0.011	0.042
	β_2	0.007	0.002	0.003	0.001	0.002	0.000
	α	0.040	0.254	0.033	0.163	0.021	0.094
	σ	0.098	0.501	0.054	0.201	0.031	0.103
	γ	−0.010	0.341	−0.005	0.119	−0.001	0.052
Logarithmic	β_0	0.030	2.320	−0.012	0.607	−0.020	0.245
	β_1	0.143	1.105	0.048	0.392	0.030	0.185
	β_2	0.016	0.009	0.006	0.003	0.003	0.001
	α	0.045	0.287	0.037	0.195	0.019	0.087
	σ	0.116	0.592	0.076	0.257	0.024	0.121
	γ	−0.037	0.320	−0.017	0.067	−0.007	0.045
Geometric	β_0	−0.053	0.668	−0.027	0.286	−0.017	0.132
	β_1	0.059	0.379	0.025	0.171	0.012	0.082
	β_2	0.007	0.003	0.003	0.001	0.002	0.001
	α	0.045	0.237	0.037	0.195	0.019	0.057
	σ	0.136	0.574	0.065	0.266	0.091	0.078
	γ	−0.019	0.219	−0.008	0.110	−0.004	0.055

Table 10: *Simulation study for PSP4 model with cure rate (continuation).*

Distribution for M		$n = 50$		$n = 100$		$n = 200$	
		bias	MSE	bias	MSE	bias	MSE
				High cure rate			
Bernoulli	β_0	−0.088	0.798	−0.079	0.355	−0.038	0.156
	β_1	0.065	0.541	0.050	0.307	0.019	0.112
	β_2	0.015	0.006	0.008	0.003	0.003	0.001
	α	0.058	0.277	0.037	0.178	0.019	0.051
	σ	0.100	0.612	0.063	0.186	0.022	0.067
	γ	−0.023	0.321	−0.004	0.009	−0.002	0.005
				Medium cure rate			
Poisson	β_0	−0.010	0.255	−0.013	0.116	−0.003	0.054
	β_1	0.030	0.134	0.013	0.061	0.006	0.029
	β_2	0.007	0.001	0.004	0.001	0.002	0.000
	α	0.034	0.233	0.029	0.136	0.015	0.087
	σ	0.087	0.452	0.051	0.186	0.027	0.092
	γ	−0.009	0.321	−0.004	0.100	−0.001	0.043
Logarithmic	β_0	0.168	2.616	0.025	0.656	0.011	0.269
	β_1	0.147	1.234	0.072	0.492	0.023	0.223
	β_2	0.014	0.010	0.007	0.004	0.002	0.002
	α	0.039	0.254	0.031	0.143	0.015	0.076
	σ	0.102	0.475	0.062	0.212	0.021	0.112
	γ	−0.034	0.287	−0.014	0.062	−0.006	0.038
Geometric	β_0	−0.009	0.518	0.005	0.225	0.001	0.103
	β_1	0.040	0.314	0.017	0.141	0.008	0.069
	β_2	0.008	0.003	0.003	0.001	0.001	0.001
	α	0.040	0.212	0.032	0.171	0.015	0.043
	σ	0.117	0.534	0.061	0.247	0.072	0.054
	γ	−0.015	0.192	−0.007	0.087	−0.003	0.049
Bernoulli	β_0	−0.074	0.542	−0.057	0.314	−0.032	0.139
	β_1	0.055	0.451	0.043	0.236	0.015	0.100
	β_2	0.011	0.005	0.006	0.002	0.002	0.001
	α	0.041	0.243	0.037	0.141	0.011	0.034
	σ	0.081	0.517	0.052	0.159	0.015	0.053
	γ	−0.019	0.259	−0.003	0.008	−0.002	0.004

Table 10: *Simulation study for PSP4 model with cure rate (continuation).*

Distribution for M		$n = 50$		$n = 100$		$n = 200$	
		bias	MSE	bias	MSE	bias	MSE
Poisson	β_0	0.045	0.231	0.017	0.098	0.009	0.045
	β_1	0.025	0.115	0.009	0.050	0.004	0.024
	β_2	0.007	0.001	0.004	0.000	0.002	0.000
	α	0.029	0.198	0.025	0.119	0.011	0.053
	σ	0.075	0.276	0.043	0.150	0.021	0.076
	γ	−0.008	0.276	−0.003	0.086	−0.000	0.033
Logarithmic	β_0	0.350	0.253	0.013	1.014	0.044	0.427
	β_1	0.253	3.837	0.082	0.789	0.044	0.341
	β_2	0.013	1.993	0.007	0.005	0.003	0.002
	α	0.032	0.214	0.028	0.113	0.012	0.059
	σ	0.089	0.429	0.053	0.189	0.018	0.097
	γ	−0.021	0.253	−0.010	0.042	−0.004	0.025
Geometric	β_0	0.063	0.012	0.030	0.232	0.013	0.106
	β_1	0.050	0.287	0.021	0.134	0.007	0.065
	β_2	0.005	0.002	0.003	0.001	0.001	0.001
	α	0.030	0.193	0.023	0.154	0.011	0.031
	σ	0.109	0.497	0.053	0.212	0.053	0.049
	γ	−0.011	0.153	−0.005	0.067	−0.002	0.032
Bernoulli	β_0	−0.049	0.417	−0.049	0.284	−0.023	0.097
	β_1	0.043	0.445	0.035	0.200	0.009	0.071
	β_2	0.007	0.004	0.004	0.002	0.001	0.001
	α	0.029	0.210	0.027	0.119	0.007	0.029
	σ	0.065	0.471	0.047	0.132	0.010	0.043
	γ	−0.017	0.212	−0.002	0.006	−0.001	0.002

Table 10 reveals an acceptable bias and MSE for all parameters and cases, except for the parameter σ for which a high bias and MSE was encountered for the small sample size. The bias and MSE decrease when the sample size is increased, suggesting that the parameter estimators are consistent. Finally, the bias and MSE decreases when the cure rate is decreased, which also is expected because for a lower cure rate, we expect more failure times observed in the sample, i.e., more precise information.

6. Final discussion

The Pareto IV power series cure rate model has been shown to outperform an analogous competing Birnbaum Saunders model for modeling a cutaneous melanoma data set. A

simulation study confirms that, with reasonable sample sizes, accurate parameter estimation is feasible within this model. An EM algorithm approach to obtaining maximum likelihood estimates can be recommended for these models. It is interesting to note that the rarely used logarithmic distribution turns out to be the distribution of choice among the four power series models considered.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions of Automatic Control*, 19, 716–723.

Arnold, B.C. (1983). *Pareto Distributions*. International Co-operative Publishing House. ISBN 0-89974-012-X.

Arnold, B.C. (2015). *Pareto Distributions*, 2nd edn. Series: Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Berkson, J. and Gage, R. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501–515.

Cancho, V.G., Louzada, F. and Ortega, E.M. (2013a). The power series cure rate model: an application to a cutaneous melanoma data. *Communications in Statistics - Simulation and Computation*, 42, 586–602.

Cancho, V.G., Louzada, F. and Barriga, G.D.C. (2013b). The Geometric Birnbaum-Saunders regression model with cure rate. *Journal of Statistical Planning and Inference*, 142, 993–1000.

Cordeiro, G.M., Cancho, V.G., Ortega, E.M.M. and Barriga, G.D.C. (2016). A model with long-term survivors: Negative binomial Birnbaum-Saunders. *Communication in Statistics - Theory and Methods*, 45, 1370–1387.

Gallardo, D.I., Bolfarine, H. and Pedroso-de-Lima, A.C. (2016a). An EM algorithm for estimating the destructive weighted Poisson cure rate model. *Journal of Statistical Computation and Simulation*, 86, 1497–1515.

Gallardo, D.I. and Bolfarine, H. (2016b). Two efficient estimation procedures for the negative binomial cure rate model with a latent activation scheme. *Statistics and Operations Research Transactions*, 40, 31–54.

Gallardo, D.I., Romeo, J.S. and Meyer, R. (2016c). A simplified estimation procedure based on the EM algorithm for the power series cure rate model. *Communication in Statistics - Simulation and Computation*. DOI: 10.1080/03610918.2016.1202276.

Gómez, Y.M. and Bolfarine, H. (2016). The geometric power half-normal regression model with cure rate. *Hacettepe Journal of Mathematics and Statistics*. DOI: 10.15672/HJMS.201613820026,

Hanin, L. and Li-Shang, H. (2014). Identifiability of cure rate models revisited. *Journal of Multivariate Analysis*, 130, 261–274.

Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.

Li, C.S., Taylor, J. and Sy, J. (2001). Identifiability of cure models. *Statistics and Probability Letters*, 54, 389–395.

Noack, A. (1950). On a class of discrete random variables. *Annals of Mathematical Statistics*, 21, 127–132.

Ortega, E.M.M., Barriga, G.D.C., Hashimoto, E.M., Cancho, V.G. and Cordeiro, G.M. (2014). A new class of survival regression models with cure fraction. *Journal of Data Science*, 12, 107–136.

Ortega, E.M.M., Cordeiro, G.M., Campelo, A.K., Kattan, M.W. and Cancho, V.G. (2015). A power series beta Weibull regression model for predicting breast carcinoma. *Statistics in Medicine*, 34, 1366–1388.

Pal, S. and Balakrishnan, N. (2016). An EM type estimation procedure for the destructive exponentially weighted Poisson regression cure model under generalized gamma lifetime. *Journal of Statistical Computation and Simulation*. DOI:10.1080/00949655.2016.1247843.

Piegorsch, W.W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, 46, 863–867.

Rodrigues, J., Cancho, V.G., Castro, M.A. and Louzada-Neto, F. (2009a). On the unification of the long-term survival models. *Statistics and Probability Letters*, 79, 753–759.

Rodrigues, J., de Castro, M.A., Cancho, V.G. and Balakrishnan, N. (2009b). COM-Poisson cure rate survival model and an application to a cutaneous melanoma data. *Journal of Planning and Inference*, 139, 3605–3611.

Rodrigues, J., Cordeiro, G., de Castro, M.A. and Nadarajah, S. (2015). A unified class of compound lifetime distributions. *Communications in Statistics - Theory and Methods*, 45, 2323–2331.

Scheike, T. (2015). *Timereg package*. R package version 1.8-9. With contributions from T. Martinussen, J. Silver and K. Holst. R package version 3.2.3.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Sen, P.K., Singer, J.M. and Pedroso-de-Lima, A.C. (2010). *From Finite Sample to Asymptotic Methods in Statistics*. New York: Cambridge University Press.

Yakovlev, A.Y. and Tsodikov, A.D. (1996). Stochastic models of tumor latency and their biostatistical applications. *World Scientific*, New Jersey.

SORT 41 (2) July-December 2017, 319-346 DOI: 10.2436/20.8080.02.62

Estimating regional social accounting matrices to analyse rural development

Alfredo Mainar-Causapé[1], José Manuel Rueda-Cantuche[2,*], M. Alejandro Cardenete[3,4], Patricia Fuentes-Saguar[2], M. Carmen Delgado[3], Fabien Santini[4] and Sergio Gómez y Paloma[4]

Abstract

This paper has two complementary objectives: on the one hand, it introduces the EURO method for the estimation of (regional) Social Accounting Matrices. This method is widely used by Eurostat for the estimation of missing national Supply, Use and Input-output tables but it has not been used before within the context of social accounting matrices or of regional statistics and/or regional impact analyses. On the other hand, this work discusses the possibility of producing non-survey-based regional Social Accounting Matrices that may eventually allow the user to carry out impact analyses such as those of rural development policies, among others. The analysis is carried out for 12 selected European regions based on clusters.

MSC: 97K80.

Keywords: Social accounting matrices, rural development, European regions, impact analysis.

1. Introduction

Social Accounting Matrices (SAMs) are datasets comprising economic transactions that allow the extraction of information on the different economic agents such as producers, consumers, the government and the foreign sector, as well as on the behaviour of productive factors and institutions. They complete the information provided by input-output tables. A Social Accounting Matrix can be defined (in a simplified form) as an extension of an input-output table with a more disaggregated structure of expenditures

* *Corresponding author.* European Commission, DG Joint Research Centre. Edif. Expo, C/Inca Garcilaso, 3. 41092 Seville (Spain). JoseM.RCantuche@ec.europa.eu

[1] Department of Applied Economy III, University of Seville (Spain).

[2] 2 Department of Economics, Quantitative Methods and Economic History, University Pablo de Olavide, Seville (Spain).

[3] Department of Economics, Universidad Loyola Andalucía, (Spain).

[4] School of Management and Business, Universidad Autónoma de Chile, (Chile).

Note: The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

Received: March 2017
Accepted: July 2017

and income, integrating the relationships between institutional sectors, estimated with information from national accounting systems. In this way, the objective of closing the full economic flow is achieved. Thus, a SAM is a consistent framework for gathering national income data, product accounts, input-output tables, reflecting the monetary flows among institutions. Therefore, a SAM is a matrix representing in a comprehensive, flexible and disaggregated way all the transactions of a socio-economic system. It reflects the process of income generation by activities, of production, and the distribution and redistribution of income between institutional groups (Pyatt and Round, 1985; Pyatt and Thorbecke, 1976). Figure A1 in Annex 1 shows the standard structure of a SAM.

The interest in SAMs is based on the fact that they illustrate the production relationships between the economic sectors as well as the transactions that take place among the different institutions of a certain economic system in terms of revenues or expenses. Besides their statistical interest, which enables us to close the circular flow of income, SAMs have become a useful tool for evaluation of policy interventions in national or regional frameworks. In this sense, it is interesting to have regional SAMs to be able to analyse the effect and impact of regional development policies, especially in rural areas. But the difficulty of obtaining databases for this purpose is an important obstacle that we attempt to overcome with the methodology presented here.

Moreover, it is possible to carry out a complete analysis of the productive structure of the economy and to obtain a general perspective of changes that might occur in the event of any shock (e.g. key sectors). Below, we present the approach used for obtaining 12 NUTS 3 level[1] regional SAMs. The estimates of the NUTS 3 SAMs are obtained using a two-step process:

1. Input-output frameworks are regionalised (i.e. Supply, Use and Symmetric tables) from the NUTS 1 regions or countries concerned, using the EURO method (Beutel, 2002, 2008; Eurostat, 2008; Temurshoev and Timmer, 2011; Valderas et al., 2016).
2. The NUTS 3 SAM estimation is calculated using the regionalised SUT and some additional information to produce the input-output tables.

Regarding policies, the Rural Development Policy, often referred to as Pillar 2, has become one of the most significant elements of the Common Agricultural Policy (CAP), representing close to one third of the total CAP budget. Before integration of flexibility between pillars and other adjustments, the amount dedicated to rural development policies over the financial period of 2014-2020 is likely to reach EUR 95 billion out of a total of EUR 348 billion for both pillars of the CAP (27% of the total). In recent years,

1. NUTS: Nomenclature of Territorial Units for Statistics. The NUTS classification is a hierarchical system for dividing up the economic territory of the EU for the purpose of socio-economic analyses of the regions: NUTS 1: major socio-economic regions; NUTS 2: basic regions for the application of regional policies; NUTS 3: small regions for specific diagnoses. The NUTS 2013 classification is valid from 1 January 2015 and lists 98 regions at the NUTS 1 level, 276 regions at the NUTS 2 level and 1,342 regions at the NUTS 3 level.

several research programmes, scientific papers and policy reports have looked at ways to assess the impacts of Pillar 2 at country and regional levels. The European Commission and the Member States carry out periodic ex ante, mid-term and ex post evaluation of the rural development policy and of the Rural Development Programmes. Several FP7 and Horizon 2020 research programmes are dedicated to the evaluation of the impact of rural development policies.

However, the diversity of rural situations across Europe has complicated the empirical studies of these impacts of rural development and often makes any comparison between regions rather trivial. Also, rural development policies do not only aim at supporting specific sectors (such as agriculture); indeed several measures are focused on non-farm actors, and others are related to the improvement of quality of life in rural areas. Hence, it is necessary to use multi-sectoral models, requiring a significant amount of data, in order to capture the full economic impact. In this sense, well-known linear multiplier models and computable general equilibrium (CGE) models use SAMs to develop their analysis. Also, given the nature of rural development (regional implementation through Rural Development Programmes and the existence of menus offered to the beneficiaries in each region), the need for modelling at a sub-regional level has led to the application of these models at the NUTS 3 level with models going as deep as modelling the rural area and the urban area of NUTS 3 regions. The challenge of such work is that it requires extensive effort in the construction of NUTS 3 SAMs, especially if the rural-urban split is modelled.

In this context, this paper builds NUTS 3 SAMs for 12 regions, following a detailed analysis of the source data rather than using an automatic approach, which would derive regional SAMs directly from superior level tables, using an optimisation method and some regional proxies. It aims to cover all types of NUTS 3 regions with significant participation of rural areas, so that the impact of rural development policies can be studied for most of the types of regions receiving aid, thus allowing the evaluation of their effectiveness. The selection of these NUTS 3 regions uses an empirical classification of NUTS 3 regions (Raggi et al. (2013)), which reflects the heterogeneity of NUTS 3 characteristics in the EU. This multidimensional classification is based on the following set of four criteria: Rural character; Accessibility; Actual economic diversification; and Total gross domestic product per capita.

So, this paper has two complementary objectives: it introduces the EURO method for the estimation of (regional) Social Accounting Matrices and illustrates the possibility of producing non-survey-based regional Social Accounting Matrices for rural development policies' impact analyses.

The rest of the paper is structured as follows: Section 2 discusses the methodology used in the regionalised SAM estimations and its application in some European Union regions. Section 3 presents the main results and, finally, Section 4 provides conclusions. Some tables and aggregated versions of the estimated SAMs are included in the Annex.

2. Methodology and data

2.1. The EURO method for estimating supply and use tables

The general balancing problem of matrices basically consists of only knowing one single base table (be it a Supply and Use Table (SUT), Symmetric Input-Output Table (SIOT) and/or Social Accounting Matrix (SAM)) and at least the row and column totals for the unknown table that has to be estimated[2]. There are different ways to approach this underdetermined problem where unknowns (e.g. elements of the interior tables) outnumber external constraints (e.g. RAS[3] or bi-proportional scaling methods, Lenzen, Gallego and Wood, 2009, among others).

However, none of these methods allows the estimation of SUTs and SIOTs whenever row and column totals are not given and with the minimum amount of information possible. Actually, to the knowledge of the authors, the EURO method is the only existing method that allows the estimation of SUTs and SIOTs without given row and column totals. The EURO method typically aims at updating SIOTs at basic prices from one year to another and is based on a previous version initially developed by Beutel (2002) for input-output tables and further explained by the Eurostat Manual of Supply, Use and Input-Output Tables (2008, Ch. 14).

The EURO method is a robust update procedure which is inexpensive and has limited data requirements. It exclusively uses official data and integrates all quadrants of SIOTs. Row and column totals for intermediate consumption and output and the corresponding final demand structure are derived endogenously, not allowing for arbitrary changes of input-output coefficients. The method is fully consistent with supply and demand through the Leontief quantity model (Eurostat, 2008). Therefore, it is sustained on economic grounds rather than on optimisation and/or pure mathematical techniques.

Recently, Temurshoev, Webb and Yamano (2011) formalised a SUT variant of the EURO method based on Beutel (2008). Beutel and Rueda-Cantuche (2012) elaborated a more detailed version to be used by Eurostat. And, in line with the pioneering works of Hewings (1969, 1977), we formulate an adapted version of the latter to be used in this project for the regionalisation of supply and use tables.

The EURO method is used in this paper as a method for regionalisation for the first time. Below, we present an adapted and more detailed explanation of the EURO method for SUT regionalisation, mostly based on Temurshoev et al.'s (2011) description of the EURO method for updating SUTs.

The initial SUTs (typically at the NUTS 1 or NUTS 2 level) consist of the following components all expressed at basic prices: domestic and imported intermediate use matri-

2. Mínguez, Oosterhaven and Escobedo (2009) and Oosterhaven and Escobedo (2011) consider several known tables as base tables but the lack of information at NUTS3 level makes this analysis inappropriate for our purpose.

3. In the original presentation of this method (i.e. working paper), the vector of row multipliers was designated by r, the table of inter-industry transactions in coefficient form in the base year by A and the vector of column multipliers by s. Hence the juxtaposition of the notation led to the nomenclature RAS (as originally in Stone, 1961).

ces (commodity × industry); domestic and imported final demand matrices (commodity × category of final use); supply matrix (commodity × industry); vector of total value added of industries (industry × 1); and a vector of total taxes less subsidies on products by industries and final use categories. The projected SUTs require the following macroeconomic statistics for the SUTs at the NUTS 3 level, based on regionalisation rates[4] of macroeconomic variables: value added by industry; total final demand by use; total taxes less subsidies on products; and total imports. The listed data requirements mean that the vectors of value added per industry, totals of final demand categories and aggregate values of taxes less subsidies on products and imports need to be known at the NUTS 3 level too.

Following Thissen, Diodato and van Oort (2010), we have used information on interregional transport flows to estimate regional imports and exports. We have used the Eurostat data on road freight transport loading (exports) and unloading (imports) in physical terms and have calculated a ratio over the whole country (in physical terms). The method uses these official statistics as exogenous inputs, and replicates them in the derived SUTs. This method involves minimum data requirements, which is appropriate given the lack of macroeconomic data at the NUTS 3 level.

Each of the iterations of the EURO method consists of two steps (see Figure 1). The first step of the first iteration defines domestic and imported intermediate and final uses, the vector of value added, the vector of taxes less subsidies on products, and the supply matrix of the projected SUTs. This first estimation of the (unbalanced) use table is basically a cell-wise arithmetic average resulting from multiplying the corresponding regionalisation rates by the rows and columns of the initial use table. Subsequently, the total commodity output (from the estimated use table) is allocated row-wise proportionally to the initial supply table (i.e. constant market shares) in order to obtain the first estimation of the supply table at the NUTS 3 level. The total industry outputs and inputs are not equal after this first step (column sums of projected supply and use tables). To make the derived SUTs consistent, it is assumed that the domestic and imported input structures of industries and the totals of commodities' final uses from the first step are valid. Given this assumption, the so-called fixed commodity sales structure model determines consistent industry output and input levels (Eurostat, 2008, Model D, p. 351). This second step ensures the consistency of the industry outputs and inputs, and commodity supply and demand, but it deviates from macroeconomic statistics, i.e. value added per industry, final uses of categories, total value added and total imports.

The regionalisation rates initially used are then adjusted in an iterative procedure in order to make the difference between the actual and projected (in each of the iterations) regionalisation rates minimal (less than 1%). The observed deviations are used to correct these rates in such a way that it should ensure that if the model overestimates (underestimates) the available macroeconomic statistics, the corresponding regionalisation rates

4. They are calculated as regional/national ratios.

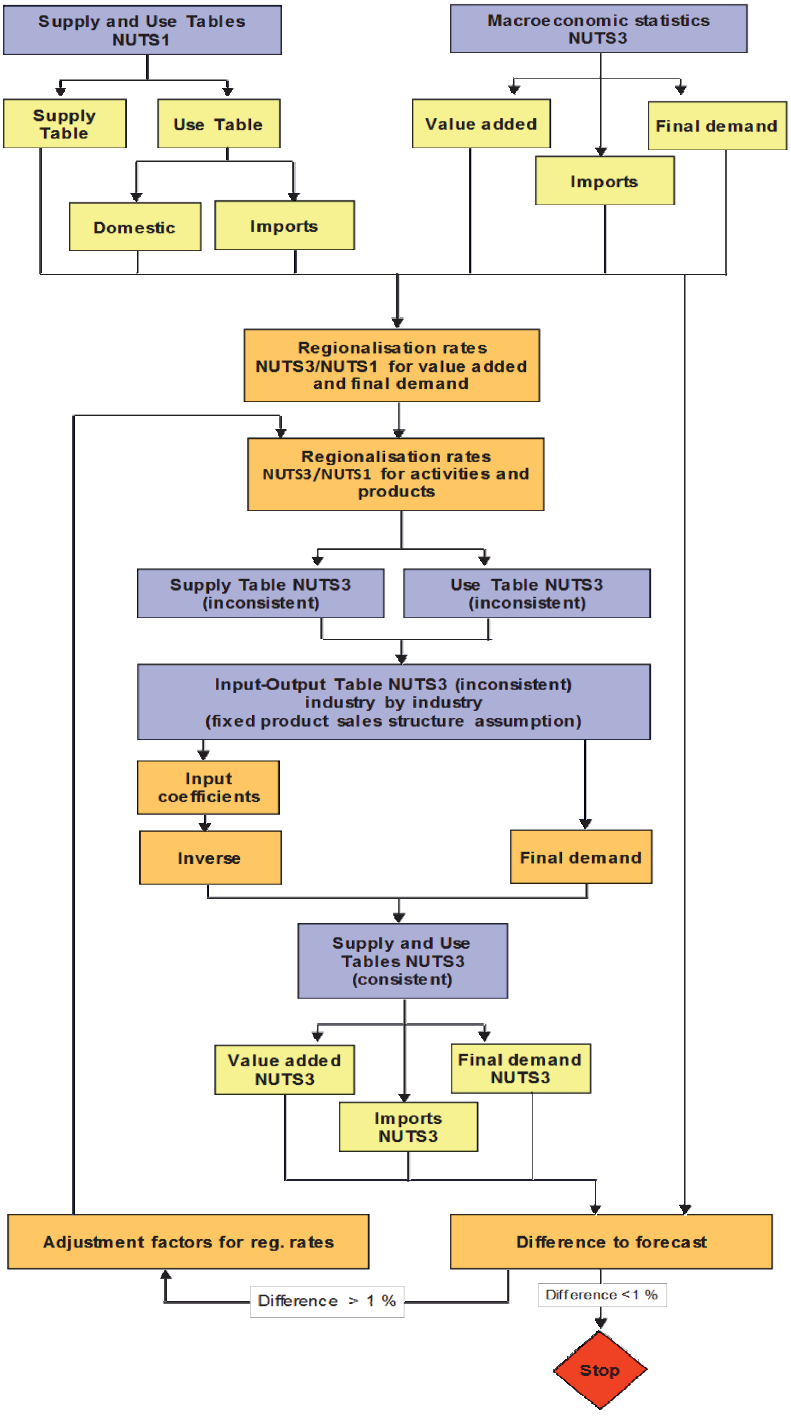

Figure 1: *EURO method for regionalising SUTs.*

Source: Own elaboration based on Beutel and Rueda-Cantuche (2012).

are decreased (increased). This is done through correction factors (see Eurostat, 2008). Then, the first step of the second iteration computes the projected SUT components as in the first iteration, i.e. domestic and imported intermediate and final uses, the vector of value added, the vector of taxes less subsidies on products, and the supply matrix of the projected SUTs. As was the case with the first step of the first iteration, the results do not ensure the equality of industry outputs and inputs. The consistent industry outputs and inputs are again found using the fixed commodity sales structure model, which is then used to derive the consistent SUTs of the second iteration in exactly the same manner as defined earlier for the first iteration.

However, note that now the domestic and imported input structure matrices are derived from the outcomes of the first step of the second iteration. As a result, one obtains a new deviation vector, which quantifies the difference between the projected regionalisation rates and the macroeconomic statistics.

If the difference between the actual and projected regionalisation rates is acceptable, the resulting SUTs are the final outcome of the EURO projection. Otherwise, the steps of the second iteration are repeated until the projected variables resemble (closely or perfectly) those of the macroeconomic statistics.

It is important to note that each such subsequent iteration begins with the computation of new correction factors, which are then used to correct the regionalisation rates from the previous iteration. The convergence in the EURO method can always be found by changing the tolerance level until convergence is reached. The last important point concerning the EURO method is that it requires that the number of industries and commodities are equal. Thus, even though the EURO method distinguishes between products and industries, it does not allow for the estimation of rectangular SUTs[5].

The data requirements of the EURO method are the following for the NUTS 3 case studies: gross value added by industry; taxes less subsidies on products (total); final demand components (totals), including exports; and total imports. The following sections explain the data sources and methods used in the calculation of the necessary data for the projections.

Gross value added by industry. It is not very common or easy to find detailed data on gross value added by industry at the NUTS 3 level. In this paper, we use a breakdown of 6 products/sectors (see below), which will be split up into 13 products/sectors according to the NUTS 1 or NUTS 2 shares, depending on the available information (see Table A2 in Annex 1 for details about the 13 products/sectors).

Taxes less subsidies on products (total). Provided that the GDP is available for the NUTS 3 regions, its difference with respect to the total sum of gross value added at basic prices (also available) makes the overall total of taxes less subsidies on products.

5. In this paper, the EURO method is programmed in the Eviews software and Excel templates are used to adapt the results to the standard Eurostat format.

Table 1: *Example of a final demand estimation using NUTS 2 or NUTS 1 data.*

	Baden-Württemberg	**Konstanz**
	Share of GDP final demand components	Values (million EUR)
GDP	**100%**	**7,961.68**
Consumption of households	54.4%	4,328.71
Consumption of Public Administration and NPISH	15.3%	1,221.22
Gross capital formation	18.4%	1,463.84
Net exports	11.9%	947.91

Source: Own elaboration.

Final demand components and imports. Gross domestic product (GDP) is defined as the sum of: final consumption of households; final consumption of government and non-profit institutions serving households; gross capital formation (investment); and net exports (exports minus imports).

Therefore, by using this definition of GDP, we split up the value of GDP for NUTS 3 regions using the shares of GDP components from the NUTS 2 or NUTS 1 regions (wherever available). As an example (see Table 1), the Baden-Württemberg (NUTS 2) shares of GDP components are given below as well as the GDP of Konstanz (NUTS 3) for 2007 and the corresponding calculation of its final demand total by category.

However, we are interested in calculating exports and imports separately and not as net exports. In order to do so, we estimate NUTS 3 exports and NUTS 3 imports according to the NUTS 3/NUTS 1 share of the Eurostat data on road freight transport loading (exports) and unloading (imports). As a result, in a second step, net exports are recalculated and the other final demand components adjusted accordingly.

2.2. Estimation and selection of representative regional SAMs

For the construction of NUTS 3 SAMs, we initially develop a basic SAM linking the input-output framework previously estimated, closing economic flows between productive sectors, commodities and institutional sectors. To do this, we use additional information, most of it from Eurostat in order to achieve greater uniformity in the estimation of the matrices for all the NUTS 3 analysed. However, when more specific information is necessary, we obtain it from local or national statistical offices. The basic sources[6] used are:

- allocation of primary and secondary income account of households by NUTS 1 and NUTS 2 regions (e. g. Baden-Württemberg/Freiburg-Konstanz) - Eurostat;

6. All Eurostat data can be found in http://ec.europa.eu/eurostat/data/database.

- income of households by NUTS 2 region (e.g. Freiburg-Konstanz) - Eurostat;
- compensation of employees by NUTS 2 region (e.g. Freiburg-Konstanz) - Eurostat;
- employment by NUTS 3 regions - Eurostat;
- non-financial transactions (e.g. Germany-Konstanz) - Eurostat;
- gross domestic product (GDP) at current market prices by NUTS 3 region - Eurostat;
- gross value added at basic prices by NUTS 3 regions (NACE_R1) - Eurostat;
- disposable income of households - national statistical offices (e.g. Konstanz: VGR der Länder: Regionaldatenbank Deutschland);
- Input-output tables at NUTS 1 or country level (e.g. Germany 2005) - Eurostat and OECD[7].

This information is incorporated into the input-output framework provided, obtaining a first version of the matrix for each NUTS 3 region. Small discrepancies that may arise in the estimation process are corrected by using a simple technical adjustment through RAS[8]. The result is a NUTS 3 level basic SAM composed of the accounts presented in Table A3 (see Annex 1).

Basic SAMs for each NUTS 3 region can be extended to successively incorporate the accounts and sectors needed to perform the required analysis of the corresponding regions. For this, the basic SAM accounts are disaggregated by block, using new information, almost entirely from Eurostat, to achieve the greatest possible homogeneity:

- farmland: number of farms and areas by economic size of farm (ESU) and NUTS 2 region;
- agricultural accounts according to EAA 97 Rev.1.1 by NUTS 2 region;
- average annual earnings by economic activity, sex, occupation - country level;.
- employment by occupation and economic activity - country level;
- structure of consumption expenditure by degree of urbanisation (COICOP level 2) (1 000) - country level;
- mean consumption expenditure by degree of urbanisation (in PPS) - country level;

7. OECD (2015).

8. The only exception in this initial procedure is the SAM for Huesca (Aragon, Spain), which comes from a previous expert's version for 2005 (elaborated by the authors) and which has simply been updated to 2007 using basic information from Eurostat and the RAS adjustment.

- household characteristics by degree of urbanisation - country level;
- population in rural areas (NUTS 2-3 level) - Eurostat Regional Statistics; Rural Development Indicators;
- employment (in persons) by rural/urban typology (NACE R1) - country level;
- gross value added at basic prices (NACE R1) - country level.

In selected regions, SAMs are estimated following an innovative methodology which allows reliable kind of database to be obtained despite the great difficulty of procuring data at this level of disaggregation, combining regionalisation and updating methods with the use of Regional and National Accounts and other socio-economic and business statistics.

The aim is to provide SAMs that are representative of rural regions of the EU, so first it is necessary to select an adequate list of NUTS 3 level regions reflecting the actual heterogeneity. With this in mind, regions have been chosen following first a cluster classification of European NUTS 3 regions (Raggi et al., 2013). This cluster classification divides the set of NUTS 3 regions into six groups with the following characteristics (the percentage of the total NUTS 3 regions is shown in brackets):

- Cluster 1 includes NUTS 3 regions classified as intermediate urban/rural, which are economically diversified, with high accessibility and a high GDP (28.2%);
- Cluster 2 contains rural NUTS 3 regions, which are dependent on agriculture, with good accessibility and a high GDP (25.8%);
- Cluster 3 takes into account NUTS 3 regions that are predominantly rural and dependent on agriculture, with low accessibility and a low GDP (13.7%);
- Cluster 4 considers NUTS 3 regions that are predominantly urban and not reliant on agriculture, with high accessibility and a high GDP (12.8%).
- Cluster 5 contains rural NUTS 3 regions, which are strongly economically dependent on agriculture, with the lowest accessibility index and a low GDP (11.3%);
- Cluster 6 consists of urban and intermediate NUTS 3 regions with a low GDP, intermediate accessibility and intermediate economic diversification (8.2%).

After discussion, and taking into account the data availability and the weight of each cluster, regions have been selected. The objective of this selection is to have significant representation of each cluster, so all the different typologies of regions will be well represented. Given that the purpose of the study is to provide databases (SAMs) to study measures of rural development, Cluster 4 regions have been excluded from the selection (no rural or agricultural component type). The selected list of regions and clusters are presented in Table A1 in Annex 1.

Here it is necessary to specify the information required to distinguish between rural and urban activities. The former are those carried out in rural areas, while the latter are those that are based in urban areas. To distinguish between urban and rural areas, we take as a reference the DGURBA2011[9] database which provides information on new classifications of urbanisation[10]. The LAU 2[11] types 1 or 3 are directly classified as urban or rural, respectively, while type 2 is classified using a threshold of 30 000 inhabitants (below this threshold is classified as rural and above is classified as urban). This typology allows fitting the objectives of the study to better distinguish between cases within 'intermediate' areas.

It is very difficult to obtain aggregated and homogeneous accurate information for this split for all cases. We have therefore used an estimate based on a private database (Orbis, developed by Bureau van Dijk) from companies at the highest level of geographical disaggregation. This database distinguishes the number of businesses by industry (NACE R1-R2) at the equivalent of the LAU 2 level or similar. We have completed the necessary information base with LAU 2 demographic data and other official statistics from Eurostat on predominantly rural, intermediate and predominantly urban areas.

With this data, the percentages of companies in rural and urban areas in each sector in each NUTS3 region are obtained, which allows the disaggregation between rural and urban sectors in the corresponding SAMs. This disaggregation based on the number of companies gives an adequate representation of the economic reality of each region.

With this statistical information, the percentage representing economic activities in rural and urban areas for each sector can be identified for each NUTS 3 region. This disaggregation criterion considers that companies that have their head office in a LAU 2 (or similar) regarded as rural (urban) are entirely allocated to the "rural" ("urban") part of the corresponding NUTS 3 region. This creates a division between rural and urban activities within each sector and NUTS 3 region. Obviously, economic activities in intermediate areas are classified as rural or urban based on the previous decision on the allocation of their place of establishment.

For the distinction between large and small farms, we have used data on the number of farms and areas by the economic size of farm (ESU) and NUTS 2 region, and agricultural accounts according to EAA 97 Rev.1.1 by NUTS 2 region, both available from Eurostat. The threshold of 16 ESU is used to distinguish between large and small farms for all regions. While we acknowledge that such an assumption may lead to inaccuracies in the description of farm sectors across the EU, it is necessary to protect a strong degree of data homogeneity.

9. http://ec.europa.eu/eurostat/ramon/miscellaneous/index.cfm?TargetURL=DSP_DEGURBA

10. The classification we use is: 1: densely populated (urban); 2: intermediate (small towns and suburbs); and 3: sparsely populated (rural). We also use population at level LAU 2 (completed with data from national statistical offices).

11. LAU: Local Administrative Units. The lowest LAU level (LAU level 2, formerly NUTS level 5) consists of municipalities or equivalent units in the 28 EU Member States.

Regarding the SAM estimations, we have also had to take into account that the time periods for which we have additional statistical information do not always coincide with the reference year (2007). In such cases, the nearest periods have been taken and we have used ratios because they are more stable than absolute values.

Next, once the accounts have been disaggregated, we have applied the Cross-Entropy Method to achieve the final adjustment for the final version of the SAMs at the NUTS 3 level. The Cross-Entropy Method (CEM) has been developed and adapted, among others, by Golan, Judge and Robinson (1994), Thissen and Lofgren (1998) and Robinson, Cattaneo and El-Said (2001). In comparison with the RAS estimation method, CEM is more flexible, cost-efficient and consistent with all the information provided by national accounts and other resources. This method has been extensively used in the literature and can also consider relationships to be incorporated into the estimation model as additional restrictions[12].

The Cross-Entropy approach involves projecting technical coefficients instead of total SAM flows. Once the new coefficients have been obtained, the new SAM can be derived in the usual way. Because CEM aims directly at estimating technical coefficients, the scaling method does not work. The problem would consist of the following minimisation problem:

$$d\left(\mathbf{A}^0, \hat{\mathbf{A}}^1\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\hat{a}_{ij}^1/X_j\right)\left(\ln\left(\hat{a}_{ij}^1/X_j\right) - \ln\left(a_{ij}^0/X_j^0\right)\right) \tag{1}$$

s.t.

$$\sum_{j=1}^{n}\hat{a}_{ij}^1 = X_i \quad \forall i$$

$$\sum_{i=1}^{n}\hat{a}_{ij}^1 = X_j \quad \forall j$$

$$a_{ij}^0 = 0 \rightarrow \hat{a}_{ij}^1 = 0$$

where $\mathbf{A} = (a_{ij})$ represents a matrix in a set $\mathbf{A}_n$ of $(n \times n)$ non-negative matrices with no row or column full of zeros. Considering a matrix $\mathbf{A}^0 \in \mathbf{A}_n$, a positive vector $\mathbf{x} \in \mathbb{R}_+^n$ and a loss function $\boldsymbol{d} : \mathbf{A}_n \times \mathbf{A}_n \rightarrow \mathbb{R}$, then $x_j^0 = \Sigma_i a_{ij}^0$ is the value for the j-th row and column sum in the original matrix; and a_{ij}^0/x_j^0 and $\hat{a}_{ij}^0/x_j$ the initial and updated technical coefficients, respectively.

Many other distances from metric spaces, besides the RAS and Cross Entropy (CE) minimands, are available to minimise the loss function but either they do not seem to

12. For further details, see Cardenete and Sancho (2004).

outperform RAS or their interpretation is not straightforward in terms of information theory or economic content (see Jackson and Murray, 2003). A possible complement to RAS and CE is suggested by classical information retrieval theory, a branch of computer science concerned with developing efficient methods of retrieving information from a data bank (Salton and McGill, 1983). Whenever a query for data is formulated, a retrieval algorithm fetches documents in a data bank that are closely related to the query in some way. The greater the similarity between the query and the information contained in the retrieved documents, the more successful the algorithm. Notice that a base SAM can be seen as a query for the true but unknown SAM document and an information retrieval algorithm will fetch from the data bank (the set of feasible SAMs) one with information content closely matching that required by the query. For a technical description of the procedure, see Cardenete and Sancho (2004).

Finally, it is necessary to stress that the final structure of the SAM accounts should be unique and wide enough to collect specific circumstances of a particular regional economy. For this reason, we leave in the SAMs accounts such as Agriculture or Forestry in urban areas, which in an ad hoc analysis of many economies would be considered negligible but are modelled for homogeneity reasons. Furthermore, the structure of the NUTS 3 SAMs comprising 63 accounts is as shown in Table A4 (see Annex 1).

In order to analyse changes in technical coefficients, the first idea is to measure some indicators of statistical distances between the I/O or SAM tables. When pairs of Input-Output or SAM tables are compared, it is possible to compute the Le Masné Index (Le Masné, 1990) for the sector j:

$$S_j = 100 * \left(1 - 0.5 \sum_i \left|a_{ij}^A - a_{ij}^S\right|\right) \tag{2}$$

The Le Masné Index will be close to 100 in cases of high similarity, and is therefore one of the many statistical distance indicators that can be analysed for the purpose of studying the similarity between tables. Table 2 shows the Le Masné index for Huesca, Konstanz and Lüneburg for analysing the similarities between a SAM built with an automatic procedure and a SAM built with an expert procedure.

Table 2: *Le Masné Index (average values) - Automatic procedure vs. Expert procedure.*

	Huesca	**Konstanz**	**Lüneburg**
Activities accounts	**90.63**	**91.78**	**91.42**
Commodity accounts	**79.07**	**86.96**	**88.73**
Rest of accounts	**81.04**	**72.97**	**77.02**
All accounts	**83.76**	**84.68**	**86.34**

Source: Own elaboration.

Table 2 shows that the total average similarity between procedures is 83.76% for Huesca, 84.68% for Konstanz and 86.34% for Lüneburg. In the case of the average for activities, it shows 90.63% for Huesca, 91.78% for Konstanz and 91.42% for Lüneburg. The similarity is higher than average for commodities: 79.07% for Huesca, 86.96% for Konstanz and 88.73% for Lüneburg. There is a high degree of similarity in the majority of accounts, with it being higher in Lüneburg and lower in Huesca. However, the case of Huesca is slightly different and the similarity indicator is the lowest. This may be due to the specific characteristics in the construction and later updating of this database. The NUTS 3 SAM of Huesca has been constructed with specific data which was available in regional statistical accounts, while the two other NUTS 3 SAMs are ultimately derived from the German national accounts.

3. Results

Following the methodology presented, the 12 SAMs for the selected regions are estimated, all referring to 2007 (for reasons of data availability at the time of the completion of paper). Such matrices are available in full upon request to the authors; Annex 2 shows only an aggregation.

However, to illustrate the validity and importance of the SAMs obtained, a summary is given in Table 3 including some of the main ratios derived from the estimated SAMs and reference to the rural or urban character of the NUTS 3 region; the importance of the activities in rural areas in general, and agricultural activities and (rural and urban) food processing industries in particular; and the trade relationship established outside the regions. The results presented in Table 3 correspond largely with the characteristics that define each of the clusters, demonstrating the importance of the rural economy and the need for its development and the importance of investment and public support (for example through Pillar 2).

Some interesting results can be obtained as illustrative examples of potential analyses that could be further developed with the estimated regional SAMs, for example to identify the economic structure of these regions.

Apart from the classification of NUTS 3 regions in one cluster or another, the importance of the economy of rural areas is fundamental, especially in the units classified in Clusters 5 and 6. With the exception of the low value in Noord-Drenthe (NL) (due to the diffuse criterion that sometimes separates both activities in this type of region), the percentage of GVA is high, surpassing 50% in all other regions except Lüneburg (DE), Norfolk (UK) and Slupski (PL), where it stands at around 40%.

Another significant aspect is the weight of public (government) investment, especially high in the provinces or regions of Cluster 1. The only low values are found in Huesca (ES) and Gorenjska (SI). The different role of trade with other regions is also remarkable, showing a possible cross-hauling effect with simultaneous imports and ex-

Table 3: *Some summary ratios of the NUTS 3 SAMs for 2007.*

Cluster	Country	NUTS 3 region	Rural activities share in total GVA	Agriculture and food (Rural + Urban) share in total GVA	Government investment/ Total investment	Imports/ GDP	Exports/ GDP
(1)	DE	Lüneburg	0.392	0.029	0.351	0.363	0.474
(1)	UK	Norfolk	0.388	0.030	0.386	0.495	0.409
(1)	DE	Konstanz	0.739	0.032	0.399	0.350	0.404
(2)	FR	Finistère	0.677	0.052	0.182	0.392	0.340
(2)	SI	Gorenjska	0.780	0.039	0.031	2.670	2.342
(2)	NL	Noord-Drenthe	0.063	0.051	0.129	1.204	1.256
(2)	SE	Örebro	0.605	0.036	0.244	0.673	0.747
(3)	HU	Heves	0.610	0.087	0.286	1.172	1.218
(3)	EE	Lääne-Eesti	0.586	0.096	0.287	1.134	0.897
(3)	PL	Slupski	0.477	0.091	0.121	0.826	0.538
(5)	ES	Huesca	0.801	0.133	0.031	0.625	0.483
(6)	PT	Setúbal	0.711	0.042	0.198	0.713	0.584

Source: Own elaboration. (GVA: gross value added; GDP: gross domestic product).

ports of the same goods, with trading totals being especially high in Gorenjska (SI), Heves (HU) and LääneEesti (EE).

Comparing the results by NUTS 3 territories, the share of Agriculture and food activities in Cluster 1 regions (Lüneburg, Norfolk and Konstanz) only represents 3% of their regional GVA with high shares of public investment, i.e. around 35-40% of the total gross fixed capital formation. Besides, the ratio of exports and imports on GDP shows similar behaviour. These results suggest strong dependency on the public sector with weak links with the rest of the national and international economies.

Regarding regions of Cluster 2, the behaviour is not so homogeneous. The agri-food activities in Gorenjska and Örebro are almost 4% of their total regional GVA, while Noord Drenthe and Finistère are over 5%. Public investment is not so important in these regions except in Örebro, where it is 24% of the total regional investment. Greater disparities are observed in the trade links with the rest of the economy provided that Gorenjska and Noord Drenthe have trade flows well above their respective regional GDP; opposite to Finistère and Örebro, where they are much lower, especially in the French region.

In Cluster 3 regions, agri-food activities represent almost 10% of their regional GVA, with public investment around 30% of total regional investment, except in Slupski. Only the Polish region shows both ratios of imports and exports below one.

In Huesca (Cluster 5), the agri-food activities represent 13% of its regional GVA, with a very small public sector share in its total regional investment (i.e. 3%) and around 40-60% of its GDP traded with other territories.

Finally, the agricultural sector and the agri-food industry in Setúbal (Cluster 6) represent together 4% of its regional GVA, with a public investment close to 20% of their total regional investment. Trade shares over GDP with other regions and countries are similar to Huesca, although slightly higher.

Therefore, in the light of the results obtained, the methodology used for estimating the SAMs at NUTS 3 level appears to be adequate and provide a significant contribution as a tool for obtaining such information, which is important for the assessment of regional economic development policies.

4. Conclusions

This paper describes a novel methodology for estimating non-survey-based regional Social Accounting Matrices with limited information for a selection of 12 NUTS 3 EU regions. For the first time, a modified version of the EURO method for Supply and Use Tables has been used as a method for regionalisation. The resulting SAMs can be further used for policy analysis, for example for modelling the impacts of rural development policies by using linear multipliers or computable general equilibrium (CGE)-based model approaches.

These SAMs used, as far as possible, existing regional/local data from their respective national and/or regional statistical offices, with consideration of the disaggregation of specific institutional sectors by degree of urbanisation (rural vs. urban areas).

Given the lack of official survey-based information to build regional (NUTS 2/NUTS 3) SAMs, we conclude that the methodology proposed in this paper can be useful (and replicated) to estimate non-survey-based regional SAMs with (optional) ad hoc specific considerations for certain sectors depending on the purpose of the analysis, i.e. rural/urban split for analysing rural development policies. Notwithstanding the caveats/assumptions made in our approach, we believe that sound impact analyses (e.g. using linear multipliers, CGE models, etc.) can be carried out in the future with regional SAMs estimated in the way we propose in this paper.

Annex 1. Tables and figures

Table A1: *NUTS 3 regions selected.*

	NUTS 3	Cluster	NUTS 2	NUTS 1	Member State
1	Lüneburg	(1)	Lüneburg	NIEDERSACHSEN	Germany
2	Norfolk	(1)	East Anglia	EAST OF ENGLAND	United Kingdom
3	Konstanz	(1)	Freiburg	BADEN-WÜRTTEMBERG	Germany
4	Finistère	(2)	Bretagne	OUEST	France
5	Gorenjska	(2)	Zahodna Slovenija	SLOVENIJA	Slovenia
6	Noord-Drenthe	(2)	Drenthe	NOORD-NEDERLAND	Netherlands
7	Örebro	(2)	Östra Mellansverige	ÖSTRA SVERIGE	Sweden
8	Heves	(3)	Észak-Magyarország	ALFÖLD ÉS ÉSZAK	Hungary
9	Lääne-Eesti	(3)	Eesti	EESTI	Estonia
10	Słupski	(3)	Pomorskie	REGION PÓŁNOCNY	Poland
11	Huesca	(5)	Aragón	NORESTE	Spain
12	Península de Setúbal	(6)	Área Metrop. de Lisboa	CONTINENTE	Portugal

Source: Own elaboration.

Table A2: *List of products/sectors.*

Original data source	Used in SAMs
1. Agriculture, forestry and fishing	1. Agriculture[13]
2. Manufacturing industry	2. Forestry
3. Construction	3. Fishing
4. Trade, transport and telecommunications	4. Mining
5. Finance, renting and business services	5. Food and beverages
6. Public services and other services	6. Other manufacturing activities
	7. Utilities
	8. Construction
	9. Trade
	10. Hotels and restaurants
	11. Transport and telecommunications
	12. Other private services
	13. Public services

Source: Own elaboration.

13. This industry still needs to be broken down further into arable crops, permanent crops and other agricultural products.

Table A3: *NUTS 3 basic SAM accounts.*

A.0-1	Agriculture, hunting and related services	C.0-6	Other manufacturing
A.0-2	Forestry, logging and related services	C.0-7	Utilities
A.0-3	Fish	C.0-8	Construction
A.0-4	Mining	C.0-9	Trade
A.0-5	Food industries	C.0-10	Hotels and restaurants
A.0-6	Other manufacturing	C.0-11	Transport and communication
A.0-7	Utilities	C.0-12	Other private services
A.0-8	Construction	C.0-13	Public services
A.0-9	Trade	L	Labour
A.0-10	Hotels and restaurants	K	Capital
A.0-11	Transport and communication	ANT	Activity net taxes
A.0-12	Other private services	CNT	Commodity net taxes
A.0-13	Public services	INT	Income net taxes
C.0-1	Prod. of agric., hunting and related services	H	Households
C.0-2	Prod. of forestry, logging and related services	E	Enterprises
C.0-3	Fish	G	Government
C.0-4	Mining	IS	I-S
C.0-5	Food industries	ROW	Rest of the world

Source: Own elaboration.

Table A4: *Structure of the NUTS 3 SAM for 2007.*

Rural activities	A.0-1_1_R	Small arable crops farms_Rural
	A.0-1_2_R	Large arable crops farms_Rural
	A.0-1_3_R	Small permanent crops farms_Rural
	A.0-1_4_R	Large permanent crops farms_Rural
	A.0-1_5_R	Small other farms_Rural
	A.0-1_6_R	Large other farms_Rural
	A.0-2_R	Products of forestry, logging and related services_Rural
	A.0-3_R	Fish_Rural
	A.0-4_R	Mining_Rural
	A.0-5_R	Food industries_Rural
	A.0-6_R	Other manufacturing_Rural
	A.0-7_R	Utilities_Rural
	A.0-8_R	Construction_Rural
	A.0-9_R	Trade_Rural
	A.0-10_R	Hotels and restaurants_Rural
	A.0-11_R	Transport and communication_Rural
	A.0-12_R	Other private services_Rural
	A.0-13_R	Public services_Rural

Table A4 (cont.)

Urban activities	A.0-1_1_U	Small arable crops farms_Urban
	A.0-1_2_U	Large arable crops farms_Urban
	A.0-1_3_U	Small permanent crops farms_Urban
	A.0-1_4_U	Large permanent crops farms_Urban
	A.0-1_5_U	Small other farms_Urban
	A.0-1_6_U	Large other farms_Urban
	A.0-2_U	Products of forestry, logging and related services_Urban
	A.0-3_U	Fish_Urban
	A.0-4_U	Mining_Urban
	A.0-5_U	Food industries_Urban
	A.0-6_U	Other manufacturing_Urban
	A.0-7_U	Utilities_Urban
	A.0-8_U	Construction_Urban
	A.0-9_U	Trade_Urban
	A.0-10_U	Hotels and restaurants_Urban
	A.0-11_U	Transport and communication_Urban
	A.0-12_U	Other private services_Urban
	A.0-13_U	Public services_Urban
Commodities	C.0-1_1	Arable crops products
	C.0-1_2	Permanent crops products
	C.0-1_3	Other agricultural products
	C.0-2	Products of forestry, logging and related services
	C.0-3	Fish
	C.0-4	Mining
	C.0-5	Food industries
	C.0-6	Other manufacturing
	C.0-7	Utilities
	C.0-8	Construction
	C.0-9	Trade
	C.0-10	Hotels and restaurants
	C.0-11	Transport and communication
	C.0-12	Other private services
	C.0-13	Public services
Factors	SL	Skilled labour
	UL	Unskilled labour
	K	Capital
Taxes (net)	ANT	Activity net taxes
	CNT	Commodity net taxes
	INT	Income net taxes
Institutional sectors	RH	Rural households
	UH	Urban households
	E	Enterprises
	G	Government
Investment/Save	IS	I-S
Rest of the world	ROW	Rest of the world

Source: Own elaboration.

	Commodities	**Activities**	**Factors**	**Households**	**Enterprises**	**Government**	**Savings-Investment**	**Rest of the World**	***Total***
Commodities		Intermediate consumption (inputs)		Household consumption		Government consumption	Fixed capital formation and change in stock (Investment)	Exports	*Total demand*
Activities	Marketed output / Domestic sales								*Activity income*
Factors		Factor income from activities						Factor income from ROW	*Factor income*
Households			Labour and mixed income	(Inter Households transfers)	Distributed benefits to Households / Other transfers	Current transfers to Households		Current transfers to Households from ROW	*Household income*
Enterprises			Operating surplus / Capital income			Current transfers to Enterprises		Current transfers to Enterprises from ROW	*Enterprise income*
Government	Net taxes on products	Net taxes on production	Factor income to Government	Direct taxes	Surplus to Government / Enterprises taxes			Current transfers to Government from ROW	*Government income*
Savings-Investment				Household savings	Enterprise savings	Government savings	(Capital accounts transfers)	Capital transfers from ROW	*Savings*
Rest of the World	Imports		Factor income to ROW	Household transfers to ROW	Surplus to ROW	Government transfers to ROW	Current external balance		*Foreign exchange outflow*
Total	*Total supply*	*Cost of production activities*	*Factor income payments*	*Household expenditures*	*Enterprise expenditures*	*Government expenditures*	*Investment*	*Foreign exchange inflow*	

Figure A.1: *Structure of a Social Accounting Matrix.*

Source: Round (2003) and own elaboration.

Annex 2. Social accounting matrices[14]-NUTS 3 regions 2007 (mio EUR, current prices)

Lüneburg

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					280								280
OtR					2,866								2,866
AgU					50								50
OtU					3,618								3,618
COM	196	1,596	36	1,530			475	1,385		673	540	1,639	8,069
FACT	93	1,198	15	2,003								23	3,332
RH						714			140	199		2	1,055
UH						1,976			388	552		5	2,920
ENT						620						360	979
GOV	-8	73	-1	85			374	1,093	109	0	50	15	3,565
I-S							193	402	50	349			995
ROW					1,254	22	14	41	292	16	405		2,043
Tot	280	2,866	50	3,618	8,069	3,332	1,055	2,920	979	3,565	995	2,043	

14. AgR: Agricultural and food activities_Rural; OtR: Other activities_Rural; AgU: Agricultural and food activities_Urban; OtU: Other activities_Urban; COM: Commodities; FACT: Factors; RH: Rural households; UH: Urban households; ENT: Enterprises; GOV: Government (incl. taxes); I-S: I-S; ROW: Rest of the world; Tot: Total.

Norfolk

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					534								534
OtR					12,019								12,019
AgU					1,017								1,017
OtU					18,735								18,735
COM	317	5,757	728	8,784			4,282	5,082		5,387	3,394	6,842	40,574
FACT	214	5,115	320	8,139								135	13,923
RH						6,233			1,366	4,027		96	11,722
UH						7,201			1,578	4,653		111	13,544
ENT						423						4,554	4,977
GOV	3	1,147	-31	1,812			5,561	6,544	655	0	776	65	32,995
I-S							1,708	1,717	355	2,372			6,151
ROW					8,269	66	171	201	1,023	92	1,982		11,804
Tot	534	12,019	1,017	18,735	40,574	13,923	11,722	13,544	4,977	32,995	6,151	11,804	

Konstanz

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					676								676
OtR					10,104								10,104
AgU					67								67
OtU					3,721								3,721
COM	458	4,891	47	1,828			559	3,615		1,311	1,456	2,968	17,134
FACT	230	4,990	21	1,802								50	7,093
RH						660			288	214		2	1,164
UH						4,046			1,766	1,312		12	7,136
ENT						2,336						768	3,104
GOV	-12	222	0	90			398	2,579	233	0	122	52	7,329
I-S							190	840	192	810			2,032
ROW					2,567	51	16	102	625	36	454		3,851
Tot	676	10,104	67	3,721	17,134	7,093	1,164	7,136	3,104	7,329	2,032	3,851	

Finistère

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					2,621								2,621
OtR					24,450								24,450
AgU					399								399
OtU					11,637								11,637
COM	1,704	12,078	288	5,411			7,601	3,282		5,298	4,452	6,682	46,795
FACT	918	11,476	110	5,793								132	18,429
RH						10,211			1,528	3,137		78	14,954
UH						5,001			748	1,536		38	7,324
ENT						3,193						2,006	5,199
GOV	-1	896	2	432			6,018	2,637	616	0	376	61	22,033
I-S							1,063	1,286	530	879		1,070	4,828
ROW					7,689	24	272	119	1,776	187			10,066
Tot	2,621	24,450	399	11,637	46,795	18,429	14,954	7,324	5,199	22,033	4,828	10,066	

Gorenjska

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					396								396
OtR					6,264								6,264
AgU					10								10
OtU					1,722								1,722
COM	294	4,275	7	1,134			1,283	484		660	1,135	6,279	15,550
FACT	115	1,845	4	549								74	2,586
RH						1,649			48	364		29	2,090
UH						615			18	136		11	780
ENT						243						77	320
GOV	-13	144	-1	39			651	245	94	0	37	23	2,419
I-S							109	33	10	37		982	1,171
ROW					7,158	78	48	18	150	22			7,475
Tot	396	6,264	10	1,722	15,550	2,586	2,090	780	320	2,419	1,171	7,475	

Noord-Drenthe

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					168								168
OtR					633								633
AgU					673								673
OtU					9,848								9,848
COM	109	399	497	5,689			1,318	761		1,342	966	5,811	16,893
FACT	58	217	178	3,964								56	4,472
RH						2,116			476	837		26	3,455
UH						1,161			261	459		14	1,895
ENT						1,085						1,952	3,037
GOV	1	16	-2	195			1,497	868	164	0	89	27	5,695
I-S							564	223	318	164			1,269
ROW					5,572	110	76	44	1,817	54	213		7,886
Tot	168	633	673	9,848	16,893	4,472	3,455	1,895	3,037	5,695	1,269	7,886	

Örebro

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					685								685
OtR					10,838								10,838
AgU					79								79
OtU					6,444								6,444
COM	420	6,083	44	3,207			1,704	1,822		2,452	1,701	6,191	23,626
FACT	289	4,388	39	2,946								53	7,715
RH						3,255			350	1,042		29	4,675
UH						3,090			332	989		27	4,438
ENT						1,303						373	1,676
GOV	-23	367	-4	290			1,929	2,062	258	0	148	27	10,090
I-S							953	459	184	515			2,111
ROW					5,579	67	89	95	552	55	262		6,700
Tot	685	10,838	79	6,444	23,626	7,715	4,675	4,438	1,676	10,090	2,111	6,700	

Heves

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					334								334
OtR					2,986								2,986
AgU					198								198
OtU					1,859								1,859
COM	216	1,878	142	1,129			659	303		483	475	2,449	7,734
FACT	132	1,059	59	697								79	2,025
RH						1,260			55	374		19	1,708
UH						515			22	153		8	698
ENT						225						178	403
GOV	-13	49	-3	33			703	314	63	0	30	25	2,406
I-S							312	65	60	175			613
ROW					2,357	25	34	15	202	17	107		2,758
Tot	334	2,986	198	1,859	7,734	2,025	1,708	698	403	2,406	613	2,758	

Lääne-Eesti

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					268								268
OtR					1,523								1,523
AgU					51								51
OtU					1,111								1,111
COM	170	905	32	625			449	255		259	546	1,094	4,334
FACT	106	590	21	466								57	1,240
RH						657			55	109		12	832
UH						359			30	59		7	455
ENT						213						77	291
GOV	-9	28	-2	19			322	176	22	0	23	12	1,172
I-S							48	17	8	163		333	569
ROW					1,383	11	14	7	176	1			1,592
Tot	268	1,523	51	1,111	4,334	1,240	832	455	291	1,172	569	1,592	

Slupski

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					503								503
OtR					2,455								2,455
AgU					316								316
OtU					3,361								3,361
COM	330	1,322	240	2,005			1,244	725		698	858	1,474	8,895
FACT	179	1,069	77	1,274								110	2,709
RH						1,551			60	236		30	1,877
UH						908			35	138		17	1,099
ENT						233						26	259
GOV	-6	63	-1	81			575	337	92	0	36	7	2,365
I-S							35	24	22	108		704	893
ROW					2,261	18	22	13	50	3			2,367
Tot	503	2,455	316	3,361	8,895	2,709	1,877	1,099	259	2,365	893	2,367	

Huesca

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					2,320								2,320
OtR					6,574								6,574
AgU					149								149
OtU					1,840								1,840
COM	1,682	3,126	107	865			998	2,558		1,307	963	2,466	14,072
FACT	690	3,061	45	870								76	4,741
RH						1,060			173	258		99	1,591
UH						2,719			445	663		254	4,081
ENT						854						357	1,211
GOV	-53	388	-3	105			409	1,049	92	0	30	259	4,361
I-S							173	442	424	33			1,072
ROW					3,189	108	11	33	77	13	79		3,511
Tot	2,320	6,574	149	1,840	14,072	4,741	1,591	4,081	1,211	4,361	1,072	3,511	

Setúbal

	AgR	OtR	AgU	OtU	COM	FACT	RH	UH	ENT	GOV	I-S	ROW	Tot
AgR					1,072								1,072
OtR					12,648								12,648
AgU					133								133
OtU					4,759								4,759
COM	751	6,954	97	2,349			1,419	4,009		2,024	2,101	4,939	24,643
FACT	311	5,302	33	2,277								874	8,797
RH						2,142			639	778		80	3,639
UH						4,614			1,377	1,675		172	7,838
ENT						2,016						689	2,705
GOV	10	392	3	133			1,149	3,201	325	0	80	-350	9,921
I-S							1,029	513		433		206	2,181
ROW					6,031	26	41	115	365	33			6,611
Tot	1,072	12,648	133	4,759	24,643	8,797	3,639	7,838	2,705	9,921	2,181	6,611	

References

Beutel, J. (2002). *The Economic Impact of Objective 1 Interventions for the Period 2000–2006*. Report to the Directorate-General for Regional Policies.

Beutel, J. (2008). *An Input-Output System of Economic Accounts for the EU Member States*. Report to the European Commission, Directorate-General Joint Research Centre, Institute for Prospective Technological Studies.

Beutel, J. and Rueda-Cantuche, J.M. (2012). The Euro method revisited for updating supply and use tables, presented at the *Third Workshop on Input-Output Analysis of the Hispanic-American Input-Output Society*, Santiago de Compostela, October 2012.

Cardenete, M.A. and Sancho, F. (2004). Sensitivity of CGE simulation results to competing SAM updates. *The Review of Regional Studies*, 34, 37–56.

Eurostat (2008). *The Eurostat Manual of Supply, Use and Input-Output Tables*, Luxembourg: Eurostat, 461-475.

Golan, A., Judge G. and Robinson, S. (1994). Recovering information from incomplete or partial multisectoral economic data. *Review of Economics and Statistics*, 76, 541–549.

Hewings, G.J.D. (1969). Regional input-output models using national data: the structure of the west midlands economy. *The Annals of Regional Science*, 3, 179–191.

Hewings, G.J.D. (1977). Evaluating the possibilities for exchanging regional input-output coefficients. *Environment and Planning A*, 9, 927–944.

Jackson, R. and T. Murray (2003). *Alternate Input-Output Matrix Updating Formulations*, Research paper 2003-16, Regional Research Institute, West Virginia University.

Le Masné, P. (1990). Le système productif français face à ses voisins européens in *La Comptabilité Nationale face au Défi International*. Ed. E. Archanmault, O. Arkhipoff (Economica, París).

Lenzen, M., Gallego B. and Wood, R. (2009). Matrix balancing under conflicting information. *Economic Systems Research*, 21, 23–44.

Mínguez, R., J. Oosterhaven and F. Escobedo (2009). Cell-corrected ras method (CRAS) for updating or regionalizing an input-output matrix. *Journal of Regional Science*, 49, 329–348.

OECD (2015). *Input-Output Tables*. http://www.oecd.org/trade/input-outputtables.htm https://stats.oecd.org/Index.aspx?DataSetCode=IOTS

Oosterhaven, J. and Escobedo, F. (2011). A new method to estimate input-output tables by means of structural lags, tested on spanish regions. *Papers in Regional Science*, 90, 829–845.

Pyatt, G. and Round, J. (1985). *Social Accounting Matrices: a Basis for Planning*. Washington: The World Bank.

Pyatt, G. and Thorbecke, E. (1976). *Planning Techniques for a Better Future*. Geneva: International Labour Office.

Raggi, M., Mary, S., Santini, F. and Gómez y Paloma S. (2013). A classification of european NUTS 3 regions. *JRC Technical Report*, JRC85163.

Robinson, S., Cattaneo, A. and El-Said, M. (2001). Updating and estimating a social accounting matrix using cross entropy methods. *Economic Systems Research*, 13.

Round, J. (2003). Constructing SAMS for development policy analysis: lessons learned and challenges ahead. *Economic Systems Research*, 15, 161–183.

Salton, G. and M.J. McGill (1983). *Introduction to Modern Information Retrieval*, McGraw-Hill: New York.

Stone, R. (1961). *Input-Output and National Accounts*. Paris, Organisation for European Economic Cooperation.

Temurshoev, U. and Timmer, M. (2011). Joint estimation of supply and use tables. *Papers in Regional Science*, 90, 863–882.

Temurshoev, U., Webb, C. and Yamano, N. (2011). Projection of supply and use tables: methods and their empirical assessment. *Economic Systems Research*, 23, 91–123.

Thissen, M., Diodato, D., and van Oort, F. G. (2010). *Integration and Convergence in Regional Europe?: European Regional Trade Flows from 2000 to 2010* (pp. 1–32). The Hague: PBL Netherlands Environmental Assessment Agency.

Thissen, M. and H. Lofgren (1998). A new approach to sam updating with an application to Egypt. *Environment and Planning A*, 30, 1991–2003.

Valderas, J.M., Rueda-Cantuche, J.M., Olmedo, E. and Beutel, J. (2016). *The SUT-EURO and the SUT-RAS methods: extensions and fair comparison*s. 24th International Input-Output Conference, Seoul, 4-8 July.

SORT 41 (2) July-December 2017, 347-372 DOI: 10.2436/20.8080.02.63

Joint models for longitudinal counts and left-truncated time-to-event data with applications to health insurance

Xavier Piulachs[1], Ramon Alemany[1], Montserrat Guillén[1]
and Dimitris Rizopoulos[2]

Abstract

Aging societies have given rise to important challenges in the field of health insurance. Elderly policyholders need to be provided with fair premiums based on their individual health status, whereas insurance companies want to plan for the potential costs of tackling lifetimes above mean expectations. In this article, we focus on a large cohort of policyholders in Barcelona (Spain), aged 65 years and over. A shared-parameter joint model is proposed to analyse the relationship between annual demand for emergency claims and time until death outcomes, which are subject to left truncation. We compare different functional forms of the association between both processes, and, furthermore, we illustrate how the fitted model provides time-dynamic predictions of survival probabilities. The parameter estimation is performed under the Bayesian framework using Markov chain Monte Carlo methods.

MSC: 62J99, 62N01, 62P05.

Keywords: Joint models, panel count data, left truncation, Bayesian framework, health insurance.

1. Introduction and motivation

The developed world is experiencing significant growth in its elderly population, which not only means people are living longer, but that they tend to face a greater number of years affected by a range of health problems. In the context of health insurance, the changing demographic structure of the population leads to a steady rise in demand for medical services, while the increasing usage of health care systems, in turn, extends longevity even further. This is especially true of private health insurance policyholders, as they are assumed to enjoy greater preventive care than the rest of the population (see e.g., Dow et al., 2010; Chen et al., 2012). Given these circumstances, assessing the

[1] *Corresponding author:* xavier.piulachs@ub.edu. Department of Econometrics, Riskcenter-IREA, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain.

[2] Department of Biostatistics, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands.

Received: May 2016

Accepted: July 2017

relationship between subject-specific medical history and time until death is of obvious interest for elderly policyholders, as they seek fair premiums. Likewise, insurance companies share this interest, as the must determine the potential costs associated with people living longer than mean expectations. Building on such a scheme, joint modelling techniques are postulated as a proper way to relate the historical information on medical records and the time-to-event outcomes.

The research is conducted on a real health insurance dataset of insured subjects aged 65 years and over, a cohort that requires critical medical care more frequently than their younger counterparts, and, consequently, they have more difficulties in finding private coverage at a reasonable price. Our data contain information both on their health care use and lifespan, and we aim to explain, at subject-level, the underlying mortality risk using the relationship between emergency medical services demanded and time until death. Specifically, the variable of interest in the longitudinal part is the annual rate of emergency claims, including ambulance services, hospitalizations, and non-routine visits. The data only consider the subjects who reach the age of 65, defined in the study as the pre-specified time zero. This assumption has two practical consequences: a) those subjects who die before 65 years of age are not observed, and therefore their time-to-event outcomes are not included, and b) all subjects entering the study after the age of 65 are considered as delayed entries, so their time-to-event data are left-truncated further than the usual censorship (Uzunogullari and Wang, 1992; Klein and Moeschberger, 2003), and not all subjects present the same number of longitudinal measurements. In order to avoid an overestimation of the survival probabilities, a proper consideration of the left truncation issue in the mortality risk is achieved by using the subject's age above 65 years as the particular time scale (Lamarca et al., 1998; Thiébaut and Bénichou, 2004).

The relationship between longitudinal and time-to-event processes can be properly analysed using a shared-parameter joint model (JM), where the corresponding outcomes are stochastically correlated by means of a common latent structure. Using this approach, longitudinal and event times are independent given the random effects, as are repeated measurements in the longitudinal process. Complete overviews of the joint modelling techniques can be found in Tsiatis and Davidian (2004) and Yu, Taylor and Sandler (2008). An exhaustive explanation of the shared-parameter JM, with different examples, is provided by Rizopoulos (2012). In the context of the application of joint modelling techniques to health insurance studies, previous work can be found in Piulachs et al. (2015), where the study focused on elderly policyholders and the counting process was approximated by a log-transformation of the longitudinal outcome.

Given the discrete nature of emergency claims per year, the longitudinal response must account for non-Gaussian data. Previous approaches of this kind have been proposed. For example, Rizopoulos and Ghosh (2011) defined a Bayesian JM to relate multiple longitudinal outcomes (discrete or continuous) to a time-to-event outcome. Murawska, Rizopoulos and Lessaffre (2012) presented a two-stage JM where the longitudinal information was summarized by either a non-linear mixed-effects model or a generalized linear mixed model (GLMM) in the first stage, while in the second the Em-

pirical Bayes estimates of the subject-specific parameters were included as predictors in the proportional hazards model. Viviani, Alfó and Rizopoulos (2012) implemented an expectation-maximization algorithm to incorporate non-Gaussian data in the longitudinal response, with particular attention to Poisson and binomial mixed models. More recently, Ivanova, Molenberghs and Verbeke (2016) formulated a JM to handle different types of responses, i.e., continuous, discrete and ordinal. Parameters were estimated under a likelihood-based approach.

A common feature of the aforementioned extensions is that they do not account for delayed entries in the time-to-event sub-model. In contrast, we consider here the lifetime elapsed from the moment a subject is 65 until his or her death. As a consequence, left truncation has to be accounted for in survival times of these subjects entering the study above the age of 65. Additionally, most event times cease to be observed at administrative closure of study, whereas some others are not completely observed due to dropout. In order to simultaneously deal with left-truncated and right-censored event times, a Cox proportional hazards model with time-dependent covariates is used for the survival analysis. Our final goal is to assess, in a personalized manner, the relationship between emergency claims per year and the time until death (i.e. subject's mortality risk) by postulating an appropriate JM. In this regard, we investigate the role played by information contained in medical records and identify a cumulative and fading effect, so that more recent records have a greater influence than older records on the hazard of death. Finally, we illustrate how the fitted JM can also be employed to obtain subject-specific survival estimates. From a statistical perspective, this problem requires an innovative application of a joint framework, where a pronounced dependency pattern between longitudinal and time-to-event outcomes for the elderly is expected. From a methodological perspective, the statistical analysis poses challenges in handling correlated counts in the longitudinal response of the JM, and to incorporate the delayed entries in the survival outcome.

The remainder of the paper is organized as follows. Section 2 includes a description of the study's health insurance dataset, which consists of 5470 policyholders aged 65 years and over. Section 3 presents the specification, under the Bayesian framework, of the proposed JM for longitudinal counts and left-truncated time-to-event outcomes. Section 4 shows the application of the derived JM to our health insurance dataset, and the results are commented. Section 5 illustrates how to obtain personalized and time-dynamic predictions for survival from the fitted JM. Finally, Section 6 presents a final discussion and some concluding remarks.

2. Health insurance dataset

The motivating dataset was provided by a Spanish medical insurance company, and consists of a cohort of 5470 policyholders (37.6% men and 62.4% women), aged 65 years and above, living in the city of Barcelona (Spain). The data contain, for each subject, historical information on emergency claims (use of ambulance services, hospitaliza-

tions, and non-routine visits) from January 1, 2006 to February 1, 2014. We also know the age of each subject upon entry into the study and their age at death or right censoring, where the latter is assumed to be independent of all other survival and covariate information.

A set of control points was fixed at the 31st of December each year throughout the study period, and we collected, for each subject, subsequent measurements of the amount of emergency claims demanded in a calendar year; this time unit is the one used by most insurance companies, and, in general, in actuarial studies. Hence, instead of directly working with the amount of observed counts, the main longitudinal outcome in our study is defined in terms of count rates. In our case, for each subject we observed repeated measurements of emergency claims per year. These measurements were recorded at each of the control points covered by each of the subject-profiles. In this regard, the entry of each subject into the study period was registered in their longitudinal response by the measurement associated with the first control point reached by his or her observed profile. We assumed a *last observation carried forward* approach for handling the longitudinal information, i.e. an observed measurement within each subject's profile remains constant between two subsequent control points. However, not all subject-profiles started to be observed at the beginning of a specific calendar year. This resulted in the first measurement of emergency claims having an exposure time less than one year. We therefore needed to explicitly consider exposure effects in order to avoid spurious effect estimates (Cameron and Trivedi, 1998). This procedure was carried out by relating the amount of emergency claims observed at the end of a calendar year to the corresponding exposure, i.e. counts/exposure, thus taking into account the real period-at-risk in which the aforementioned amount was collected. Since this premise assumes that the likelihood of a emergency claim is constant over time, very large (and therefore unrealistic) values of count rates could be obtained in case of very small exposures. These cases were avoided by imposing a rule that each of the values registered for a subject must have been obtained from an exposure above half a year.

Table 1: *Descriptive statistics of observed emergency claims per year stratified by event indicator.*

Death	Subjects	*Emergency claims per year* summary				
		Mean	SD	Min	Max	% Zeros
No	4961	0.80	1.55	0	20	63.8
Yes	509	1.50	2.45	0	18	52.4
Overall	5470	0.84	1.63	0	20	63.1

The longitudinal outcome across all count rates ranges from 0 to 20 emergency claims per year, and the overall mean and variance values are 0.84 claims/year and 2.66 $(\text{claims/year})^2$, respectively, suggesting a marked heterogeneity in the response (see Table 1). A large number of zeros are exhibited in the longitudinal outcome, representing 63.1% of the overall measurements. Here, it must be pointed out that the Spanish health system offers universal coverage, so a rate count of zero may occur either because insur-

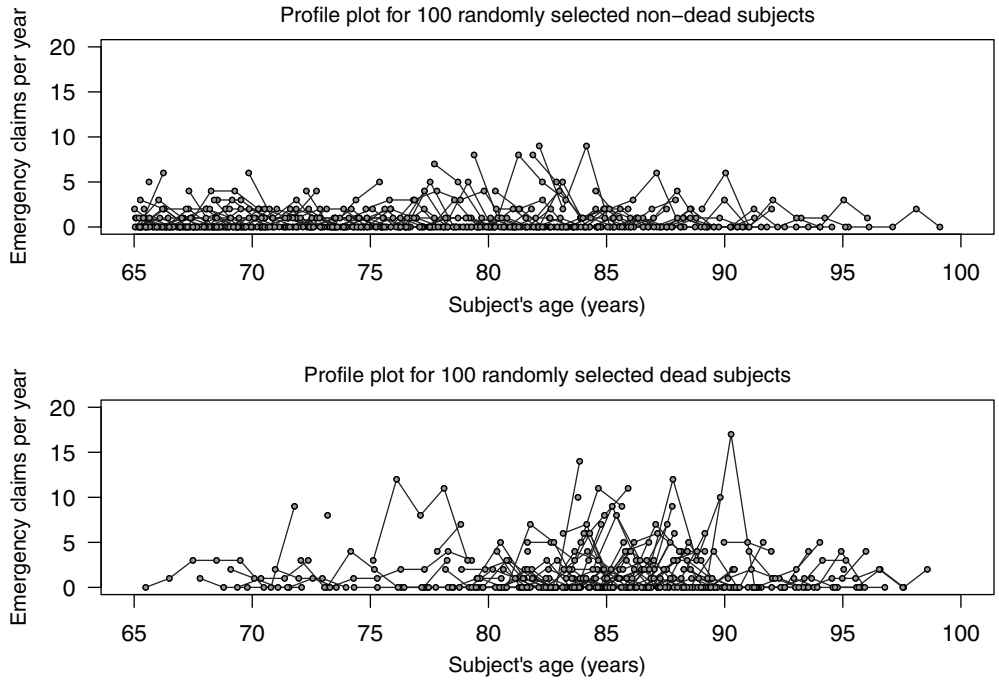

Figure 1: *Subject-profiles of emergency claims per year across time (subject's age) for 100 randomly selected subjects who are still alive after their follow-up interval (top panel) and for 100 randomly selected subjects whose death is observed (bottom panel).*

ance coverage is used solely for routine medical care, or due to the fact that policyholders have only been treated in public medical centers. This circumstance is, in general, an important source of overdispersion in the longitudinal response.

Figure 1 shows various subject-profiles of emergency claims per year, where measurements collected for each subject are connected by line segments. The top panel shows the trajectories for a random sample of 100 subjects alive after their follow-up interval, while the bottom panel shows 100 randomly selected profiles of subjects whose death event is recorded during the study period. Notice that the group of subjects who died during the study presents, in average terms, higher longitudinal responses than those presented by the subjects who remain alive.

Following the suggestions of Charpentier (2015), we also analysed the evolution of the average demand for emergency claims per year according to policyholder's age. We fitted the average values by a generalized additive model (GAM) under the Poisson (PO) and negative binomial (NB) distributions (see Figure 2), and a changing trend was detected around the age of 90 years. Thus, our data show that the use of emergency services in the health insurance company decreases among those subjects of an advanced age. This may reflect the fact that a fraction of the elderly population have taken up residence in nursing homes at older ages, and thus, receive personalized care, or it might be a result of a preference for public over private treatment for severe conditions.

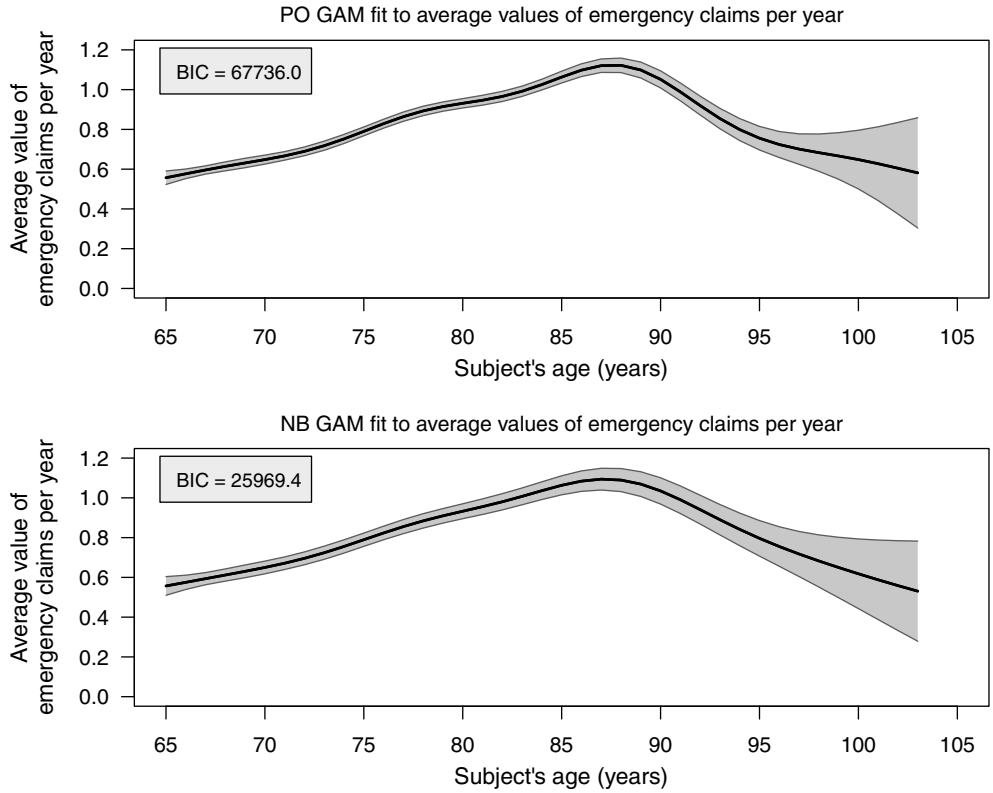

Figure 2: *Observed annual rates of emergency claims by age, with PO and NB GAM fittings. The 95% confidence regions are presented.*

Only policyholders living above the age of 65 are considered within the study period 2006-2014, which means that 79.8% of subjects are registered as late entries. The mean age of policyholders entering the study is 75.4 years (i.e., 10.4 years above the pre-specified time zero), with an average follow-up of 5.1 years. Furthermore, a classic right censoring mechanism arises, which is assumed to be independent of all other survival information. During the study period, death is recorded for a total of 509 (9.3%) individuals, entailing that 4961 policyholders survive or are no longer under observation at the end of the study, representing 90.7% of right censoring. Of these, 3429 (69.1%) are alive at the administrative closure of study, on February 1, 2014. The remaining 1532 right-censored survival times (30.9%) are attributable to insurance cancellations caused by different reasons not related to the event of death (e.g., dissatisfaction with the medical services, a change of company, or an unwillingness to pay), which in practice means that the subject is no longer covered by the insurance policy. Figure 3 displays a non-parametric survival curve estimate of the overall sample (on the left) and one stratified by gender (on the right). Although higher survival estimates are registered for women, the corresponding log-rank test does not suggest a significant improvement in women's survival when stratifying by gender ($p = 0.242$).

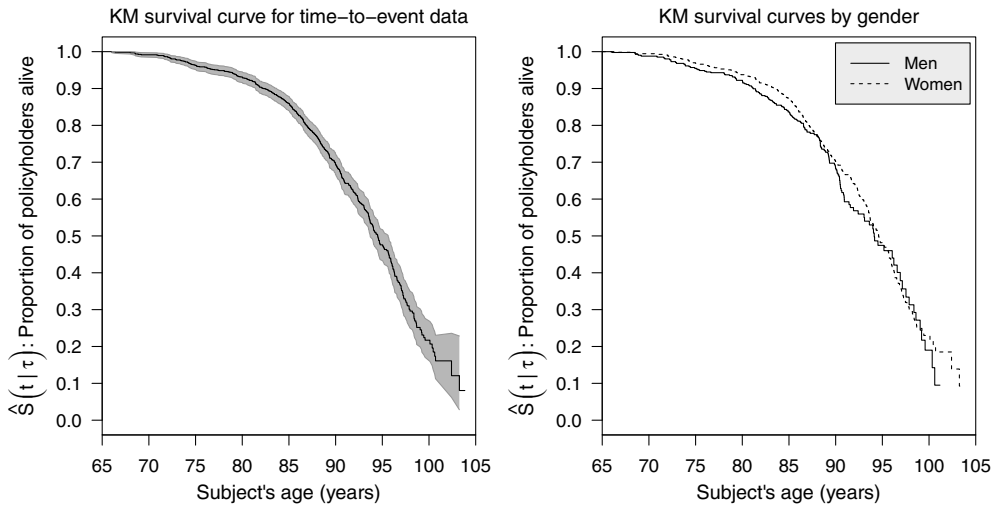

Figure 3: *Kaplan-Meier estimate of the survival function of time until death (with 95% confidence intervals) for our overall health insurance dataset (left panel), and stratified by gender (right panel).*

3. Joint model specification

3.1. Longitudinal approach to panel count data

Let us assume a panel data context with repeated measurements over time, where $\boldsymbol{y}_i = \{y_i(t),\ i = 1,\ldots,n\}$ denote the observed responses for the i-th subject, recorded at a fixed set of time points $t_{ij},\ j = 1,\ldots,n_i$. Given the vector $\boldsymbol{b}_i$ of random effects for the i-th subject, we assume that the observed measurements on this individual derive from a counting process generated by an exponential family (EF) distribution, $y_i(t) \,|\, \boldsymbol{b}_i \sim \text{EF}\{\psi_i(t), \phi\}$, with probability mass function:

$$p_y\{y_i(t) \,|\, \boldsymbol{b}_i; \psi_i(t), \phi\} = \exp\Big(\phi^{-1}\big[y_i(t)\,\psi_i(t) - b\{\psi_i(t)\}\big] + c\{y_i(t), \phi\}\Big). \qquad (1)$$

Here, $b(\cdot)$ and $c(\cdot)$ are known functions, and $\psi_i(t)$ and ϕ are termed the canonical and scale parameters, respectively. It can be shown straightforwardly that $\mathbb{E}\{y_i(t) \,|\, \boldsymbol{b}_i\} = \mu_i(t) = b'\{\psi_i(t)\}$ and $\mathbb{V}\{y_i(t) \,|\, \boldsymbol{b}_i\} = \sigma_i^2(t) = \phi\, b''\{\psi_i(t)\}$ (Molenberghs and Verbeke, 2005).

In many longitudinal studies, the subject-specific count responses are observed within a pre-specified time interval, and can be implicitly interpreted as frequency rates. In such cases, modelling the count rates is more relevant than working with the raw counts, thus considering the expected longitudinal outcome $\mu_i(t)$ in terms of counts per time unit. In our case, a calendar year is taken as the reference time frame during which emergency claims uniformly occur, but a small percentage of subjects start to be observed after the

beginning of a calendar year (i.e., their first longitudinal measurement is not recorded for the duration of a whole year). With this data pattern, the set of observations for a specific subject at their corresponding control points might have occurred during different lengths of time, making it necessary to explicitly consider exposure effects. For the i-th subject at time t, an exposure term $e_i(t)$ is included as a predictor variable of the expected longitudinal outcome $\mu_i(t)$. In addition, it is necessary to introduce a continuous and differentiable link function $g(\cdot)$ in order to relate $\mu_i(t)$ to a linear combination $\eta_i(t)$ of a set of fixed and random covariates. The most common choice for modelling panel count rates is a logarithmic link, $g(\cdot) = \log(\cdot) \Rightarrow g^{-1}(\cdot) = \exp(\cdot)$, which ensures positive outcomes and provides a straightforward interpretation of the estimated regression parameters:

$$\begin{cases} \log\{\mu_i(t)\} = \log\{e_i(t)\} + \eta_i(t) = \log\{e_i(t)\} + \boldsymbol{x}_i^{\intercal}(t)\boldsymbol{\beta} + \boldsymbol{z}_i^{\intercal}(t)\boldsymbol{b}_i \\ \mathbb{E}\{y_i(t) \,|\, \boldsymbol{b}_i\} = \mu_i(t) = e_i(t)\exp\{\eta_i(t)\}, \;\; \boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D}_{q+1}). \end{cases} \tag{2}$$

Note in the above equation that the exposure term is logged and included as an offset variable, i.e., a predictor whose coefficient is fixed at one. If we move the exposure to the left side of the equation, we evince the fact that our expected outcome is divided by the length of time, $\mu_i(t)/e_i(t)$, so we are effectively modelling the expected response in terms of rate counts. The terms $\boldsymbol{x}_i^{\intercal}(t)$ and $\boldsymbol{z}_i^{\intercal}(t)$ denote the row vectors of the fixed and random design matrices, respectively, while $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\intercal}$ and $\boldsymbol{b}_i = (b_{i0}, b_{i1}, \ldots, b_{iq})^{\intercal}$ are the corresponding fixed-effects and random-effects vectors. The random effects allow for the expression of individual deviations from the overall trend, and in most cases they can be assumed to follow a multivariate normal distribution with zero mean and unspecified variance-covariance matrix $\boldsymbol{D}_{q+1}$.

The basic option for modelling panel counts in equation (2) is to consider a PO mixed model, defined as:

$$\begin{cases} y_i(t) \,|\, \boldsymbol{b}_i \sim \text{PO}\{\mu_i(t)\}, \;\; \mu_i(t) > 0 \\ \mu_i(t) = e_i(t)\exp\{\eta_i(t)\} = e_i(t)\exp\{\boldsymbol{x}_i^{\intercal}(t)\boldsymbol{\beta} + \boldsymbol{z}_i^{\intercal}(t)\boldsymbol{b}_i\} \\ p_y\{y_i(t) \,|\, \boldsymbol{b}_i; \mu_i(t)\} = \dfrac{\exp\{-\mu_i(t)\}\,\mu_i(t)^{y_i(t)}}{y_i(t)!} \\ \mathbb{E}\{y_i(t) \,|\, \boldsymbol{b}_i\} = \mathbb{V}\{y_i(t) \,|\, \boldsymbol{b}_i\} = \mu_i(t). \end{cases} \tag{3}$$

The PO mixed model allows for robust parameter estimates, even if the underlying distribution is not true, provided that the expectation is correctly specified (Gourieroux, Monfort and Trognon, 1984). However, the observed response usually has a variance greater than the mean, so the longitudinal outcome is affected by overdispersion. This is a common issue when dealing with counts or count rates, primarily due to missing information, aggregate data, or even an excess of zeros in the longitudinal outcome (Harrison,

2014). In such cases, the derived inference under the PO mixed model could lead to erroneous conclusions about parameter significance. A detailed discussion of this issue can be found in Zuur et al. (2009) and Hilbe (2011).

Although there are several alternative models for dealing with the overdispersion related to correlated counts, the NB mixed model appears in the literature as being the most natural choice; see, for example, Ismail and Jemain (2007), Greene (2008), and Hilbe (2011). The NB distribution for longitudinal data can be easily derived from the PO distribution by placing a multiplicative gamma random noise ε_i in the conditional mean response. Specifically, such a latent variable is defined in terms of shape and rate parameters by $\varepsilon_i \sim G(\kappa,\kappa)$, $\kappa > 0$, with $\mathbb{E}(\varepsilon_i) = 1$ and $\mathbb{V}(\varepsilon_i) = 1/\kappa$, so that the longitudinal counts are modelled by $y_i(t)\,|\,\boldsymbol{b}_i \sim \text{PO}\{\varepsilon_i\,\mu_i(t)\}$. This Poisson-gamma mixture has a closed-form solution, leading to a NB mixed model with dispersion parameter κ:

$$\begin{cases} y_i(t)\,|\,\boldsymbol{b}_i \sim \text{NB}\{\mu_i(t),\kappa\},\ \mu_i(t) > 0, \kappa > 0 \\ \mu_i(t) = e_i(t)\exp\{\eta_i(t)\} = e_i(t)\exp\{\boldsymbol{x}_i^{\intercal}(t)\boldsymbol{\beta} + \boldsymbol{z}_i^{\intercal}(t)\boldsymbol{b}_i\} \\ p_y\{y_i(t)\,|\,\boldsymbol{b}_i;\mu_i(t),\kappa\} = \dfrac{\Gamma\{\kappa + y_i(t)\}}{\Gamma(\kappa)\,y_i(t)!}\ \dfrac{\mu_i(t)^{y_i(t)}\,\kappa^{\kappa}}{\{\mu_i(t)+\kappa\}^{\kappa+y_i(t)}} \\ \mathbb{E}\{y_i(t)\,|\,\boldsymbol{b}_i\} = \mu_i(t);\ \ \mathbb{V}\{y_i(t)\,|\,\boldsymbol{b}_i\} = \mu_i(t) + \mu_i(t)^2/\kappa, \end{cases} \tag{4}$$

where $\Gamma(\cdot)$ denotes the gamma function.

The NB distribution has the general canonical form of the exponential family equations for any fixed κ. Because of the quadratic expression for the variance, it is sometimes referred to as NB type 2 in the literature. Note that the NB distribution can actually be understood as an extension of the PO distribution when overdispersion is accounted for by parameter κ, since it can be proven that NB converges to PO as $\kappa \to \infty$. This result is well-documented by Lawless (1987) and Hinde and Demétrio (1998); see also Boucher, Denuit and Guillén (2008) for a numerical application in the field of insurance studies.

3.2. Joint model for counts and delayed entries

Assuming the age above 65 years as our particular time scale, let T_i^* be the true event time for the i-th subject. We define an independent random variable $\tau_i \geq 0$ as the time at which a policyholder enters the study after the age of 65, giving rise to left-truncated event times for those subjects whose $\tau_i > 0$. In addition, once a subject enters the study, the event time is affected by the usual right censorship mechanism, denoted by a potential censoring time C_i. This means we can only know the observed survival time for the i-th recruited individual, $T_i = \min\{T_i^*, C_i\} > \tau_i$, and a dichotomous event indicator $\delta_i = \mathbb{I}(T_i^* \leq C_i)$. We use a time-dependent proportional hazards model to simultaneously account for left truncation and right censoring in the time-to-event sub-model. Conse-

quently, the probabilistic distribution of the event times has to be defined according to the proportion of subjects living beyond time point t, and is conditional on their being older than the corresponding left truncation time, $S(t \mid \tau) = \Pr(T^* > t \mid T^* > \tau) = \Pr(T^* > t) \,/\, \Pr(T^* > \tau) = S(t) \,/\, S(\tau)$.

Building on the longitudinal analysis considered in section 3.1, repeated count rates and time-to-event responses can be coupled by assuming independence between both processes given the shared random effects (conditional independence hypothesis). The JM for the i-th subject, $i = 1, \ldots, n$, is expressed by means of a relative risk model where the hazard of death at time t takes into account the expected longitudinal response until t, $M_i(t) = \{\mu_i(s), 0 \leq s \leq t\}$:

$$h_i\{t \mid M_i(t), \boldsymbol{w}_i\} = h_0(t) \exp\left[\boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{w}_i + \alpha F\{\mu_i(t)\}\right]. \tag{5}$$

As in the standard proportional hazards model, $h_0(t)$ in equation (5) denotes the baseline risk function, $\boldsymbol{w}_i$ the subject's baseline survival covariates, and $\boldsymbol{\gamma}$ the vector of the corresponding regression parameters. The functional form $F(\cdot)$ specifies a proper manner in which the longitudinal information provided by $\mu_i(t)$ is accounted for in survival. Because $\mu_i(t) > 0$ in a counting process, $F(\cdot)$ is positively defined and increases with t. The parameter α quantifies the strength of association between the particular longitudinal evolution until time t, and the corresponding mortality risk. Specifically, the quantity $\exp(\alpha)$ returns the hazard ratio for a one-unit increase in the value $F\{\mu_i(t)\}$ at time snapshot t.

Although $h_0(t)$ traditionally remains unspecified in the Cox proportional hazards model, this constraint is usually lifted when using joint modelling techniques. In particular, the logarithm of baseline hazard function can be approximated using penalized B-splines. As a preliminary step, we define a knot sequence $\boldsymbol{\xi}$ of Q increasing and equally-spaced knots, $\xi_1 < \cdots < \xi_Q$, over the time range $[0, T_{max}]$. Accordingly, the baseline hazard on the log-scale is approximated through a linear combination of d-th degree B-splines:

$$\log\{h_0(t)\} = \sum_{r=1}^{R} \gamma_{h_0,r} B_{d,r}(t, \boldsymbol{\xi}), \tag{6}$$

where $\{B_{d,r}(t, \boldsymbol{\xi}),\ r = 1, \ldots, R\}$ denotes the set of d-th degree B-spline basis functions, $\boldsymbol{\gamma}_{h_0} = (\gamma_{h_0,1}, \ldots, \gamma_{h_0,R})^{\mathsf{T}}$ is the vector of B-spline coefficients (also called control points), and $R = Q + d - 1$. The r-th B-spline function is locally defined on a support spanned by the $d+2$ adjacent knots, and, to achieve boundary conditions of a B-spline curve, the original knot vector is extended so that the end-knots ξ_1 and ξ_Q have multiplicity $d+1$ (the total number of knots will be $Q + 2d$). A major concern at this point is the number Q of knots that should be employed. A too-small number of knots (and, consequently, of the number R of B-spline basis functions) could lead to biased results, while too many knots might result in an overly flexible curve with random fluctuations (small "wig-

gles"). Following the indications of Eilers and Marx (1996), a relatively large number of knots should be used, and the potential overfitting problems can be circumvented by considering a roughness penalty based on finite differences of adjacent B-spline coefficients, i.e., by means of a P-splines regression. A complete overview of recent research in P-splines can be found in Eilers, Marx and Durbán (2015).

A standard approach to relate longitudinal rate counts to survival is undertaken by associating the current expected longitudinal outcome with the hazard of an event using the identity function: $F\{\mu_i(t)\} = Id\{\mu_i(t)\} = \mu_i(t)$. However, instead of taking just a single time point, in some cases it may be more relevant to consider the whole path of the longitudinal outcome. In particular, an extension of the basic option is to include the entire background previous to the measurement at time t (Abrahamowicz, Beauchamp and Sylvestre, 2011). Furthermore, we assume that historical effects of subject's health fade over time, so the more distant history is less relevant than the more recent. Thus, $F(\cdot)$ transformation can be defined to account for the recency-weighted cumulative effect of the longitudinal outcome:

$$F\{\mu_i(t)\} = \int_0^t \overline{\omega}(t-s)\,\mu_i(s)\,\mathrm{d}s, \quad s \leq t, \tag{7}$$

where $\overline{\omega}(\cdot)$ is the selected average weighting function. Due to the importance of the most recent information for explaining the current health status, we introduced an exponential function with rate parameter ν in order to assign different weights for each of the past observed longitudinal values: $\overline{\omega}(t-s) = \nu \exp\{-\nu(t-s)\},\ \nu > 0$.

3.3. Bayesian estimation for the JM

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_t, \boldsymbol{\theta}_b)^{\mathsf{T}}$ be the JM full parameter vector that collects the longitudinal parameters, the survival parameters, and the parameters for the random effects covariance matrix, respectively. In addition, let $D_n = \{(\boldsymbol{y}_i, \tau_i, T_i, \delta_i),\ i = 1, \ldots, n\}$ denote the information from our original dataset with n policyholders. Taking advantage of the conditional independence assumption, the overall joint likelihood conditioned on the random effects $\boldsymbol{b}_i$ can be properly formulated to tackle left truncation as

$$p(D_n \,|\, \boldsymbol{b}_i, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} p_y\{y_i(t_{ij}) \,|\, \boldsymbol{b}_i, \boldsymbol{\theta}\}\, \frac{p_t(T_i, \delta_i \,|\, \boldsymbol{b}_i, \boldsymbol{\theta})}{\Pr(T_i > \tau_i \,|\, \boldsymbol{b}_i, \boldsymbol{\theta})}, \tag{8}$$

where $p_y(\cdot)$ is the conditional probability mass function to handle longitudinal rate counts, and $p_t(\cdot)$ is the conditional probability density function for the event times.

The mean estimates of parameters and random effects are then derived by Markov chain Monte Carlo (MCMC) algorithms, which enable inferences to be made by efficiently drawing a sample from the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{b}_i)$ conditioned on the observed data:

$$\pi(\boldsymbol{\theta}, \boldsymbol{b}_i \,|\, D_n) \;\propto\; p(D_n \,|\, \boldsymbol{b}_i, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}, \boldsymbol{b}_i) = p(D_n \,|\, \boldsymbol{b}_i, \boldsymbol{\theta})\, p_b(\boldsymbol{b}_i \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \tag{9}$$

where $p_b(\cdot)$ is the conditional probability density function of the random effects, and $\pi(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$.

The models' fitting was performed using a Bayesian approach, with non-informative priors being used whenever possible. Specifically, for the longitudinal analysis, we used independent univariate vague normal priors for the fixed effects, defined in terms of mean and precision parameters, $\{\beta_0, \beta_1, \beta_{t_{90}}\} \sim N\left(\mu_\beta = 0, \tau_\beta = 10^{-3}\right)$. In the time-to-event sub-model, the log-baseline hazard was approximated by using third-degree B-splines and $Q = 15$ knots, uniformly allocated over the time range $[0,\, T_{max} = 38.5]$. The joint prior for the baseline hazard coefficients was assumed to be normally distributed, $\boldsymbol{\gamma}_{h_0} \sim N(\boldsymbol{\mu}_{\boldsymbol{\gamma}_{h_0}} = \boldsymbol{0},\, \boldsymbol{T}_{\boldsymbol{\gamma}_{h_0}} = \tau_{Bs}\boldsymbol{M}_{\boldsymbol{\gamma}_{h_0}})$, where $\boldsymbol{M}_{\boldsymbol{\gamma}_{h_0}}$ is an appropriate penalty matrix to control the amount of roughness in the precision parameter $\tau_{Bs} \sim G(a_{Bs}, b_{Bs})$. In general, the penalty matrix is defined as $\boldsymbol{M}_{\boldsymbol{\gamma}_{h_0}} = \boldsymbol{\Delta}_K^{\intercal}\boldsymbol{\Delta}_K + 10^{-6}\boldsymbol{I}$, where $\boldsymbol{\Delta}_K$ is the difference matrix of order K, whereas the term $10^{-6}\boldsymbol{I}$ introduces a small "ridge penalty" to avoid a linearly dependent system. A common choice for cubic B-splines is $K = 2$, while for the hyper-prior parameters of τ_{Bs} we used $a_{Bs} = 1$ and $b_{Bs} = 0.005$ as a standard choice for a non-informative prior. The subject's gender was included in the time-dependent proportional hazards model as a dichotomous baseline covariate, w_{gi} (man $= 0$, woman $= 1$), and we made the assumption that the corresponding coefficient follows an improper normal distribution, $\gamma_g \sim N\left(\mu_g = 0, \tau_g = 10^{-3}\right)$.

For the constant association parameter, we assumed $\alpha \sim N\left(\mu_\alpha = 0, \tau_\alpha = 10^{-3}\right)$. When considering the functional form to link the recency-weighted area under the expected longitudinal profile to the time-to-event outcome, a flat prior was assumed for the rate parameter of the exponential weighting function, $\nu \sim U(a_\nu, b_\nu)$. Because $\nu > 0$, we set $a_\nu = 0$, while for the second hyper-parameter it is common to set a large enough positive value to express the uncertainty around ν, say $b_\nu = 20$.

Finally, for the random effects, we used a bivariate standard normal distribution as a prior function, $(b_{i0},\, b_{i1})^{\intercal} \sim N(\boldsymbol{0}, \boldsymbol{D}_2)$, where the terms of the 2×2 unstructured covariance matrix are summarized by $\boldsymbol{D}_2[1,1] = \sigma^2_{b_0}$, $\boldsymbol{D}_2[1,2] = \boldsymbol{D}_2[2,1] = \rho\,\sigma_{b_0}\,\sigma_{b_1}$, and $\boldsymbol{D}_2[2,2] = \sigma^2_{b_1}$. We assumed that the inverse matrix follows a standard Wishart distribution, $\boldsymbol{D}_2^{-1} \sim W(\boldsymbol{I}_2, k_w)$, where the degrees of freedom were established at $k_w = 3$. In the particular case of considering 1 RE, we have $b_{i1} = \sigma_{b_1} = 0$, assuming that $b_{i0} \sim N(\mu_{b_0}, \tau_{b_0})$ with $\mu_{b_0} = 0$ and $\tau_{b_0} \sim G\left(10^{-3}, 10^{-3}\right)$.

3.4. Bayesian model assessment

To compare both the different longitudinal models and joint models, we focused on the analysis of the Bayesian deviance term, which in generic form can be expressed as $\mathrm{D}(\boldsymbol{\theta}, \boldsymbol{b}_i) = -2\sum_{i=1}^{n} \log\{p\,(D_n \,|\, \boldsymbol{b}_i, \boldsymbol{\theta})\}$. In particular, we assessed the goodness-of-fit of a specific model by using the deviance information criterion (DIC) suggested by

Spiegelhalter et al. (2002). This criterion evaluates the fit of a model by balancing model adequacy with model complexity:

$$\mathrm{DIC}(\boldsymbol{\theta}, \boldsymbol{b}_i) = \mathrm{D}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{b}}_i) + 2\,\mathrm{p_D}, \tag{10}$$

where $\mathrm{D}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{b}}_i) = -2\sum_{i=1}^{n} \log\{p\,(D_n \,|\, \bar{\boldsymbol{b}}_i, \bar{\boldsymbol{\theta}})\}$, and the term $\mathrm{p_D} = \overline{\mathrm{D}(\boldsymbol{\theta}, \boldsymbol{b}_i)} - \mathrm{D}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{b}}_i)$ is the effective number of parameters, calculated as the difference between the posterior mean of the deviance and the deviance at the posterior means of the JM parameters. The aforementioned criterion can be reformulated as $\mathrm{DIC}(\boldsymbol{\theta}, \boldsymbol{b}_i) = \overline{\mathrm{D}(\boldsymbol{\theta}, \boldsymbol{b}_i)} + \mathrm{p_D}$, thus reinforcing the idea that this criterion takes into account both the adequacy of the model, assessed through the posterior mean estimate of the deviance, and the number of parameters required, assessed through the penalty term $\mathrm{p_D}$. The score provided by DIC serves in general as the basis for ranking the fitted models, where lower scores correspond to a better model fit. To conclude this section, it is important to point out that the DIC score obtained for a specific model is not a fixed value, but it can be subject to a certain amount of random variability due to its dependency on the MCMC output of the model. Consequently, it will become a key point to get a DIC value derived from a relatively large number of iterations in the MCMC process before reaching convergence in each of the JM parameters.

4. Results from the health insurance dataset

4.1. Longitudinal data analysis

The fixed effects of the longitudinal outcome were set at $\{\beta_0, \beta_1, \beta_{t_{90}}\}$, respectively motivated by the intercept term, the observation time (directly linked to the subject's age), and a binary variable which takes into account the observed downward trend of medical emergency demand after the age of 90, $t_{90} = \mathbb{I}(t \geq 25)$. The longitudinal measurements were fitted using PO and NB mixed models, testing these models with random intercepts (1 RE) and with ρ-correlated random intercepts and random slopes in time (2 RE). In all cases, the general expression for the expected response can be written as $\mu_i(t) = e_i(t)\exp\{\eta_i(t)\} = e_i(t)\exp\{\beta_0 + b_{i0} + (\beta_1 + b_{i1})\,t + \beta_{t_{90}}\mathbb{I}(t \geq 25)\}$, with $(b_{i0},\, b_{i1})^{\mathsf{T}} \sim N(\mathbf{0}, \boldsymbol{D}_2)$.

Setting $\boldsymbol{\theta}_\ell = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_b)^{\mathsf{T}} = \left(\beta_0, \beta_1, \beta_{t_{90}}, \kappa, \sigma_{b_0}, \sigma_{b_1}, \rho\right)^{\mathsf{T}}$, we approximated the posterior distribution $\pi(\boldsymbol{\theta}_\ell, \boldsymbol{b}_i \,|\, \boldsymbol{y}_i)$ by using the Bayesian software JAGS (Plummer, 2003), version 4.2.0. We ran two parallel chains of 25,000 iterations, with a burn-in period (here included in the total number of iterations) of 10,000 iterations. We then kept every 15-th iteration from each chain in order to reduce the autocorrelation in the samples from the posterior distribution, resulting in 2000 total samples. We checked that the chains had a good mixing, and also their convergence to the stationary distribution. For

each model, the DIC was calculated by means of the corresponding Bayesian deviance $\mathrm{D}(\boldsymbol{\theta}_\ell, \boldsymbol{b}_i) = -2\sum_{i=1}^{n} \log p(\boldsymbol{y}_i \,|\, \boldsymbol{b}_i, \boldsymbol{\theta}_\ell)$.

Table 2 provides the posterior summaries of the parameters' distribution for each longitudinal model fitted, and also the corresponding values of DIC. In the case of the two PO mixed models, the DIC provides a lower score for the option which considers 2 RE in the linear predictor. On the other hand, the NB mixed model with 2 RE also provides evidence of an improvement in the DIC score in comparison with the model using 1 RE. Therefore, it is preferable to use a random intercept and a random slope to summarize subjects' profiles. Along with the PO results, the estimates obtained from the NB mixed model with 2 RE highlight the importance of the variability around the intercept fixed effect compared to fluctuations in the slope. Hence, accounting for baseline heterogeneity indeed plays a much more important role to explain the subject's particularities. This could lead us to consider the extent to which the introduction of a random slope is necessary in this case (i.e., when a linear trend is assumed in the specification of $\eta_i(t)$).

Table 2: *Posterior summaries of all parameters for PO and NB mixed models with a different number of random effects. Mean, standard error, 95% credible interval, and DIC are sampled for each parameter from the corresponding posterior distribution.*

	PO mixed model 1 RE				NB mixed model 1 RE			
Parameter	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$
β_0	−1.085	0.0007	−1.147	−1.026	−0.976	0.0015	−1.041	−0.910
β_1	0.032	0.0001	0.029	0.036	0.031	0.0001	0.027	0.035
$\beta_{t_{90}}$	−0.117	0.0008	−0.193	−0.045	−0.189	0.0023	−0.298	−0.086
κ	–	–	–	–	0.998	0.0004	0.948	1.050
σ_{b_0}	1.075	0.0004	1.045	1.107	0.963	0.0006	0.929	0.996
$\mathrm{DIC}(\boldsymbol{\theta}_\ell, \boldsymbol{b}_i)$	76961				74201			
	PO mixed model 2 RE				NB mixed model 2 RE			
Parameter	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$
β_0	−1.158	0.0009	−1.240	−1.075	−0.974	0.0009	−1.049	−0.900
β_1	0.035	0.0001	0.029	0.040	0.029	0.0001	0.024	0.034
$\beta_{t_{90}}$	−0.190	0.0022	−0.281	−0.103	−0.197	0.0015	−0.328	−0.071
κ	–	–	–	–	1.058	0.0007	1.003	1.120
σ_{b_0}	1.750	0.0011	1.662	1.848	1.157	0.0015	1.074	1.239
σ_{b_1}	0.112	0.0001	0.106	0.118	0.074	0.0001	0.069	0.079
ρ	−0.796	0.0003	−0.818	−0.773	−0.614	0.0007	−0.671	−0.547
$\mathrm{DIC}(\boldsymbol{\theta}_\ell, \boldsymbol{b}_i)$	74993				74095			

The overall comparison between the fitted models suggests that the NB mixed models are more adequate to capture the characteristics of the longitudinal data. This result is unsurprising since the two NB mixed models account for response heterogene-

ity through parameter κ, whose mean estimate exhibits strong evidence for overdispersion for both one and two random effects, $\kappa_{1\,\mathrm{RE}} = 0.998$ (95% CI: 0.948, 1.050) and $\kappa_{2\,\mathrm{RE}} = 1.058$ (95% CI: 1.003, 1.120). In particular, the NB mixed model with 2 RE is the one which presents the lowest DIC score among the tested options. Consequently, we decided to include the effect of random slope when overdispersion is accounted for.

In what follows, the longitudinal approach in our JM framework is carried out using a count mixed model with two random intercepts per subject, one for the intercept and the other for the slope. Additionally, the longitudinal analysis will be carried out using either a PO or a NB mixed model, thus allowing us to directly assess how the goodness-of-fit changes when considering the overdispersion effect.

4.2. JM analysis

The JM that we propose is summarized by:

$$\begin{cases} \mu_i(t) = e_i(t)\exp\left\{\beta_0 + b_{i0} + \beta_1 t + b_{i1} t + \beta_{t_{90}} \mathbb{I}(t \geq 25)\right\} \\ h_i\left\{t \,|\, M_i(t), w_{gi}\right\} = h_0(t)\exp\left[\gamma_g w_{gi} + \alpha F\left\{\mu_i(t)\right\}\right] \\ (b_{i0}, b_{i1})^{\intercal} \sim N(\boldsymbol{0}, \boldsymbol{D}_2). \end{cases} \tag{11}$$

The starting point to carry out the JM fits under a Bayesian approach was the `R` package `JMbayes` (Rizopoulos, 2016), taking advantage of the structure of the function `jointModelBayes`$(\cdot)$. However, the code to fit the different joint models in this article was finally written in `JAGS` software, and executed within the `R`-environment. Setting $\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_t, \boldsymbol{\theta}_b)^{\intercal} = (\beta_0, \beta_1, \beta_{t_{90}}, \kappa, \nu, \boldsymbol{\gamma}_{h_0}, \gamma_g, \alpha, \sigma_{b_0}, \sigma_{b_1}, \rho)^{\intercal}$, the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{b}_i \,|\, D_n)$ was approximated by running the MCMC algorithm for 2 parallel chains with a total of 35,000 iterations each, with the first 15,000 discarded as the burn-in period. We kept every 20-th iteration from each chain, resulting in 2000 total samples from the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{b}_i)$. A good mixing and convergence of the 2 chains were assessed, and no autocorrelation was detected in the lag plots.

First, the estimation of JM parameters was conducted by quantifying the degree of association between the current expected value of emergency claims per year at any time t, denoted by $\mu_i(t) = \mathbb{E}\left\{y_i(t) \,|\, \boldsymbol{b}_i\right\}$, and the mortality risk at the same t. The results (given in Table 3) point to a strong association between the annual rate of emergency claims and survival, so each unit increase in the current value of the emergency claims per year involves a $\exp(\overline{\alpha}_{\mathrm{V,PO}}) = 1.47$-fold increase (95% CI: 1.41, 1.54) in the policyholder's mortality risk under the PO longitudinal sub-model, whereas this association parameter leads to a $\exp(\overline{\alpha}_{\mathrm{V,NB}}) = 1.59$-fold increase (95% CI: 1.49, 1.71) if we assume a NB longitudinal sub-model. Thus, we have an increasing relationship between the frequency of use of emergency medical services and the corresponding mortality risk. From a goodness-of-fit perspective, the comparison between the fitted joint models is performed using the DIC, where the use of the NB distribution provides a better fit.

Table 3: *Posterior summaries of all parameters for the JM when accounting for the current value of emergency claims per year. Mean, standard error, 95% credible interval, and DIC are sampled for each parameter from the corresponding posterior distribution.*

	JM with PO sub-model				JM with NB sub-model			
Parameter	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$
Longitudinal								
β_0	−1.174	0.0010	−1.254	−1.085	−1.000	0.0008	−1.072	−0.920
β_1	0.036	0.0001	0.030	0.041	0.031	0.0001	0.026	0.037
$\beta_{t_{90}}$	−0.130	0.0010	−0.212	−0.043	−0.117	0.0018	−0.236	−0.004
κ	–	–	–	–	1.067	0.0004	1.012	1.125
σ_{b_0}	1.780	0.0011	1.685	1.872	1.169	0.0019	1.080	1.247
σ_{b_1}	0.115	0.0001	0.109	0.121	0.076	0.0001	0.070	0.081
ρ	−0.800	0.0003	−0.821	−0.775	−0.611	0.0002	−0.626	−0.595
Survival								
γ_g	−0.287	0.0018	−0.449	−0.132	−0.326	0.0019	−0.483	−0.164
Association								
α	0.387	0.0005	0.342	0.431	0.464	0.0011	0.397	0.534
Goodness-of-fit								
DIC($\boldsymbol{\theta}, \boldsymbol{b}_i$)	93050				86938			

Table 4: *Posterior summaries of all parameters for the JM when accounting for the recency-weighted area under the expected profile of emergency claims per year. Mean, standard error, 95% credible interval, and DIC are sampled for each parameter from the corresponding posterior distribution.*

	JM with PO sub-model				JM with NB sub-model			
Parameter	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$	Mean	SE	$q_{2.5\%}$	$q_{97.5\%}$
Longitudinal								
β_0	−1.166	0.0010	−1.250	−1.089	−1.002	0.0015	−1.068	−0.936
β_1	0.035	0.0001	0.030	0.041	0.031	0.0001	0.026	0.036
$\beta_{t_{90}}$	−0.129	0.0016	−0.212	−0.044	−0.116	0.0020	−0.234	−0.003
κ	–	–	–	–	1.066	0.0005	1.011	1.124
ν	9.572	0.0218	8.154	11.060	9.691	0.0275	8.246	11.218
σ_{b_0}	1.770	0.0016	1.698	1.846	1.175	0.0014	1.096	1.261
σ_{b_1}	0.114	0.0001	1.110	1.119	0.076	0.0002	0.071	0.082
ρ	−0.798	0.0003	−0.820	−0.774	−0.607	0.0003	−0.647	−0.561
Survival								
γ_g	−0.269	0.0019	−0.433	−0.092	−0.298	0.0021	−0.489	−0.124
Association								
α	0.398	0.0005	0.354	0.443	0.480	0.0010	0.422	0.537
Goodness-of-fit								
DIC($\boldsymbol{\theta}, \boldsymbol{b}_i$)	92983				86892			

One of the more interesting features of the JAGS software is its flexibility in choosing the structure association $F(\cdot)$ that captures the relationship between the longitudinal and time-to-event sub-models. The JM estimates in Table 4 were conducted by associating

the recency-weighted area under the expected longitudinal profile with the mortality risk. Specifically, an exponential weighting function was employed, again showing a strong relationship between both processes, so that a $\exp(\overline{\alpha}_{\text{W,PO}}) = 1.49$-fold increase (95% CI: 1.42, 1.56) in the policyholder's mortality risk is inferred for each unit increase in the exponentially-weighted area under the expected profile of emergency claims per year with the PO distribution, and a $\exp(\overline{\alpha}_{\text{W,NB}}) = 1.62$-fold increase (95% CI: 1.53, 1.71) if we assume an underlying NB distribution. The estimated mean rates of the exponential weighting functions for the PO and NB longitudinal outcomes are $\overline{\nu}_{\text{PO}} = 9.57$ (95% CI: 8.15, 11.06) and $\overline{\nu}_{\text{NB}} = 9.69$ (95% CI: 8.25, 11.22), respectively. Thus, in practice, it is shown that only the 0.25 years (i.e., three months) prior to t are strongly related to the current mortality risk. In this regard, note the broad similarity between the association parameters of these results and those obtained in Table 3 for the current-value association structure, thus emphasizing that only the most recent past emergency claims have a real influence on the survival. Once again the DIC indicates that a more accurate claims distribution is achieved under the NB longitudinal sub-model.

Among all the fitted joint models presented in this section, the results indicate that the lower DIC scores are obtained for the functional form which links the recency-weighted area under the expected longitudinal outcome with survival. This becomes an adequate manner to capture the fading effect of emergency medical demand on mortality risk. In particular, the JM which considers a NB longitudinal outcome is the one which provides the lowest DIC score of all, since it includes the overdispersion effect.

4.3. Residual diagnostics and model assessment

After fitting the joint models, it is a primary step to validate all the necessary model assumptions before performing inference. To achieve this validation, we need plots of residuals for each of the two components of the JM, i.e. the longitudinal and the time-to-event sub-models.

For the longitudinal part, the analysis of the plots of residuals is focused on the non-Gaussian mixed models assumed for the joint models with a recency-weighted cumulative effect. We will consider both PO and NB distributions in order to compare their results. However, a direct graphical interpretation of the residuals under these distributions is usually difficult, since the normality and homoscedasticity of the residuals derived from a count model is, in general, not expected. When longitudinal response takes a limited number of low count rates, the scatterplot of the residuals versus the fitted values typically shows a non-homogeneous configuration, the data being grouped on a set of quasi-parallel and curvilinear traces of points according to distinct response values. In such circumstances, it becomes difficult to evaluate the residual plot, even if the model is correctly specified. To solve this limitation, we can obtain continuously distributed residuals by taking advantage of the idea of randomized quantile residuals (Dunn and Smyth, 1996). The underlying idea is based on applying a transformation scale to the original residuals that standardizes them to continuous values between 0

and 1, so that the residuals are obtained by finding the equivalent standard normal deviate for each subject-specific observation in the original data. By delving deep into this work scheme, the quantile residuals can also be directly obtained through a simulation-based approach. In particular, this task has been recently implemented in the R package DHARMa (Hartig, 2017), which standardizes the residuals to uniformly distributed values in the unit interval. As a first step, from JM results we use the fitted longitudinal sub-model $p_y\{y_i(t)\,|\,\boldsymbol{b}_i,\overline{\mu}_i(t),\overline{\kappa}\}$ to simulate a relatively large number M of new longitudinal datasets, $\{\tilde{\boldsymbol{y}}_i^{(m)}\}_{m=1}^{M}$. Then, for a particular subject-specific observation $y_i(t)$ in the original data, we have a set of M simulated values $\{\tilde{y}_i^{(1)}(t),\ldots,\tilde{y}_i^{(M)}(t)\}$, allowing us to obtain its corresponding empirical cumulative distribution function, $\tilde{P}_y\{\tilde{y}_i(t)\,|\,\boldsymbol{b}_i,\overline{\mu}_i(t),\overline{\kappa}\}$. Finally, the quantile residual associated with the original observation $y_i(t)$ is calculated as:

$$r_i^q(t) = \tilde{P}_y\{y_i(t)\,|\,\boldsymbol{b}_i,\overline{\mu}_i(t),\overline{\kappa}\} = \Pr\{\tilde{y}_i^{(m)}(t) \leq y_i(t)\} \in (0,1). \tag{12}$$

Recall that, if the longitudinal model is correctly specified, there will be no difference between the original dataset and the M simulated datasets, so all the values in the empirical cumulative distribution will have the same probability. In such a case, this would lead to a uniform distribution of the residuals, regardless of the longitudinal model employed to fit the data. Once the described process has been repeated for each of the original observations, a residual analysis can easily be carried out, detecting deviations from the uniform distribution, residual dependency on a predictor, or overdispersion.

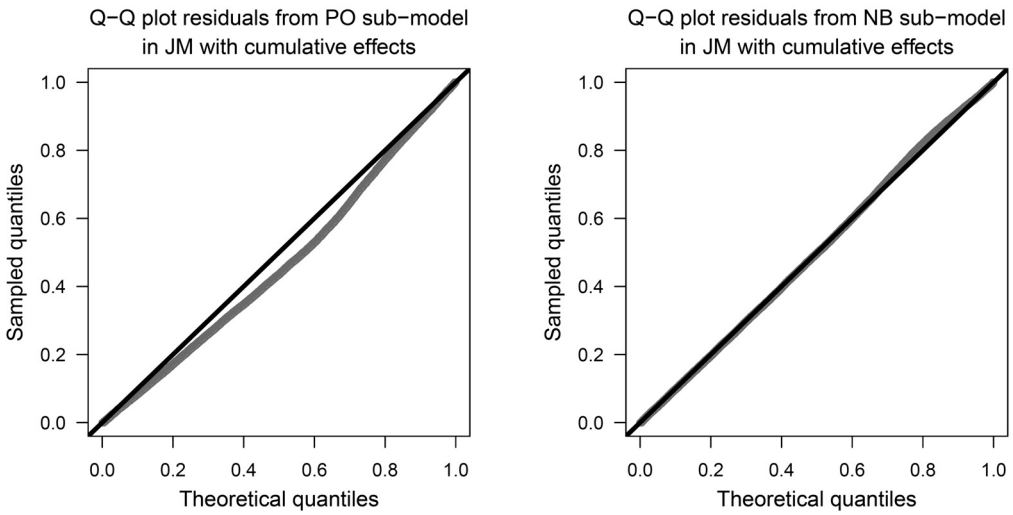

Figure 4: *Randomized quantile residuals for the longitudinal sub-model of the joint models with a recency-weighted cumulative effect, for both PO longitudinal sub-model (left panel) and NB longitudinal sub-model (right panel).*

Figure 4 depicts the randomized quantile residuals derived from both PO and NB longitudinal sub-models in the fitted JM with a recency-weighted cumulative effect parameterization, where $M = 500$ datasets were simulated for each sub-model. The quantile residuals of the estimated PO longitudinal sub-model show clear evidence of lack of fit, exhibiting a quadrature pattern, probably due to the overdispersion effect. In order to confirm this trend, we also plotted the residuals against the time predictor, and again a systematic deviation quadrature was obtained in the corresponding Q-Q plot. By contrast, the quantile residuals of the NB sub-model are almost perfectly uniformly distributed, lying approximately on the diagonal line. These graphical results strongly suggest the adequacy of the NB longitudinal sub-model.

To check the quality of the survival model's predictions, the analysis of martingale residuals (Barlow and Prentice, 1988) is a very common graphical method. Let $R_i(t) = \mathbb{I}(T_i \geq t)$ be the indicator that the i-th subject is at risk at time t, and $N_i(t)$ be the corresponding cumulative number of events until time t. In general, the martingale residuals for subject i at time t is defined by the mean estimates $(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{b}}_i)$ as

$$r_i^m(t) = N_i(t) - \int_0^t R_i(s)\, h_i\{s \,|\, \overline{M}_i(s), \overline{\boldsymbol{\theta}}\}\, \mathrm{d}s, \quad r_i^m(t) \in (-\infty, 1]. \tag{13}$$

Here, $h_i\{s \,|\, \overline{M}_i(s), \overline{\boldsymbol{\theta}}\} = \overline{h}_0(s) \exp\left[\overline{\gamma}_g w_{gi} + \overline{\alpha} F\{\overline{\mu}_i(s)\}\right]$, where $\overline{h}_0(\cdot)$ is the estimated baseline hazard function of the relative risk model, and $\overline{\mu}_i(s) = e_i(s) \exp\{\overline{\beta}_0 + \overline{b}_{i0} + \overline{\beta}_1 s + \overline{b}_{i1} s + \overline{\beta}_{t_{90}} \mathbb{I}(s \geq 25)\}$. The martingale residuals are calculated, for each subject-specific measurement, as the difference between the observed number of events

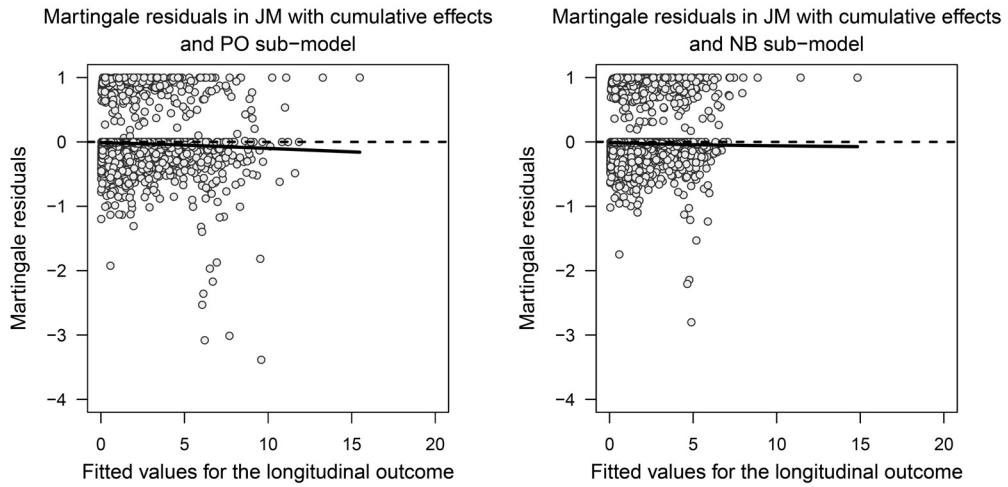

Figure 5: *Martingale residuals derived from the joint models with a recency-weighted cumulative effect. The left panel shows the martingale residuals versus the subject-specific values fitted with the PO longitudinal sub-model, and the right panel shows the results obtained with the NB longitudinal sub-model. In both plots, a loess function has been overlaid to determine the trend.*

(0 or 1) and the number of events expected to happen. This implicitly allows for the identification as outliers of those cases where the survival function predicts an event either too early (values near 1) or too late (extreme negative values). Using an adequate model specification, these residuals should be uncorrelated with one another and have a zero mean, even though they are not symmetrically distributed around zero value.

Taking the results of the JM approach with a recency-weighted cumulative effect, we have calculated the martingale-based residuals for both PO and NB longitudinal sub-models. Figure 5 shows the corresponding plots of the martingale residuals versus the fitted values for the emergency claims per year. In general, these residuals are skewed towards negative values, and for this reason it is very useful to superimpose a loess function (solid line) to better assess the shape of the relationship between the residuals and the fitted values. In the JM with the PO longitudinal sub-model, the loess trend deviates from zero as the fitted value increases. By contrast, in the plot for the JM with NB longitudinal response, the loess curve shows almost no evidence of a trend across all the fitted values.

5. Individualized survival predictions

Using the Bayesian joint framework, personalized and dynamically-updated survival predictions can be obtained by considering each subject-specific longitudinal profile (Proust-Lima and Taylor, 2009; Rizopoulos, 2012; Serrat et al., 2015). Let us consider a new subject, denoted by $k = i+1$, not included in the original dataset but sampled from the target population. If emergency claims per year are recorded until time t, we implicitly know that this new subject is alive at least until t, thus providing a historical set of observed measurements, $Y_k(s) = \{y_k(s_{kj}),\ \tau_k \leq s_{kj} \leq t,\ j = 1,\ldots,n_k\}$, as well as a specific value for gender factor w_{gk}. From this information, we want to predict the conditional subject-specific survival probabilities at any future time $u > t$, given survival up to t: $\tilde{p}_k(u\,|\,s) = p_t\left(T_k^* \geq u\,|\,T_k^* > s, Y_k(s), w_{gk}, D_n\right)$. This prognosis task can be carried out quite straightforwardly by adopting a Bayesian strategy. Let $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{b}_k)$ denote the full vector of uncertainties in the joint model and the random effects of the new subject. Then, the the conditional survival probability can be estimated from the posterior predictive distribution of the observed data:

$$
\begin{aligned}
&\tilde{p}_k(u\,|\,s) = \\
&= \iint_{\boldsymbol{\Omega}} p_t\left(T_k^* \geq u\,|\,T_k^* > s, Y_k(s), w_{gk}, \boldsymbol{b}_k, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\,|\,D_n\right) \mathrm{d}\boldsymbol{b}_k\, \mathrm{d}\boldsymbol{\theta} \qquad (14)\\
&= \iint_{\boldsymbol{\Omega}} p_t\left(T_k^* \geq u\,|\,T_k^* > s, \boldsymbol{b}_k, \boldsymbol{\theta}\right) p_b\left(\boldsymbol{b}_k\,|\,T_k^* > s, Y_k(s), w_{gk}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\,|\,D_n\right) \mathrm{d}\boldsymbol{b}_k\, \mathrm{d}\boldsymbol{\theta}\\
&= \iint_{\boldsymbol{\Omega}} \frac{\Pr\left(T_k^* \geq u\,|\,M_k(u), \boldsymbol{b}_k, \boldsymbol{\theta}\right)}{\Pr\left(T_k^* > s\,|\,M_k(s), \boldsymbol{b}_k, \boldsymbol{\theta}\right)} p_b\left(\boldsymbol{b}_k\,|\,T_k^* > s, Y_k(s), w_{gk}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\,|\,D_n\right) \mathrm{d}\boldsymbol{b}_k\, \mathrm{d}\boldsymbol{\theta}.
\end{aligned}
$$

Using the previous result and the MCMC sampling information from $\pi(\boldsymbol{\theta}, \boldsymbol{b}_i \,|\, D_n)$ for a particular JM fit (we assume that the inclusion of a new subject does not entail the updating of $\bar{\boldsymbol{\theta}}$), a simulation scheme can be applied to obtain a Monte Carlo estimate of $\tilde{p}_k(u \,|\, s)$. By way of an example, let us consider a male and female policyholders, both aged 70 years upon entering the study ($\tau_k = 5$), and not included in the original dataset. The same number of emergency claims per year is assumed for them during the next decade, those being observed annually between the ages of 70 and 80, providing $\{Y_k(s), s = 5, \ldots, 15\}$. Moreover, we assume a NB counting sequence within the JM approach with a recency-weighted cumulative effect. We first focus on estimating the survival probability of both subjects at 90 years of age, conditioned on their being alive at s, $\tilde{p}_k(u = 25 \,|\, s)$. The results are obtained by adapting the code of `survfitJM(·)` function from the `JMbayes` package, and they show how the Monte Carlo estimates update dynamically as new longitudinal information is considered (Table 5). Time-dynamic updating of this kind emphasizes the need for a well-characterized follow-up for each policyholder when we aim for personalized decisions and an accurate prediction of the insurance capital needed to cover the corresponding health insurance plan.

Table 5: *Time-dynamic probabilities of being alive at 90 years for a man and a woman with the same longitudinal information collected between the ages of 70 and 80. The results are estimated from the JM with a recency-weighted cumulative effect parameterization and a NB longitudinal sub-model.*

Age (yr.)	Emergency claims per year, $y_k(s)$	Man's survival at 90 yr.			Woman's survival at 90 yr.		
		Mean	$q_{2.5\%}$	$q_{97.5\%}$	Mean	$q_{2.5\%}$	$q_{97.5\%}$
70	0	0.783	0.478	0.864	0.796	0.509	0.871
71	0	0.804	0.601	0.868	0.817	0.628	0.874
72	1	0.790	0.550	0.863	0.803	0.563	0.872
73	0	0.806	0.635	0.868	0.818	0.647	0.874
74	2	0.776	0.556	0.854	0.790	0.585	0.863
75	0	0.793	0.616	0.861	0.807	0.632	0.871
76	0	0.807	0.660	0.867	0.822	0.682	0.875
77	8	0.671	0.359	0.812	0.692	0.367	0.826
78	1	0.690	0.404	0.816	0.708	0.410	0.829
79	2	0.688	0.406	0.812	0.706	0.414	0.827
80	0	0.719	0.479	0.825	0.736	0.501	0.839

We conclude that there is an increasing probability of being alive at the age of 90 when no emergency claims are observed, whereas this probability decreases sharply when a large number of emergency claims are annually reported. The survival probabilities for the female are higher than those for the male policyholder, since the gender regression coefficient indicates that *ceteris paribus* males have a higher mortality risk than females. Hence, by the age of 80, the survival estimate at 90 years of age for the male policyholder is $\tilde{p}_{k,m}(u = 25 \,|\, s = 15) = 0.719$, whereas a woman with the same demand presents an estimate of $\tilde{p}_{k,w}(u = 25 \,|\, s = 15) = 0.736$.

Table 6: *Time-dynamic probabilities of being alive above 80 years for a man and a woman with the same longitudinal information collected between the ages of 70 and 80. The results are estimated from the JM with a recency-weighted cumulative effect parameterization and a NB longitudinal sub-model.*

Age (years)	Man's survival above 80 yr.			Woman's survival above 80 yr.		
	Mean	$q_{2.5\%}$	$q_{97.5\%}$	Mean	$q_{2.5\%}$	$q_{97.5\%}$
80	1.000	1.000	1.000	1.000	1.000	1.000
82	0.972	0.945	0.983	0.974	0.945	0.985
84	0.934	0.871	0.960	0.939	0.875	0.964
86	0.882	0.773	0.929	0.890	0.779	0.935
88	0.812	0.640	0.884	0.824	0.655	0.896
90	0.719	0.479	0.825	0.736	0.501	0.839

If we know for certain that both subjects from the previous example remain alive when they are 80 years old, then we can also assess their future survival from the information contained in our dataset of policyholders above the age of 80. Table 6 provides the survival estimates of these two subjects. Recall that logically the last row in this table provides the same results as those in Table 5, since both survival estimates at the age of 90 are performed under the same assumptions.

6. Conclusions

Health insurance companies have access to a valuable source of information for use in follow-up studies, as they keep records of the medical claims made by each of their policyholders, in what is a discrete counting process. In this article, we have assessed the degree of relationship between an elderly policyholder's annual demand of medical emergency claims (as a longitudinal discrete response) and his or her time until death. We defined elderly people as those with a chronological age of 65 years or above, so the event times are left-truncated for all subjects whose observation time starts after this age threshold.

A correct statistical analysis of the association between the longitudinal and time-to-event outcomes entails a joint modelling approach. The longitudinal analysis was carried out using either a PO or a NB mixed model, whereas for the survival analysis, a time-dependent Cox model was used. The JM for annual rate counts and delayed entries was fitted under the Bayesian paradigm via JAGS software, entailing the challenging task of applying it to a large dataset. First, we examined the influence of the current longitudinal outcome on mortality risk. Then, in a second stage, we considered the effect of the recency-weighted area under the longitudinal profile on survival. In both cases, the results show that relatively high cumulative demand for ambulance services, hospitalizations, and non-routine visits is positively related to a deterioration in the subject's health status and, consequently, it entails higher mortality risk (i.e., lower survival probabilities). The most interesting conclusion is that the most recent critical medical

demand has the greatest impact on the current survival. This is what the JM is able to capture by means of the functional form which relates the recency-weighted area under the expected longitudinal profile with the time-to-event outcome, this approach being preferable to the one which only takes into account the effect of the current expected longitudinal value. Moreover, the results confirm the adequacy of assuming a NB distribution in the longitudinal sub-model as a first step to account for overdispersion in the longitudinal response. However, further extensions in the longitudinal part can be considered to specifically deal with zero inflation, such as different versions of zero-inflated and Hurdle models. To conclude, subject-specific survival predictions have been obtained as an example of the enormous potential of joint analysis as a predictive tool.

Acknowledgments

The authors thank the two anonymous reviewers of this article, whose valuable suggestions greatly contributed to the improvement of the manuscript.

References

Abrahamowicz, M., Beauchamp, M.-E. and Sylvestre, M.-P. (2011). Comparison of alternative models for linking drug exposure with adverse effects. *Statistics in Medicine*, 31, 1014–1030.

Barlow, W. and Prentice, R. (1988). Residuals for relative risk regression. *Biometrika*, 75, 65–74.

Boucher, J.-P., Denuit, M. and Guillén, M. (2008). Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Variance*, 2, 135–162.

Cameron, C. and Trivedi, K. (1998). *Regression Analysis of Count Data*. Number 30. Cambridge University Press. Cambridge, England.

Charpentier, A. (2015). *Computational Actuarial Science with R*. Boca Raton, Florida: Chapman and Hall/CRC The R Series.

Chen, B., Covinsky, K., Stijacic, C., Adler, N. and Williams, B. (2012). Subjective social status and functional decline in older adults. *Journal of General Internal Medicine*, 27, 693–699.

Dow, W., Schoeni, R., Adler, N. and Stewart, J. (2010). Evaluating the evidence base: policies and interventions to address socioeconomic status gradients in health. *Annals of the New York Academy of Sciences*, 1186, 240–251.

Dunn, K. and Smyth, G. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5, 236–244.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–121.

Eilers P., Marx, B. and Durbán, M. (2015). Twenty years of P-splines. *SORT*, 39, 149–186.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica*, 52, 681–700.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters*, 99, 585–590.

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

Hartig, F. (2017). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models*. R package version 0.1.5.

Hilbe, J. (2011). *Negative Binomial Regression*, 2nd Edition. Cambridge: Cambridge University Press.

Hinde, J. and Demétrio, C. (1998). Overdispersion: models and estimation. *Computational Statistics and Data Analysis*, 27, 151–170.

Ismail, N. and Jemain, A. (2007). Handling overdispersion with negative binomial and generalized Poisson regression models. In *Casualty Actuarial Society Forum*, Winter 2007, pp. 103–158.

Ivanova, A., Molenberghs, G. and Verbeke, G. (2016). Mixed models approaches for joint modelling of different types of responses. *Journal of Biopharmaceutical Statistics*, 26, 601–618.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.

Lamarca, R., Alonso, J., Gómez, G. and Muñoz, A. (1998). Left-truncated data with age as time scale: an alternative for survival analysis in the elderly population. *Journal of Gerontology: Medical Sciences*, 53, M337–M343.

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, 15, 209–225.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

Murawska, M., Rizopoulos, D. and Lessaffre, E. (2012). A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies. *Journal of Probability and Statistics*, vol. in press, 2012, 1–18.

Piulachs, X., Alemany, R., Guillén, M. and Serrat, C. (2015). Joint modelling of health care usage and longevity uncertainty for an insurance portfolio. In *Scientific Methods for the Treatment of Uncertainty in Social Sciences*, pp. 289–297. Springer International Publishing.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Technische Universität Wien, Vienna, Austria.

Proust-Lima, C. and Taylor, J. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modelling approach. *Biostatistics*, 10, 535–549.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data with Applications in R*. Chapman and Hall/CRC Biostatistic Series. Boca Raton, Florida.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72, 1–46.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30, 1366–1380.

Serrat, C., Rue, M., Armero, C., Piulachs, X., Perpiñán, H., Forte, A., Páez, A. and Gómez, G. (2015). Frequentist and Bayesian approaches for a joint model for prostate cancer risk and longitudinal prostate-specific antigen data. *Journal of Applied Statistics*, 42, 1223–1239.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64, 583–639.

Thiébaut, A. and Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine*, 23, 3803–3820.

Tsiatis, A. and Davidian, M. (2004). Joint Modelling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14, 809–834.

Uzunogullari, U. and Wang, J. (1992). A comparison of hazard rate estimators for left-truncated and right-censored data. *Biometrika*, 79, 297–310.

Viviani, S., Alfó, M. and Rizopoulos, D. (2012). Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing*, 24, 417–427.

Yu, M., Taylor, J. and Sandler, H. (2008). Individual prediction in prostate cancer studies using a joint longitudinal survival-cure model. *Journal of the American Statistical Association*, 103, 178–187.

Zuur, A., Ieno, E., Walker, N., Saveliev, A. and Smith, G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

SORT 41 (2) July-December 2017, 373-392 DOI: 10.2436/20.8080.02.64

Statistical and machine learning approaches for the minimization of trigger errors in parametric earthquake catastrophe bonds

Laura Calvet[1,2], Madeleine Lopeman[3], Jésica de Armas[4], Guillermo Franco[3] and Angel A. Juan[1,2]

Abstract

Catastrophe bonds are financial instruments designed to transfer risk of monetary losses arising from earthquakes, hurricanes, or floods to the capital markets. The insurance and reinsurance industry, governments, and private entities employ them frequently to obtain coverage. Parametric catastrophe bonds base their payments on physical features. For instance, given parameters such as magnitude of the earthquake and the location of its epicenter, the bond may pay a fixed amount or not pay at all. This paper reviews statistical and machine learning techniques for designing trigger mechanisms and includes a computational experiment. Several lines of future research are discussed.

MSC: 62-07.

Keywords: Catastrophe bonds, risk of natural hazards, classification techniques, earthquakes, insurance.

1. Introduction

Catastrophe (CAT) bonds are financial instruments that package catastrophe risk in a tradeable security. These tools are in effect responsible for the existence of a new market for trading risk at the frontier between finance and insurance, the so-called convergence market (Cummins and Weiss, 2009), which promises an enormous supply of capital for CAT risk transfer as long as pricing remains attractive for all parties involved. By purchasing a CAT bond, investors take the risk from a sponsor (risk ceding party) in exchange for some interest or spread. This spread constitutes the premium that compensates the risk-taking party.

[1] IN3-Department of Computer Science, Open University of Catalonia, 08018 Barcelona, Spain.

[2] Euncet Business School, 08225 Terrassa, Spain. {lcalvetl, ajuanp}@uoc.edu

[3] Guy Carpenter & Company, LLC, Dublin, Ireland, {madeleine.lopeman, guillermo.e.franco}@guycarp.com

[4] Faculty of Economics and Business Sciences, Pompeu Fabra University, 08005 Barcelona, Spain. jesica.dearmas@upf.edu

Received: December 2016
Accepted: August 2017

CAT bonds can be of different types depending on how their payment behaviour is structured. Earthquake CAT bonds in particular can base their payments on a variety of proxies (Wald and Franco, 2017). While some base payments on actual, experienced losses (indemnity), others (parametric) base them on the observable and measurable parameters that describe an event. Strategies to provide coverage for large losses ensuing after earthquakes through these parametric tools have been in use since the 1990s (Franco, 2014). Their popularity in the market is due to historically lower prices relative to traditional (re)insurance and their appeal among investment and hedge funds is due to their transparency. Lately, as traditional reinsurance pricing has decreased significantly, the price differential between traditional and alternative risk transfer (sometimes referred to as ART) is very small and is no longer the driving rationale for seeking parametric coverage. Rather, sponsors now look to parametric risk transfer for the flexibility and the ease of payment it provides.

Parametric earthquake CAT bonds employ a kind of trigger mechanism, typically a numerical check of some sort, to determine the payment that should take place when an earthquake occurs. These trigger mechanisms rely on obtainable physical characteristics of the event via respected third parties, often public agencies (Cummins, 2007; Croson and Kunreuther, 1999).

Since neither the investor nor the sponsor has the ability to manipulate this information, the risk transfer process is without moral hazard (the risk that the parties involved influence the payment outcome). Earthquakes around the world cause enormous losses, of which only about 30% have insurance coverage (Guy Carpenter, 2014). These financial impacts often disrupt individual livelihoods and national economies. Therefore, the possibility of expanding the coverage of insurance to minimize these impacts is very appealing. Making earthquake insurance more accessible, however, is difficult for traditional providers since their operations are typically resource- and time-consuming. Parametric risk transfer, in contrast, can be seamless, fast and cheap but in order to be viable, parametric solutions need to be accurate. They also need to be designed and customized without much effort so they can be easily industrialized.

Despite the popularity of parametric CAT bonds in the reinsurance market, the number of scientific works discussing these financial tools is rather limited. Thus, one of the main contributions of this paper is to help to fill this gap by proposing simple, fast, and automatic approaches able to design accurate trigger mechanisms. While a few authors have proposed ad hoc complex approaches based on the use of genetic algorithms, we explore here the use of more general tools based on statistical and machine learning methods. In particular, we review eight techniques to classify events as to whether they should trigger a payment or not, following a binary payment scheme often used in the industry. Events are classified using the fundamental parameters of focal location (i.e., longitude, latitude and depth) and the moment magnitude. Note that all the approaches need to be trained with a given dataset. These data need to constitute a large sample of events and have to include a monetary loss for each earthquake. Therefore, we turn to an earthquake CAT model to obtain a viable training dataset since historical catalogues

usually do not contain a large enough sample of this type of information. In order to test the performance of these techniques against a known benchmark, we recover the analysis presented in Franco (2010) and solve the same problem using the same dataset. We then compare across methods on such issues as accuracy, computational effort, and spatial correlation of the classifier results. According to the obtained results, the techniques employed can produce trigger mechanisms of equal or better accuracy than the technique described in the aforementioned paper. Moreover, several techniques provide huge efficiency gains in terms of decreasing classification time. Additionally, they provide scalability, being easily adapted to a larger parameter space and larger catalogues without losing much efficiency. Finally, they are relatively easy to implement using modern programming languages and software such as Python and R.

The rest of this work is structured as follows. While Section 2 contextualizes this work and describes the characteristics of the trigger mechanism, Section 3 reviews the related work. Afterwards, the techniques considered are introduced in Section 4. Section 5 explains the computational experiments. Finally, our conclusions and suggestions for future research are collected in Section 6.

2. CAT bonds and the trigger mechanism

CAT bonds have allowed insurers, reinsurers, governments, private entities and catastrophe pools to cede risks of earthquake losses to the capital markets via transparent mechanisms associated with physical event features. Since they bypass the claims adjusting process, these tools provide a very fast recovery of funds to their sponsor after an event. Within the realm of parametric earthquake CAT bonds there are also several classes of tools. Some first-generation parametric CAT bonds, or so-called "CAT-in-a-box" triggers, rely on the main physical descriptors of an earthquake event (see, for instance, Cardenas et al., 2007; Franco, 2010; Franco, 2013). Others, second-generation indexes, rely on spatially-distributed features such as ground motions recorded at sensors located throughout a region (see for instance Goda, 2013; Goda, 2014; Pucciano, Franco and Bazzurro, under review). From here, this paper focuses on first-generation triggers.

Consider a set of l earthquake events in a geographic region of interest A. An earthquake event i is characterized by a magnitude m_i, a hypocenter depth d_i, and epicenter coordinates (x_i,y_i) within A. A binary trigger will determine whether a payment should be disbursed due to event i. This response is represented by the variable B', whose values $1/0$ indicate trigger/no-trigger (payment/no-payment). Two situations may arise: (1) at least one earthquake i triggers the bond ($B'_i = 1$) during its contract life, which means that the entire bond principal has to be disbursed and, as a consequence, the buyers of the bond lose their investment (and the bond sponsors receive compensation), or (2) no earthquake triggers the bond during its life, in which case the principal is returned to the investors with interest.

Since the payment of a large sum of money is at stake, it is important that the trigger performs as desired, i.e. that the trigger responds positively to events that cause a large loss beyond a design threshold and that it does not respond to events that cause a loss below this threshold. The accuracy of the trigger determines its success in the market. Triggers that behave erratically erode the confidence of the markets in these tools and therefore jeopardize the risk transfer process. It is crucial to design triggers that behave as they should.

To describe the accuracy of the trigger, first consider a reference variable B that represents its idealized behaviour and that depends on a measure based on the losses (typically monetary). For an earthquake event i, this variable can be described as follows:

$$B_i = \begin{cases} 0 & \text{if } L_i \leq L \\ 1 & \text{otherwise} \end{cases}$$

where L_i is the actual loss caused and L is a loss threshold specified by the sponsor, usually expressed in terms of a specific return period. In this idealized scenario, events trigger this CAT bond only if the corresponding loss is above a given pre-specified threshold L.

The objective of parametric trigger mechanism design is to develop a classification mechanism that uses physical parameters of events to determine the trigger behaviour B'. Discrepancies between variables B and B' or the sum of errors $(E = \sum_{i=1}^{I} I(B_i \neq B'_i))$, represent lack of correlation between the output of the trigger and the ideal trigger. Effective parametric trigger mechanism design aims to minimize these discrepancies.

A database including a set of events, their characteristics and the variable B can be used to calculate trigger errors for this specific set of events. A measure of the loss has to be obtained or estimated to compute B. It is preferable to have a reliable historical dataset including a high number of events but in earthquake research, this is not possible due to the low frequency of earthquakes, and the great uncertainty surrounding their associated losses, and the evolution of insured portfolios over time. For this reason, the design of triggers for seismic risk relies on simulated CAT model output.

According to the description offered in this section, the development of a trigger mechanism can be labelled as a binary classification problem, allowing us to employ a wide range of techniques to address it. In the following sections, some of them are introduced and tested, and their use is illustrated.

3. Related work

The literature related to CAT bonds has increased during the last few years due to their growing popularity. Combining instruments in finance and insurance fields with engineering seismic risk assessment, Tao and Tao (2005) propose a method to set the rate

for earthquake property and personal insurances with two kinds of deductibles. Moreover, the authors present a framework to set the annual coupon rate for earthquake CAT bonds, which considers the probability of a catastrophe occurrence from seismic risk assessment, the yields of reinvestment, the principal protected ratio and the issuance fee ratio. An illustrative example focused on an urban area of China is described. Zimbidis, Frangos and Pantelous (2007) produce a model for the risk dynamics of the magnitude of the earthquakes by using advanced techniques from the extreme value theory. The model is tested on historical data of earthquakes in Greece. Moreover, the theory of incomplete markets and price CAT bonds is discussed. Tao, Tao and Li (2009) builds a pricing model, which employs the probability of an earthquake, estimated by a seismic risk assessment method. The cash flows of the insurance in complete and incomplete markets are described by Geometric Brownian Motion and Jump-Diffusion processes, respectively. Wu and Zhou (2010) reviews the state-of-the-art approaches in modelling losses for CAT bonds' modelling and pricing. They are compared by using a catalogue of earthquakes in China from 1966 to 2008. The double exponential Jump-Diffusion model fits better. Damnjanovic, Aslan, and Mander (2010) propose an integrative model linking engineering design parameters with financial indicators. The authors explain a framework based on a four-step structural loss model and a transformed survival model, which estimates excess returns. Härdle and Cabrera (2010) study the calibration of a CAT bond for Mexican earthquakes, which proves that a hybrid strategy combining traditional reinsurance and CAT bonds presents a better performance in the sense that provides coverage for a lower cost and lower exposure in comparison with a strategy without CAT bonds. Goda (2013) compares the effectiveness of two trigger mechanisms for parametric earthquake CAT bonds: scenario-based and station intensity-based approaches. The results indicate that the latter method performs at least as well as the former. Additionally, different spatial correlation models of peak ground motions are studied. Later, Goda (2014) extends the station intensity-based trigger method, which uses direct observation of ground motions at recording stations, by promoting a flexible multiple-discrete payment structure. Gunardi and Setiawan (2015) present a study case for Indonesia, in which formulas are proposed for pricing three types of CAT bonds. A generalized extreme value distribution is used to model the probability of maximum magnitude for Indonesian earthquakes. Shao, Pantelous and Papaioannou (2015) investigate the pricing process for CAT bonds considering financial and catastrophe-independent risks. An application for earthquakes is considered employing extreme value theory, and a numerical example based on California is detailed. Finally, Cummins (2007) reviews the status of the market for CAT bonds and other risk-linked securities. It discusses the complementarity between CAT bonds and the reinsurance market. In addition, the role of other modern financing mechanisms such as risk swaps, industry loss warranties, and sidecars is explained.

4. Statistical and machine learning approaches

Classification techniques (Kotsiantis, 2007) constitute a set of procedures from statistics and machine learning (more specifically, supervised learning) to determine a category or class for a given observation. Having a dataset of l observations composed of explanatory or independent variables $(X_1, X_2, \ldots, X_n)$, and a response or dependent variable Y, these techniques attempt to explain the relationships between variables and/or classify new observations based on the explanatory variables. In the problem of CAT bond trigger design, the response of the mechanism is the dependent variable Y, while the characteristics of an earthquake event (i.e., the magnitude, the hypocenter depth, and the epicenter coordinates) represent the independent variables.

Nowadays, there are plenty of classification techniques. Some of the most employed, e.g., Linear Discriminant Analysis or Logistic Regression, have been applied for more than five decades. These are mainly linear methods. Boosted by the computing advances in the 1980s and 1990s, non-linear methods such as Classification Trees, Neural Networks and Support Vector Machines emerged and started to attract attention. This section introduces some well-known and powerful techniques that we propose to automatically design a trigger. The reader interested in comprehensive and practical descriptions is referred to the books written by Hastie, Tibshirani and Friedman (2009) and Lantz (2015).

4.1. The nearest neighbours classifier

The Nearest Neighbours classifier is a simple technique that assigns a new observation to the class of the most similar observations, so-called neighbours. Therefore, it is suitable when observations of the same class tend to be homogeneous. Its main weaknesses are: not producing a model (which hinders the exploration of relationships among variables), taking a relatively high amount of time, and consuming a large amount of memory. This classifier depends on a parameter k representing the number of neighbours. The neighbours are selected according to a distance function, usually Euclidean. This parameter allows the balance between overfitting and underfitting (also known as bias-variance trade-off): a large k reduces the variance caused by noisy data or outliers but may ignore small/local patterns; conversely, a small value may introduce too much bias.

4.2. The naïve Bayes classifier

The naïve Bayes classifier is based on Bayes' theorem. "Naïve" refers to the assumption that all variables are independent and equally important. Even if this condition is not usually met in real-life applications, this classifier frequently provides competitive results. The posterior probability for a given class y is computed as:

$$P(Y = y \mid X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n) =$$

$$\frac{P(X_1 = x_1 \mid Y = y)P(X_2 = x_2 \mid Y = y) \cdots P(X_n = x_n \mid Y = y)P(Y = y)}{P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n)}$$

The classification for a given observation is obtained by comparing the probabilities of each class given the values of the explanatory variables, and selecting the class associated to the highest probability. There are many classifiers differing in the assumption made regarding the distribution of $P(X_j = x_j \mid Y = y)$. Gaussian distributions constitute a typical choice. This technique employs frequency tables and, consequently, each variable must be categorical. Numeric variables are usually discretized.

4.3. Linear and quadratic discriminant analyses

In Linear Discriminant Analysis, the distribution of the explanatory variables is separately modelled in each of the classes, and then Bayes' theorem is used to flip these around into estimates for the probability of the response variable taking a specific value given the explanatory variables. Commonly, these distributions are assumed to be Gaussian. In this case, the resulting models are similar to those provided by Logistic Regression. Linear Discriminant Analysis is more commonly employed when there are more than two classes. While this technique assumes that observations are drawn from a distribution with a common covariance matrix in each class (which leads to linear decision boundaries), Quadratic Discriminant Analysis does not make assumptions on the covariance matrices (producing quadratic decision boundaries).

4.4. Classification trees

Contrary to global models (where a predictive formula is supposed to hold in the entire data space) such as those of Logistic Regression, Classification Trees try to partition the data space into small enough parts where a simple model can be applied. The results can be represented as a tree composed of internal and terminal (or leaf) nodes, and branches. Its non-leaf part is a procedure to determine for each observation which model (i.e., terminal node) will be used to classify it. At each internal node of the tree, the value of one explanatory variable is checked and, depending on the binary answer, the procedure continues to the left or to the right sub-branch. A classification is made when a leaf is reached.

The most relevant advantage of this classifier is the easiness to understand what trees represent. They mirror human decision-making more closely than other techniques. Furthermore, trees require little data preparation, are able to handle both numerical and categorical data, and perform well (i.e., use standard computing resources in reasonable time) with large datasets.

4.5. Logistic regression

Logistic Regression techniques are designed to model the posterior probabilities of each class by means of linear functions. These probabilities, such as the one shown below, must be non-negative and sum to one.

$$P(Y = y \mid X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}$$

These models are usually fitted by maximum likelihood employing Newton's method. The previous expression can be rewritten in terms of log-odds as follows:

$$\log\left(\frac{P(Y = y \mid X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n)}{1 - P(Y = y \mid X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

This technique is especially useful when the aim is to explain (i.e., not only classify) the outcome based on the explanatory variables. Non-linear functions can be considered including interactions and transformations of the original variables.

4.6. Clusterwise logistic regression

While Regression Analysis consists of fitting functions to analyse the relationship between variables, Clustering seeks subsets of similar observations (or variables) in a dataset. Thus, the aim of Clusterwise Regression is to combine both techniques in order to discover trends within data when more than one trend is likely to exist (DeSarbo and Cron, 1988). This technique is highly flexible because different functions can be estimated. It is considered a "white-box technique" in that its mathematical systems are not complex and its results are relatively easy-to-interpret.

4.7. Neural networks

Neural Networks model the relationship between the explanatory variables and the response variable using a model inspired by how a biological brain responds to stimuli from sensory inputs. They extract linear combinations of the explanatory variables as derived variables and model the response variable as a non-linear function of these transformed variables. These models have several kinds of layers: the input layer, the output layer, and one or more hidden layers between them. Each layer contains neurons representing the variables. Increasing the number of hidden layers and/or neurons adds complexity and may improve computational capacity. With too few layers, the model may lack the flexibility to capture non-linearities in data. Neural Networks tend to have many weights, which can cause problems of overfitting. Weight decay is a method of regularization to prevent it. The "backpropagation" algorithm is a technique commonly employed for parameter estimation or training a Neural Network.

4.8. Support vector machines

A Support Vector Machine can be imagined as a surface that defines a boundary between various points of data that represent observations plotted in a multidimensional space. The goal is to create a flat boundary, called a hyperplane, which leads to fairly homogeneous partitions of data on either side. Among all potential hyperplanes, the one that creates the greatest separation between classes (a soft margin may be considered for the case on non-linearly separable data) is selected. The support vectors are the points from each class that are the closest to the hyperplane; each class must have at least one. In many real-life applications, the relationships between variables are non-linear. A key feature of this technique is its ability to efficiently map the observations into a higher dimension space by using the kernel trick. As a result, a non-linear relationship may be transformed into a linear one.

4.9. Discussion of classification techniques

Several techniques have been presented in the literature to design trigger mechanisms that determine – from an earthquake's physical characteristics – whether a principal bond should be paid (Franco, 2010, 2013). As mentioned, the aim of this work is to introduce and illustrate the application of simple, well-known, and efficient techniques that have heretofore not been explored in this context. Neural Networks and Support Vector Machines constitute two relatively modern and powerful techniques. Typically, they are able to reach high levels of accuracy by capturing nonlinear relationships between variables. However, this same characteristic makes them prone to overfitting. There are many procedures to avoid this problem such as the addition of a parameter to limit the growth of the weights or the introduction of randomness into the training data or the training algorithm. Sometimes it may be difficult to avoid overfitting and underfitting. Training Neural Networks often takes a long time, and both techniques require a non-trivial process of fine-tuning parameters. Furthermore, the resulting models are difficult if not impossible to interpret. For this reason, application of these techniques is almost always limited to classification/prediction purposes. Techniques such as Nearest Neighbours and Naïve Bayes Classifiers are easier to understand and implement and may provide relatively high accuracy. While the first is non-parametric and, consequently, flexible or unstable, the second relies on some assumptions that may be quite unrealistic in most cases.

Logistic Regression is a well-established technique, which enables the understanding of the effects of the explanatory variables on the response. Clusterwise Logistic Regression aims to incorporate the strengths of Logistic Regression while offering more flexibility, which should lead to a better understanding of the relationships among variables and higher accuracy. Classification Trees constitute an efficient technique that only uses the most important variables and results in a logic model. As other techniques studying non-linear relationships, these three techniques are particularly susceptible to overfitting or underfitting the model. Typically, small changes in training data may lead to signif-

icant modifications. In addition, Classification Trees may derive decisions that seem counterintuitive or are unexpected.

Closely related to Logistic Regression, the classic Linear/Quadratic Discriminant Analysis techniques search for the linear/quadratic combination of variables that explains the data the best. Logistic Regression is preferred if the assumption of normally-distributed explanatory variables does not hold. Otherwise, Discriminant Analysis can provide better results.

All these techniques have different features worthy of consideration when addressing a classification problem. Consequently, all are included in the following computational experiments.

5. Computational experiments

This section illustrates the application of the techniques introduced in Section 4 and compares the results with those obtained in Franco (2010). A framework for evaluation is presented such that the techniques can be compared to one another and to the reference methodology along the dimensions of accuracy, efficiency, and spatial correlation.

The dataset analysed is an earthquake catalogue representing 10,000 years of simulated seismic activity in and around Costa Rica. The catalogue contains a total of 24,957 earthquakes. These records should include the four main physical parameters enumerated before and the corresponding simulated loss. For each synthetic earthquake event in the catalogue, the model computes a ground motion footprint, which is in turn translated into estimated levels of damage to a user-defined portfolio of properties distributed in space. CAT Models have been discussed in previous studies (e.g. Grossi and Kunreuther, 2005) and we will not discuss the CAT modelling process here. The target of the classifier algorithms, in short, is to discriminate events based on their physical parameters to identify large loss-producing events. A more detailed description of the catalogue can be found in the aforementioned work. In this case study, the events in the database are assumed to be triggering events if their loss is equalled or exceeded with an annual probability of 1%.

5.1. Evaluation framework

In the case of parametric trigger design, it is difficult a priori to select the "best" classification technique for two main reasons. First, it is a multi-objective problem. Although from a statistical perspective, the sole objective may be to maximize accuracy, in real-life applications many other characteristics will likely play an important role. These may include ease of implementation, ease of explanation to non-experts, popularity, and existence of graphical representations or summaries of the outputs, among many others. The second reason is that, assuming we are only interested in the accuracy, the best technique will depend on the data at hand. Consequently, we present a general discussion of all techniques, and evaluate the trigger mechanisms they produce in three ways.

Table 1: *Structure of a confusion matrix. Note: This table summarizes the alignments and discrepancies between the behaviour of the designed trigger mechanism and the idealized trigger behaviour.*

		Predicted Class	
		$B' = 0$	$B' = 1$
Idealized Class	$B = 0$	True Positive (TP)	False Positive (FP)
	$B = 1$	False Negative (FN)	True Negative (TN)

First, the confusion matrix (Table 1) is obtained for each trigger mechanism. This table summarizes the alignments and discrepancies between the behaviour of the designed trigger mechanism and the idealized trigger behaviour (described in Section 2). In the context of parametric triggers, B' is a function representing the predicted trigger behaviour and B is a function representing the idealized trigger behaviour. In both cases, the function is equal to 1 if the bond triggers and is equal to 0 otherwise.

Next, several metrics are computed from the confusion matrix to quantify performance of each technique's trigger mechanism: error, sensitivity, and specificity. The formulas for computation thereof are shown below.

$$Error = \frac{FP + FN}{TP + FP + FN + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Both false positive and false negative are equally penalized in this framework. In other words, we simply focus on minimizing the total number of errors. The Error metric above quantifies the rate at which the trigger mechanism misclassifies events[1]. Sensitivity characterizes how often the mechanism triggers when it should trigger, and specificity characterizes how often the mechanism does not trigger when it should not. The time required to design the trigger mechanism is also reported for each technique. The metrics described above constitute the numerical evaluation of the trigger mechanisms. Moreover, maps of the resulting trigger patterns are produced for a subset of techniques. This exercise is intended to assess whether classification techniques produce trigger mechanisms with realistic geospatial trigger patterns.

1. Note that error is equal to one minus accuracy.

5.2. Application of classification techniques

As mentioned in Section 2, the design of a parametric trigger mechanism is driven by the minimization of discrepancies between its outputs and those from a trigger with an idealized behaviour (one based directly on the losses). If the resulting trigger mechanism is expected to be useful for new or unseen observations, one should avoid employing the same observations for developing the mechanism and assessing its performance. This could lead to a problem of overfitting (i.e., obtaining complex models that capture specificities of the data but do not generalize well for other observations). An effective technique to avoid this problem is to split the dataset into three subsets: a training set used for constructing the triggers, a validation set employed to tune the parameters, and a test set required to assess their performance. We apply this approach using 50% of the observations for training, 25% for validation, and the remaining 25% for testing. z-score standardization has been applied for all techniques except Classification Trees, Logistic Regression and Clusterwise Logistic Regression. A confidence level of 95% has been considered for the statistical tests. Details of the application of each of the classification techniques are provided in the following paragraphs. The R program (R Core Team, 2012) has been used.

The nearest neighbours classifier. This technique requires a choice of the number of nearest neighbours to consider. Values ranging from 3 to 10 have been tested, and the corresponding accuracies associated to each value have been assessed using the validation set. Ultimately, 5 nearest neighbours are considered for construction of the trigger mechanism, since this provides the highest accuracy but is still small enough to reduce both the variance and the computational time required to make predictions.

Classification trees. Construction of a Classification Tree relies on the selection of the complexity parameter (a parameter that measures the tree cost-complexity). A total of 20 equidistant values from 0.01 up to 0.20 have been tested, and the corresponding accuracies associated with each value have been assessed using the validation set. The value 0.05 has been selected, since it provides the most accurate result. The representation of the tree is shown in Figure 1. Observations which satisfy the condition shown for each internal node terminate to the left; otherwise, they proceed to the right. The percentage shown at the bottom of each node indicates the proportion of observations that reach that node. The value above that percentage refers to the binary classification. For instance, the first condition is '$m < 6.15$', and it is evaluated for all observations (i.e., 100%). Approximately 70% of the earthquakes satisfy this condition and their prediction (i.e., B') is set to 0. The remaining earthquakes are further divided according to the condition '$y < 9.785$'. The same steps are iteratively repeated until a prediction is set for all earthquakes. Thus, eight conditions are considered and only 0.22% of the earthquakes are assigned a value of 1.

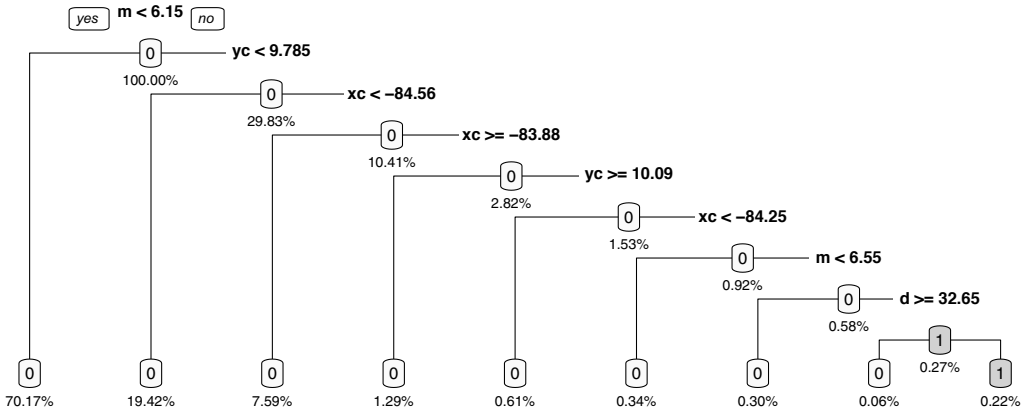

Figure 1: *Classification Tree. Note: Representation of the Tree. Observations which satisfy the condition shown for each internal node terminate to the left; otherwise, they proceed to the right. The percentage shown at the bottom of each node indicates the proportion of observations that reach that node. The value above that percentage refers to the binary classification.*

Neural Networks. Even if complex and powerful Neural Networks exist, we focus on a topology characterized by only one hidden layer. Despite its minimalism, this approach is commonly used, tends to provide good results and is conceptually simple. The number of units in the hidden layer (26) has been tuned by testing the set of values ranging from 10 to 40.

Table 2: *Kernels considered for Support Vector Machines. Note: This table presents some of the most popular kernels for Support Vector Machines in the literature, which are considered in the computational experiments.*

Linear	$k(a,b) = a^T b$
Polynomial [2]	$k(a,b) = (\alpha a^T b + c)^d$
Radial Basis [3]	$k(a,b) = exp(-\gamma \lvert a-b \rvert^2)$
Sigmoid [4]	$k(a,b) = tanh(\sigma a^T b + e)$

Support Vector Machines. In order to efficiently employ this technique, it is required to select a kernel and tune the corresponding parameters. The most popular kernels have been considered and are shown in Table 2. There is also a parameter cost related to the cost of a misclassification for which the following values have been considered: 0.01, 0.1, 1, 5, and 10. Using the validation set, each combination of cost and kernel (including

2. Values tested for α, c, and d, respectively: $\{0.1, 0.2, 0.3, 0.4\}$, $\{0, 0.2, 0.4, 0.6\}$, and $\{2, 3, 4, 5\}$.
3. Values tested for γ: $\{0.1, 0.2, 0.3, 0.4\}$.
4. Values tested for σ and e, respectively: $\{0.1, 0.2, 0.3, 0.4\}$ and $\{0, 0.2, 0.4, 0.6\}$.

the type and the corresponding set of parameters) has been tested by computing the associated accuracy level. According to the results, the best option is a polynomial kernel with the following parameters: $cost = 10$, $\alpha = 0.4$, $c = 0.4$, $d = 4$.

5.3. External validation

In order to validate the application of these techniques to the development of parametric triggers for earthquake catastrophe bonds, we compare our results with those provided by the methodology in Franco (2010). In this paper, the author proposes the construction of binary "cat-in-a-box" trigger mechanisms, where the geographical space is discretized in square boxes or sub-regions of the same size. Each sub-region belongs to a specific zone denoted as k. This approach relies on the concept of optimization and its aim is to determine the parameters of a trigger mechanism for each zone as well as the zone assignment of each sub-region such that the total trigger error is minimized. Concretely, the trigger mechanism for zone k has the following structure:

$$\forall (x_i, y_i) \in A_k, \qquad B_i' = \begin{cases} 0 & \text{if } m_i \leq M_k \quad \text{or } d_i \geq D_k \\ 1 & \text{if } m_i \geq M_k \quad \text{or } d_i \leq D_k \end{cases}$$

where M_k and D_k represent the parametric triggers for the zone, namely the magnitude and depth thresholds, respectively. All sub-regions belonging to zone k have the same trigger structure. An Evolutionary Algorithm (EA) is implemented to address this optimization problem and is executed for different combinations of geographic resolution and number of zones. Although the paper does not report computational times, these methods may consume several hours to perform the parameter optimization.

5.4. Performance

The performance of the trigger mechanisms designed using all nine statistical and machine learning techniques and using the EA employed in Franco (2010) is reported and discussed here. Performance measures are shown in Table 3. Total time takes into account the time to construct the trigger, fine-tune its parameters and test its performance.

A suitable trigger mechanism design should exhibit low error and high specificity and sensitivity and should require minimal computational effort. It can be concluded from the table that the non-linear and non-parametric techniques obtain the best performances of the statistical techniques in terms of accuracy, sensitivity and specificity. In particular, Nearest Neighbours classifier, Classification Trees, Neural Networks and Support Vector Machines are all consistently superior across the three metrics. The results reveal a high variability with respect to computational time, ranging from a few seconds to several minutes. There tends to be trade-off between accuracy and time-required, particularly in the cases of Neural Networks and Support Vector Machines, both of which require significantly more time than the other techniques.

Table 3: *Parametric trigger mechanism performance for ten design techniques. Note: This table shows the performance of the trigger mechanisms designed using all nine statistical and machine learning techniques and using the EA employed in Franco (2010). It considers the error, the sensitivity, the specificity, and the computational time.*

Technique	Error	Sensitivity	Specificity	Time (sec.)
Nearest Neighbours classifier	0.18%	99.84%	**94.44%**	7.22
Naïve Bayes classifier	0.77%	99.58%	4.35%	1.62
Linear Discriminant Analysis	0.64%	99.57%	0.00%	0.28
Quadratic Discriminant Analysis	0.42%	99.63%	57.14%	**0.12**
Classification Trees	0.24%	99.79%	87.50%	2.62
Logistic Regression	0.45%	99.58%	33.33%	0.87
Cluster-wise Logistic Regression	0.43%	99.57%	undefined	5.7
Neural Networks	**0.14%**	**99.94%**	82.14%	190.86
Support Vector Machines	0.27%	99.78%	81.25%	161.25
Evolutionary Algortihm (Franco, 2010)	0.34%	99.86%	55.56%	hours

Several techniques exhibit superior performances to the EA in terms of accuracy, sensitivity and specificity. While EA produces relatively low error rates, the time required is significantly longer than all of the statistical and machine learning techniques.

The triggering events in the idealized trigger mechanism (those for which $B = 1$) comprise less than 0.5% of the total test catalogue, while the other 99.5% of catalogue events do not trigger the idealized bond. Hence, a supposed "null" trigger mechanism in which no events ever trigger the bond would exhibit 99.5% accuracy (0.5% error), 100% sensitivity and 100% specificity. The burden in this case is therefore on any designed trigger mechanisms to outperform this null trigger mechanism benchmark. Eight out of the ten techniques produce trigger mechanisms superior to the null trigger mechanism in terms of accuracy, while the Naïve Bayes Classifier and Linear Discriminant Analysis perform worse by a small margin.

That so few events trigger the bond in the idealized scenario suggests that a larger catalogue might produce more informative and nuanced results using the statistical and machine learning techniques for parametric trigger mechanism design. With a larger catalogue to "learn" from, the techniques would have more triggering events from which to decipher patterns and connections. Reduction of the loss threshold used to construct the idealized trigger scenario would also generate more triggering events from which the statistical techniques could "learn", but since CAT bonds are typically constructed for relatively high return period losses (greater than 100 years), these solutions would not be relevant from a practical standpoint.

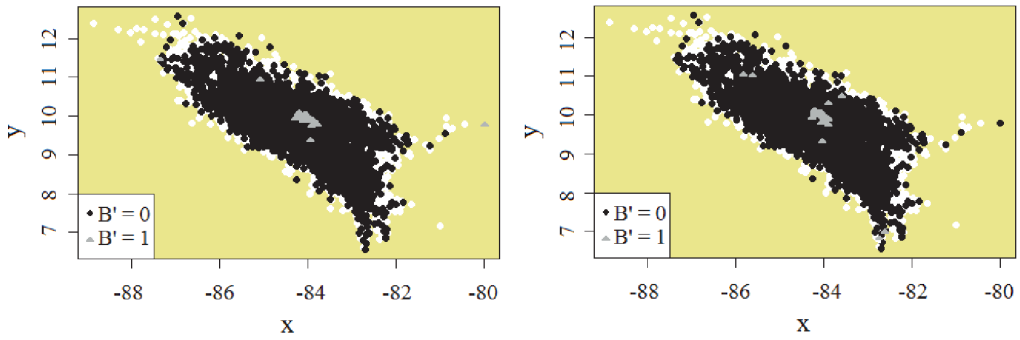

Figure 2: *Map of predictions obtained with Neural Networks (left) and the EA described in Franco (2010). Note: Evaluation of two techniques that produce trigger mechanisms: Neural Networks and the EA from Franco (2010). The first layer is composed of white points, where each point represents an earthquake of the catalogue. The second layer includes the black points, which identify those earthquakes belonging to the test set with* $B' = 0$. *Similarly, the third layer covers the gray triangles, i.e., earthquakes from the test set with* $B' = 1$, *respectively.*

While accuracy is certainly an indispensable feature of any suitable technique for design of parametric trigger mechanisms, a technique should also produce trigger behaviour that is meaningful from a physical perspective. Namely, a suitable technique for parametric trigger mechanism design should produce trigger behaviour that reflects the seismic hazard and/or development patterns in the region of study. For this reason, the physical performance of the techniques trigger mechanisms was evaluated representing earthquakes falling into the test set in maps. Figure 2 shows the evaluation of two techniques that produce trigger mechanisms that are suitable from a numerical perspective: Neural Networks and the EA from Franco (2010). The first layer is composed of white points, where each point represents an earthquake of the catalogue. The second layer includes the black points, which identify those earthquakes belonging to the test set with $B' = 0$. Similarly, the third layer covers the gray triangles, i.e., earthquakes from the test set with $B' = 1$. Note that the plot on the left (Neural Networks) gathers all gray triangles in the centre, while the plot on the right (EA) shows more dispersion.

6. Conclusions and future research

Natural catastrophes continue to cause enormous losses that remain largely uninsured, leaving populations vulnerable to severe financial impacts. The insurance and reinsurance industry, governments and catastrophe pools have started to employ financial instruments such as parametric CAT bonds to cede these catastrophic risks to the capital markets. Were these tools extended for more widespread usage at the retail level, we could progressively and massively reduce the “insurance gap” for earthquake risks. However, this requires the construction of accurate and unbiased parametric triggers

with extreme efficiency and automation, something that is not available in the industry today.

To address this problem, we have explored solving the trigger design challenge as a classification problem, employing well-known and powerful techniques from statistics and machine learning. From a numerical perspective, it has been shown that these techniques can produce trigger mechanisms of equal or better accuracy than previously published techniques (Franco, 2010). Furthermore, several statistical and machine learning methods provide huge efficiency gains in terms of decreasing classification time. Additionally, they provide scalability, being easily adapted to a larger parameter space and larger catalogues without losing much efficiency, and ease of implementation since there is a wide range of programs and programming languages that enable free and simple implementation of these statistical and machine learning techniques such as R (R Core Team, 2012), Octave (Eaton et al., 2014) and Scilab (Scilab Enterprises, 2012). Application of these statistical and machine learning techniques to the problem of parametric trigger design is not without complication, however, because while these methods provide accuracy and efficiency improvements, some of the examples shown in this paper produce trigger mechanisms with relatively low specificity values.

Several lines of future research emerge from the introduction of classification techniques to the development of trigger mechanisms for earthquake CAT bonds. First, it is apparent from the experiments in this paper that more meaningful insights as to the applicability of classification techniques to the development of trigger mechanisms could be gleaned from the use of a larger earthquake catalogue. It would also be worthwhile to examine the behaviour of the trigger mechanisms at multiple return periods, particularly lower ones. There is a natural imbalance in the data at high return periods since very few events trigger the bond. Consequently, there are two groups of events subjected to classification (depending on whether they should trigger a given CAT bond), but they greatly differ in size. Techniques may present low accuracy with respect to the minority (triggering) group and still have a good global accuracy. Analysis of the same simulated earthquake catalogue at lower return periods would reduce this classification group imbalance but would not produce a usable trigger mechanism, since CAT bonds are typically constructed to protect against high return period losses. Therefore, such an experiment could provide valuable insights into the different classification techniques but would not produce directly usable trigger mechanisms. A popular numerical alternative to this complication is to oversample events in the minority group, which would constitute an artificial expansion of the original earthquake catalogue.

Introduction of such a large number of alternative techniques for parametric trigger mechanism design motivates the development of a selection framework. From the standpoint of practical implementation, it would be interesting to identify the most desirable characteristics for a trigger mechanism and order them. For instance, if accuracy is supreme, one should explore the use of more modern and complex techniques such as Random Forests and Multi-Layer Neural Networks (provided a larger catalogue was

available). In contrast, if the interpretability plays the largest role, it would make sense to employ more classical techniques and study graphical tools.

The technological developments characterizing the era of Big Data and the Internet of Things have potentially fascinating implications in this field. These avenues open the possibility of designing triggers not only based on few physical characteristics of an earthquake but on much more information obtained through broad networks of sensors. Metaheuristics, simheuristics (i.e., algorithms combining metaheuristics and simulation techniques) and other classical instruments may be used to perform a feature selection or extraction. Finally, the capacity of simulators to create larger catalogues is ever-increasing, constantly being able to generate more and more data, more and more reliably. In this scenario, non-linear approaches such as Deep Learning would be worth exploration.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness, FEDER (TRA2013-48180-C3-P, TRA2015-71883-REDT) and the Erasmus+ program (2016-1-ES01-KA108-023465). The authors are also thankful to AIR Worldwide for permitting the usage of the AIR Earthquake Model for Central America version 1.1 for the reproduction of the example contained in Franco (2010).

References

Cardenas, V., Hochrainer, S., Mechler, R., Pflug, G. and Linnerooth-Bayer, J. (2007). Sovereign financial disaster risk management: the case of Mexico. *Environmental Hazards*, 7 , 40–53.

Croson, D.C. and Kunreuther, H. (1999). Customizing reinsurance and cat bonds for natural hazard risks. In Conference on Global Change and Catastrophic Risk Management.

Cummins, J.D. (2007). CAT bonds and other risk-linked securities: state of the market and recent developments. *Social Science Research Network*. http://ssrn.com/abtract=105740.

Cummins, J.D. and Weiss, M.A. (2009). Convergence of insurance and financial markets: hybrid and securitized risk-transfer solutions. *Journal of Risk and Insurance*, 76, 493–545.

Damnjanovic, I., Aslan, Z. and Mander, J. (2010). Market-implied spread for earthquake CAT bonds: financial implications of engineering decisions. *Risk Analysis*, 30, 1753–1770.

DeSarbo, W.S. and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, 5, 249–282.

Eaton, J.W., Bateman, D., Hauberg, S. and Wehbring, R. (2014). GNU Octave version 3.8.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. ISBN 1441413006. http://www.gnu.org/software/octave/doc/interpreter.

Franco, G. (2010). Minimization of trigger error in cat-in-a-box parametric earthquake catastrophe bonds with an application to Costa Rica. *Earthquake Spectra*, 26, 983–998.

Franco, G. (2013). Construction of customized payment tables for cat-in-abox earthquake triggers as a basis risk reduction device. In *Proceedings of the International Conference on Structural Safety and Reliability (ICOSSAR)* (pp. 5455–5462). New York, NY.

Franco, G. (2014). Earthquake mitigation strategies through insurance. Chapter *Encyclopedia of Earthquake Engineering*. (pp. 1–18). Springer Berlin Heidelberg.

Goda, K. (2013). Basis risk for earthquake catastrophe bond trigger using scenario-based versus station intensity-based approaches: a case study for southwestern British Columbia. *Earthquake Spectra*, 29, 757–775.

Goda, K. (2014). Seismic risk management of insurance portfolio using catastrophe bonds. *Computer-Aided Civil and Infrastructure Engineering*, 30, 570–582.

Grossi, P. and Kunreuther, H. (2005). *Catastrophe Modeling: A New Approach to Managing Risk*. Springer.

Gunardi and Setiawan, E.P. (2015). Valuation of Indonesian catastrophic earthquake bonds with generalized extreme value (GEV) distribution and Cox-Ingersoll-Ross (CIR) interest rate model. In *AIP Conference Proceedings* (Vol. 1692, pp. 020024).

Guy Carpenter (2014). Chart: gap between economic and insured losses. http://www.gccapitalideas.com/2014/01/20/chart-gap-between-economic-and-insured-losses/.

Härdle, W.K. and Cabrera, B.L. (2010). Calibrating CAT bonds for Mexican earthquakes. *Journal of Risk and Insurance*, 77, 625–650.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. (2nd ed.). Springer.

Kotsiantis, S. (2007). Supervised machine learning: a review of classification techniques. *Informatica*, 31, 249–268.

Lantz, B. (2015). *Machine Learning with R*. Packt Publishing Ltd.

Pucciano, S., Franco, G. and Bazzurro, P. (under review). Loss predictive power for strong motion networks for usage in parametric risk transfer: Istanbul as a case study. *Earthquake Spectra*.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. ISBN 3- 900051-07-0. http://www.R-project.org/.

Scilab Enterprises (2012). *Scilab: Free and Open Source Software for Numerical Computation*. Scilab Enterprises Orsay, France.

Shao, J., Pantelous, A. and Papaioannou, A.D. (2015). Catastrophe risk bonds with applications to earthquakes. *European Actuarial Journal*, 5, 113–138.

Tao, X.X. and Tao, Z.R. (2005). Financial instruments integrated with engineering risk assessment for earthquake disaster reduction. *WIT Transactions on The Built Environment*, 82.

Tao, Z., Tao, X. and Li, P. (2009). Pricing Model for Earthquake CAT Bonds. In *International Conference on Business Intelligence and Financial Engineering* (pp. 740–744).

Wald, D. and Franco, G. (2017). Financial decision-making based on near-realtime earthquake information. In *Proceedings of the 16th World Conference on Earthquake Engineering*. Santiago de Chile. Paper: 3625.

Wu, D. and Zhou, Y. (2010). Catastrophe bond and risk modeling: a review and calibration using Chinese earthquake loss data. *Human and Ecological Risk Assessment*, 16, 510–523.

Zimbidis, A.A., Frangos, N.E. and Pantelous, A.A. (2007). Modeling earthquake risk via extreme value theory and pricing the respective catastrophe bonds. *ASTIN bulletin*, 37, 163–183.

SORT 41 (2) July-December 2017, 1-22

Horizontal collaboration in freight transport: concepts, benefits, and environmental challenges

Adrián Serrano-Hernández[1], Angel A. Juan[2], Javier Faulin[1]
and Elena Perez-Bernabeu[3]

Abstract

Since its appearance in the 1990s, horizontal collaboration (HC) practices have revealed themselves as catalyzers for optimizing the distribution of goods in freight transport logistics. After introducing the main concepts related to HC, this paper offers a literature review on the topic and provides a classification of best practices in HC. Then, the paper analyses the main benefits and optimization challenges associated with the use of HC at the strategic, tactical, and operational levels. Emerging trends such as the concept of 'green' or environmentally-friendly HC in freight transport logistics are also introduced. Finally, the paper discusses the need of using hybrid optimization methods, such as simheuristics and learnheuristics, in solving some of the previously identified challenges in real-life scenarios dominated by uncertainty and dynamic conditions.

MSC: 90B06.

Keywords: Horizontal collaboration, freight transport, sustainable logistics, supply chain management, combinatorial optimization.

1. Introduction

Terms such as 'joint venture', 'network', 'alliance', 'coalition', 'cooperation', 'agreement', or 'partnership' are frequently used in modern business activities. Due to their relevance, they are often accompanied by the 'strategic' adjective. Specifically, the concepts of 'cooperation' and 'collaboration' are occasionally used as synonymous by some authors (as it will be the case in this paper), while others consider that the latter extends the former by also including mutual trust, a higher stage of commitment, etc. Several researchers have tried to rank these terms, obtaining different results depending on the economic sector and criteria considered (Mentzer, Foggin and Golicic, 2000; Golicic, Foggin and Mentzer 2003). As Barratt (2004) concluded, "cooperation is an amorphous

[1] *Corresponding author.* Dept. of Statistics and OR, Institute of Smart Cities, Public University of Navarre, Spain. adrian.serrano@unavarra.es, javier.faulin@unavarra.es

[2] IN3 – Computer Science Dept., Open University of Catalonia, Spain. ajuanp@uoc.edu

[3] Dept. of Applied Statistics, OR and Quality, Universitat Politecnica de Valencia, Spain. elenapb@eio.upv.es

Received: March 2017
Accepted: October 2017

Table 1: *Well-known works providing general overviews on collaboration practices.*

	Unrelated	**Vertical**	**Horizontal**
Marketing	Rokkan, Heide and Wathne (2003)	Zhang et al. (2013)	Czernek (2013)
R&D	Teirlinck and Spithoven (2013)	Zeng et al. (2015)	Roijakkers and Hagedoorn (2006)
NPD	Yam and Chan (2015)	Petersen, Handfield and Ragatz (2005)	Chen (2005)
L&T Maritime	N/A	Álvarez-SanJaime et al. (2013b)	Álvarez-SanJaime et al. (2013a)
L&T Air	N/A	Fu, Homsombat and Oum (2011)	Kuchinke and Sickmann (2005)
L&T Landside	N/A	Bahinipati and Deshmukh (2012)	Cruijssen, Cools and Dullaert (2007b) and Cruijssen, Dullaert and Fleuren (2007c)

meta-concept that has been interpreted in many different ways". According to Hammant (2011), 95% of the companies surveyed implemented some type of collaboration strategy. However, as pointed out by Raue and Wieland (2015), misunderstanding of a collaboration agreement can lead to problems in the inter-firm relationship derived from unmet expectations from one of the sides. On the one hand, inter-firm agreements imply maintaining an independent legal personality while, on the other hand, they also entail the establishment of formulas, protocols, and frameworks that enable the collaboration in some business-related areas: finance, new product development (NPD), research and development (R&D), marketing, logistics and transportation (L&T), etc. Therefore, multiple variants of collaboration practices can occur in these areas. Table 1 classifies some representative works that offer general overviews on the concept of collaboration in different areas, including Marketing, R&D, NPD, and different variants of L&T.

Companies involved in collaboration practices might be related somehow: for example, they might belong to different levels in a supply chain (vertical collaboration) or to the same level in different supply chains (horizontal collaboration or HC). In vertical collaboration, or supply chain management (SCM), agreements take place among companies belonging to different levels inside a supply chain (Chopra and Meindl, 2007). On the contrary, HC refers to joint actions performed by several companies working at the same level of the supply chain and oriented to obtain an enhanced performance in terms of economic and ecologic impact (Bahinipati, Kanda and Deshmukh, 2009). Lambert, Emmelhainz and Gardner (1996) defined HC as a tailored relationship that is based on mutual trust and openness, with the aim of obtaining a competitive advantage – that is, assuming that conjoint performance is higher than the one each partner would achieve on its own. Cruijssen et al. (2007b) consider HC to be an interesting approach to decrease costs, improve service quality, and protect market positions. HC relies on the sharing of

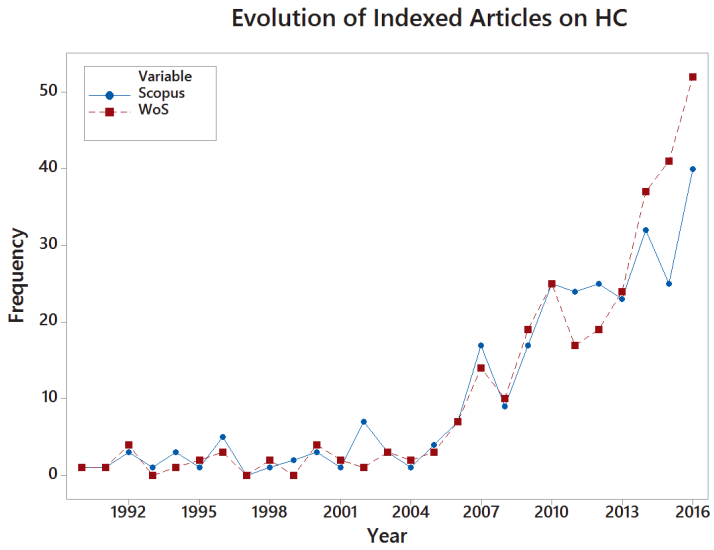

Figure 1: *Evolution of indexed publications related to HC in L&T.*

activities and information, which would necessarily imply sharing operation costs. Through information-sharing, small and medium enterprises expect to act as if they were a large enterprise able to benefit from economies of scale. However, sharing information implies mutual trust, which uses to be a major drawback in most HC practices (Zeng et al., 2015). Vertical collaboration inside supply chains has been intensively studied in the literature (Soosay and Hyland, 2015). There are also studies related to inter-modal transportation, establishing collaborations between truck and ship operators to provide inter-modal services (Saeed, 2013; López-Ramos, 2014). As noticed by some authors (Leitner et al., 2011), the scientific literature related to HC practices is still scarce in comparison with the one dedicated to vertical collaboration, specially in the L&T field. Despite this, during the last decade there has been an increasing interest among researchers in analysing HC practices in L&T. This trend can be observed in Figure 1, which shows the historical evolution of Scopus- and WoS-indexed articles related to the concept of HC in L&T.

This paper aims at partially close this gap in the literature on HC by providing the following contributions: *(i)* it offers an updated literature review on the topic and provides a classification of best practices in HC; *(ii)* it analyses the main benefits and optimization challenges related to the use of HC at the strategic, tactical, and operational levels; *(iii)* it introduces the concept of environmentally friendly, sustainable, or 'green' HC (GHC) in freight transport logistics; and *(iv)* it discusses the need of using hybrid optimization methods, such as simheuristics and learnheuristics, in solving some of the previously identified challenges in real-life scenarios dominated by uncertainty and dynamic conditions. To construct this survey, an intensive search was carried out in Scopus and Web of Science. In this search, the following terms were used: "Horizontal cooperation",

"Horizontal collaboration", "coalition", and "alliance". The search was limited by using keywords such as "logistics", "transportation", and "carrier". In addition, recent articles from well-known authors in the area of HC were analysed in order to complete our set of papers. All in all, a total of 175 references were analysed.

The remaining of this article is structured as follows: Section 2 offers an updated literature review on HC practices; Section 3 offers a classification of HC practices; Section 4 discusses potential benefits of HC at the strategic, tactical, and operational levels, respectively; Section 5 analyses the emergent research field of GHC; Section 6 proposes the use of simheuristics and learnheuristics algorithms for optimizing HC practices in real-life scenarios; finally, Section 7 summarizes the main findings of this work and outlines some future research lines.

2. Literature review on HC concepts

This section offers an exhaustive review of existing works on horizontal collaboration. In order to improve its readability, the review has been organized in the following two subsections: groundworks on horizontal collaboration and works discussing its benefits and challenges.

2.1. Groundworks on horizontal collaboration

In their work related to the grocery sector, Caputo and Mininno (1996) are among the first authors in addressing HC in L&T. these authors highlighted the potential benefits that "cooperation between institutions placed in the same level" could provide. Before 2006, only a few publications explicitly refer to HC in the land-side transportation. Table 2 lists those publications and briefly summarizes their main contributions to the HC field. A turning point took place around 2007, when the topic became much more popular. Distinguished works, such as the ones by Cruijssen et al. (2007b, c), boosted HC and laid the groundwork for upcoming research. Afterwards, the remarkable article by Ballot and Fontane (2010) was published, being the first paper that clearly discussed the environmental impact associated with HC policies. As suggested in Bengtsson and Kock (1999), HC may arise due to trade-offs between cooperation and competition (Figure 2). Two or more companies are 'coexisting' when there are no economic exchanges, i.e., they are neither competing nor cooperating. A 'pure cooperative' scenario takes place among non-competing companies which aim at increasing their value chain through cooperation. A good example is presented in Hsu and Wee (2005), where two non-related manufacturers share information about production, inventory, and delivery in a stochastic environment with the aim of reducing risks. Schmoltzi and Wallenburg (2011) list six different factors of cooperation: contractual scope (type of agreements used), organizational scope (number of participant partners), functional scope (contributors for each functional area), geographical scope (where it will work), ser-

Table 2: *Initial works covering horizontal collaboration.*

Article	**Contributions to the HC field**
Caputo and Mininno (1996)	Propose policies to take advantage of HC in the grocery sector: order management, inventory management, warehousing handling, packaging, and transportation.
Lambert et al. (1996)	Propose a partner-selection model to build horizontal alliances. Define different co-operation types based on facilitator and driver points from surveys.
Zinn and Parasuraman (1997)	Define a framework and a taxonomy to deal with horizontal relationship in logistics activities based on scope and intensity. Discuss the concepts of integrated, extensive, focused, and limited logistics alliances.
Bengtsson and Kock (1999)	Define a framework and describe four types of horizontal relationship that companies might have: coexistence, competition, cooperation, and co-opetition.
Lambert, Emmelhainz and Gardner (1999)	Implement a partner selection model in logistics.
Lau and Liu (2000)	Propose a solving procedure for an inventory management problem and a vehicle routing problem with time windows in a collaborative environment.
Bahrami (2002)	Discusses the possibility of considering HC within supply chains as an option to increase productivity. It shows a real case of two German companies that merged their distribution network, comparing a traditional situation against two alternative HC scenarios (one preserving the current logistics network and other modifying it).
Golicic et al. (2003)	Describe a series of focus-group practices aimed at discussing and identifying inter-organizational relationships. A chaotic paradigm of cooperation is presented as a result of the variety of opinions.
Barratt (2004)	Identifies elements of collaboration (joint decision making, supply chain metrics, etc.) as well as the consequences of misunderstanding cooperation concepts.
Hageback and Segerstedt (2004)	Propose HC in rural areas as a way to stop depopulation.
Groothedde, Ruijgrok and Tavasszy (2005)	Quantify economies of scale achieved through cooperation
Krajewska and Kopfer (2006)	Explain how to perform HC practices between partners having similar characteristics. Propose a model that includes the re-distribution of profit. The model is based on the combinatorial auction theory and on game theory.

vice scope (which services are offered), and resource scope (corporate characteristics of each partner). 'Competition' arises among companies focused on the same target group. Relationships among competitors are based on action-reaction patterns, and they involve a limited information flow. 'Co-opetition' occurs when HC is jointly developed by competing firms. Trust and commitment become key elements to achieve fruitful relationships while keeping competition. In the L&T sector co-opetition is probably the most usual context (Limoubpratum, Shee and Ahsan, 2015).

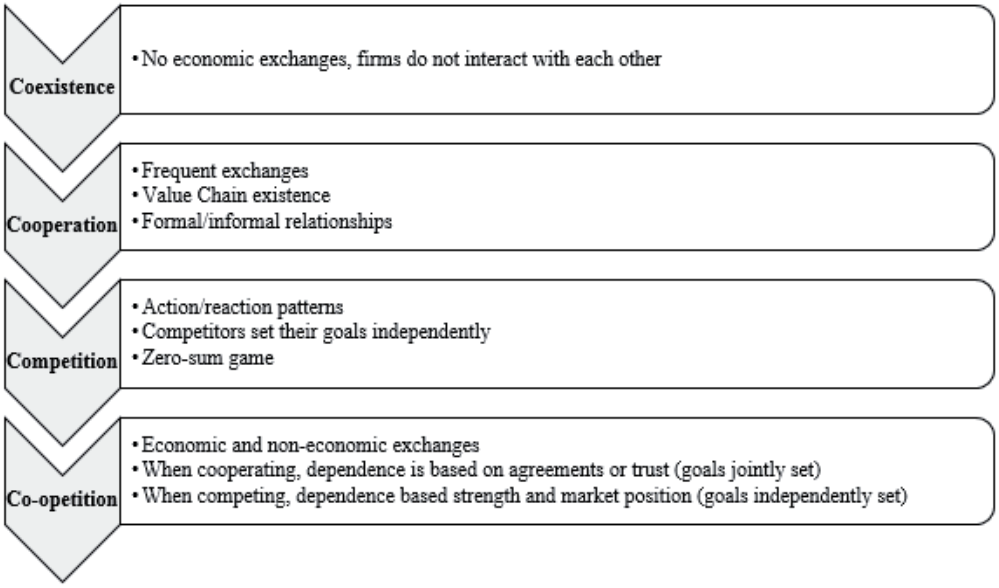

Figure 2: *Horizontal relationships among enterprises, based on Bengtsson and Kock (1999).*

2.2. Works discussing benefits and challenges of HC

Reducing transportation costs is one of the most pursued goals in HC. However, many other benefits may be achieved: for example, improving service quality, diminishing environmental impact, reducing risk, and enhancing market share. Table 3 shows relevant references covering some of the previous purposes. The existing literature also contains experiences describing the use of HC practices in non-profit associations, as in Schulz and Blecken (2010). These authors try to adapt HC practices to disaster relief logistics, describing both benefits and issues related to these practices. According to them, the main challenges when implementing HC strategies are related to: *(i)* how to establishing mutual trust among cooperating firms; and *(ii)* how to achieving a fair redistribution of both costs and profits among the partners. Due to their complex nature, HC practices offer high potential for conflicts or disagreements (Raue and Wieland, 2015; Wallenburg and Raue, 2011; Adenso-Díaz, Lozano and Moreno, 2014). Difficulty to find a suitable partner is another issue when dealing with HC (Lambert et al., 1999). On the one hand, a good knowledge of the potential partners' assets is required to evaluate the candidates. On the other hand, companies must share a common goal. A survey on profits / costs allocation is provided in Guajardo and Rönnqvist (2016), whereas Liu et al. (2010) focus on the less-than-truckload segment. These authors review over 40 different methodologies to share costs and profits in a coalition. However, as noticed by Yengin (2012), the Shapley's method is the most recurrent approach in the literature due to its clarity and simplicity. Table 4 summarizes recent references covering some of the main challenges

Table 3: *Main HC goals in the scientific literature.*

Objectives	**Discussed in**
Reducing transportation costs	Soysal et al. (2016), Fernández, Fontana and Speranza (2016), Bottani, Rizzi and Vignali (2015), Vornhusen, Wang and Kopfer (2014), Verdonck et al. (2013), Audy et al. (2012)
Improving service quality	Ghaderi, Dullaert and Amstel (2016), Lehoux, Damours and Langevin (2014)
Reducing environmental impact	Danloup et al. (2015), Perez-Bernabeu et al. (2015), Juan et al. (2014), Pan et al. (2014), Pradenas, Oportus and Parada (2013), Peetijade and Bangviwat (2012)
Reducing risk	Stojanović and Aas (2015), Li et al. (2012), Bahinipati et al. (2009)
Enhancing market share	Wei, Zhao and Li (2015), Gou et al. (2014)

Table 4: *Main HC challenges discussed in the scientific literature.*

Challenges	**Discussed in**
Difficulty to ensure relationships based on trust	Zeng et al. (2015), Raue and Wieland (2015), Schmoltzi and Wallenburg (2012), Wilhelm (2011)
Difficulty to find suitable partners	Ayadi, Halouani and Masmoudi (2016), Dao, Abhary and Marian (2014), Raue and Wallen-burg (2013), Audy et al. (2012), Asawasakulsorn (2015), Bahinipati et al. (2009)
Difficulty to share profits/losses	Guajardo and Rönnqvist (2016), Kimms and Kozeletskyi (2016), Guajardo and Rönnqvist (2015), Defryn, Sörensen and Cornelissens (2016), Karsten, Slikker and Van Houtum (2015), Vanovermeire et al. (2014), Lozano et al. (2013), Frisk et al. (2010), Dai and Chen (2012), Liu et al. (2010), Massol and Tchung-Ming (2010), Dai and Chen (2015), Frisk et al. (2010), Xu et al. (2009)
Difficulty to establish an appropriate framework	Lehoux et al. (2014), Leitner et al. (2011), Cruijssen et al.(2010), Pomponi et al. (2013), Nadarajah and Bookbinder (2013), Audy et al. (2012)

associated with HC practices. Older references can be found in Cruijssen (2006), Cruijssen et al. (2007b), Cruijssen et al. (2007c) and Pomponi et al. (2013).

3. Classification of HC practices

Several criteria have been proposed to classify HC practices. In this paper, we focus on the taxonomies proposed by Zinn and Parasuraman (1997), Lambert et al. (1999), and Pomponi et al. (2013) since they offer complete and easy-to-implement classification systems. In order to compare these taxonomies, some common factors and levels have

Table 5: *Factors and levels to classify HC practices.*

Factor	**High**	**Mid**	**Low**
Time Frame	More than 3 years	Between 1 and 3 years	Less than 1 year
Amplitude	Whole company involved	Just a division	Few aspects of the company involved
Stamina	Legal contract	No contract but formal rules	Just relational rules
Organizational level	Strategic	Tactical	Operational

been identified in Table 5. The main factors are: time frame, amplitude, stamina, and closeness. Time frame refers to the duration of the agreement. Amplitude refers to the level of commitment in terms of range of pooled services: for example, fleet, information, orders, warehouses, etc. Stamina is the ability of the coalition to survive by means of legal contracts, conjoint investments, etc. Finally, the organizational level denotes characteristics of the conjoint project, such as operational, tactical, or strategic ones. For each factor, three intensity levels are presented.

One of the first attempts to categorize HC practices in L&T was presented in Zinn and Parasuraman (1997). These authors proposed a taxonomy based on the intensity and scope of the coalition. The former relates to the extent of direct involvement among allies, whereas the latter refers to the range of involved services. By combining intensity and scope, four types of cooperation arise (Table 6).

Table 6: *Taxonomy proposed by Zinn and Parasuraman (1997) for HC practices.*

	Time Frame	**Amplitude**	**Stamina**	**Organizational level**
Limited	Low	Low	Low	Low
Extensive	Low	Mid	Mid	Low
Focused	Low	Low	Mid	Mid
Integrated	Mid-High	High	High	Mid-High

Another approach for classifying HC practices is provided by Lambert et al. (1996), who consider three types of cooperation (Table 7). Type I cooperation represents agreements in which the involved companies recognize each other as partners and coordinate their activities on a limited basis for a very short time. Type II cooperation denotes a medium-term relationship for an entire project duration and a greater level of cooperation. In contrast, in Type III cooperation firms have a high level of integration for an unlimited duration, thus involving the entire organization. In that classification, an increasing level of trust is assumed: that is, a Type I cooperation is required before a Type II one.

Table 7: *Taxonomy proposed by Lambert et al. (1996) for HC practices.*

	Time Frame	**Amplitude**	**Stamina**	**Organizational level**
Type I	Low	Low	Low	Low
Type II	Mid	Mid	Mid	Mid-High
Type III	Mid-High	High	High	Mid-High

Finally, Pomponi et al. (2013) did not consider time restrictions and designed a framework in which cooperation is categorized based on its organizational level: operational, tactical, or strategic (Table 8).

Table 8: *Taxonomy proposed by Pomponi et al. (2013) for HC practices.*

	Time Frame	**Amplitude**	**Stamina**	**Organizational level**
Operational	Low-Mid-High	Low	Low	Low
Tactical	Low-Mid-High	Mid	Mid	Mid
Strategic	Low-Mid-High	High	High	High

As in many other areas, it is not easy to find a universal classification for all HC practices in L&T. However, this section has identified several key factors that are common in the several works and which refer to a correct understanding of a collaboration agreement in terms of duration, amplitude, legal form, and organizational level involved.

4. Quantifying the benefits of hc in freight transport logistics

By taking advantage of economies of scale, HC practices contribute to increase firms' efficiency and competitiveness. Hence, cost reduction, improvement of service quality, and mitigation of CO_2 emissions are the main benefits of HC in road freight transportation. Table 9 summarizes recent outcomes of different research works, including the approaches adopted and their impact on costs. Notice that in some cases there is a high variability depending on factors such as the topology of the distribution network, the degree of cooperation, and the specific cooperative mechanism adopted. In those cases, a short explanation is provided as a footnote to the table. Since the existing literature presents several ways of achieving benefits depending on the decision level involved (strategic, tactical, or operational), the following subsections discuss preeminent approaches used in HC for each of these levels.

Table 9: *Summary of recent outcomes applying HC approaches and their impact.*

Level	Advantages	Disadvantages	References	Impact on costs
Tactical (conjoint routes)	It does not require a high level of integration	Revenue contracts are required	Dahl and Derigs (2011)	−14%
			Wang and Kopfer (2014)	−11%
			Muñoz-Villamizar, Montoya-Torres and Vega-Mejía (2015)	−25%
			Perez-Bernabeu et al. (2015)	−5% to −90% (1)
			Wang, Kopfer and Gendreau (2014b)	−89%
			Cruijssen et al. (2007a)	−31%
			Özener, Ergun and Savelsbergh (2011)	−26% to −30% (2)
Strategic (consolidation centers)	Relatively easy to apply	A large capital investment is required	Groothedde et al. (2005)	−14%
			Vornhusen et al. (2014)	−18%
			Verdonck et al. (2013)	−22%
			Wang et al. (2014a)	−5% to −50% (3)
			Cruijssen et al. (2010)	−8%
Operational (load factors)	Relatively easy to apply	A high level of trust and commitment is required Revenue contracts are required	Li (2013)	−28%
			Bailey, Unnikrishnan and Lin (2011)	−27%
			Sprenger and Mönch (2012)	−25%
			Hernández and Peeta (2014)	−2% to −55% (4)

(1): −5% in a clustered topology and −90% in scattered topology
(2): −26% without a mechanism of side payments and −30% with that mechanism
(3): −5% when companies look for a high profit margins and −50% when it is low
(4): −2% when low degree of collaboration and −55% when it is high

4.1. Strategic level – consolidation centres

Strategic decisions are carried out for a long-time period and involve the whole company. Determining the best location for the distribution centres of a firm is a typical example of such a strategic decision. Figure 3 describes an illustrative case in which firms must serve all the customers placing orders to them. In a collaborative scenario, some consolidation centres are selected to distribute products among customers in the nearby. As described in Verdonck et al. (2016), fixed assets such as warehouses and distribution centres can be shared in order to consolidate production from several manufactures, thus reducing the number of long-trip deliveries required. Collaborative hubs are proposed by Groothedde et al. (2005) to deal with a real case developed in The Netherlands. These authors also provide a methodology to assess the benefits obtained through collaboration. Transshipments, as a collaborative strategy in shared warehouses, are explored in

Vornhusen et al. (2014). The introduction of transshipments reaches significant cost reductions when compared against isolated planning and even to centralized planning. A similar collaborative hub is proposed in Cruijssen et al. (2010), where a step-wise approach is formulated considering potential savings in infrastructures that require large investments.

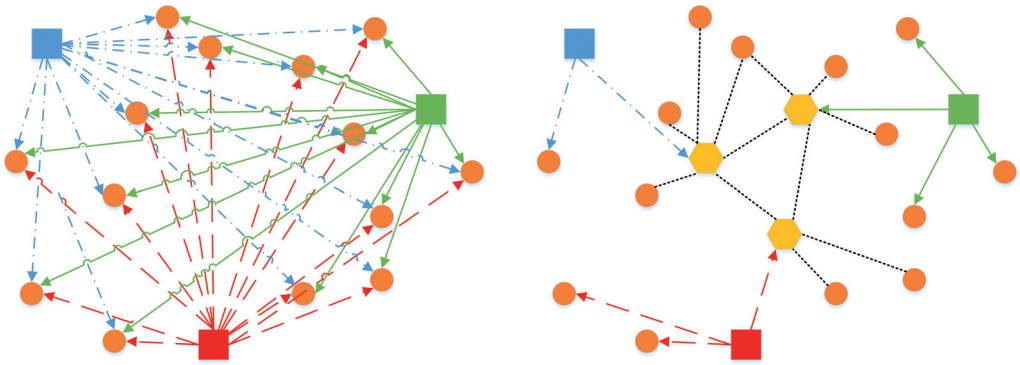

Figure 3: *Non-collaborative (left) vs. collaborative scenarios (right) for freight consolidation.*

4.2. Tactical level – conjoint routes

Tactical decisions are focused on the mid-term and they typically require a high level of synchronization among the departments of a firm. In this context, the use of conjoint routes emerges as the primary source of cost saving: two or more companies pool their customers to serve them from a shared depot. Therefore, clients' orders are exchanged to get a better match between customers and depots. Most articles start with a non-collaborative scenario, after which they analyse the potential benefits that could be obtained if a collaborative scenario was used instead. That is the case of Perez-Bernabeu et al. (2015), who compared clustered and scattered non-collaborative scenarios against the collaborative one. Similarly, Muñoz-Villamizar et al. (2015) focused on the last-mile distribution to develop a collaborative planning for carriers and assuming stochastic demands. Considering a less-than-truckload framework, Wang and Kopfer (2014) introduced a pick-up and delivery problem with time windows to illustrate HC benefits. Similarly, Nadarajah and Bookbinder (2013) considered a two-stage framework for less-than-truckload transportation: firstly, collaboration between multiple carriers at the entrance of a city is considered; secondly, there is a carrier collaboration for transshipment to finalize the initial routes. Finally, Dahl and Derigs (2011) developed a real-time collaborative decision support system in the express carrier network. Their main purpose is assessing potential benefits obtained by sharing customers. Broadly speaking, it represents moving from several vehicle routing problems to one multi-depot vehicle routing problem, as depicted in Figure 4.

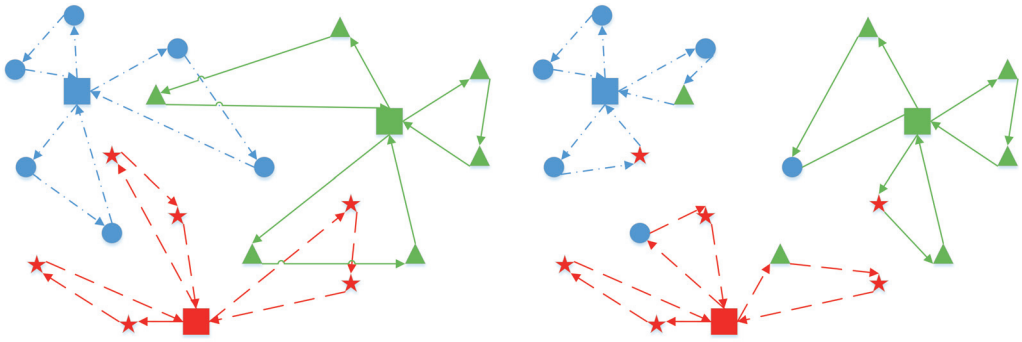

Figure 4: *Non-collaborative (left) vs. collaborative (right) scenarios for conjoint routes.*

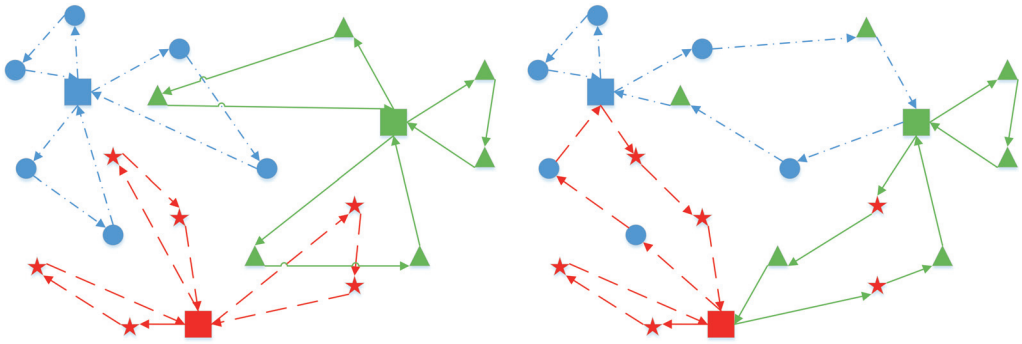

Figure 5: *Non-collaborative (left) vs. collaborative (right) scenarios for back-hauling.*

4.3. Operational level – load factors

Cooperation is an efficient way to increase load factors, thus avoiding lack of efficiency in transport activities. HC approaches can help to raise these load factors in several ways, e.g.: *(i)* by sharing the vehicle capacity among different companies; and *(ii)* by employing collaborative back-hauling. As pointed out by Hernández and Peeta (2014), sharing the vehicle capacity can significantly increase load factors, since it generates the potential to gain revenue on non-full haul trips. These authors run several sensitivity analyses based on the degree of cooperation and fuel prices. In a similar way, Sprenger and Mönch (2012) discussed the concept of vehicles sharing within a German food industry. They also proposed a methodology for a collaborative transportation planning problem in a rolling horizon setting. For this problem, they used simulation to characterize the dynamic and stochastic transport system. Usually, customers are widespread over the geography, which generates long empty back-hauls after deliveries. Thus, load factors can be easily improved by collaborating to reduce empty back-hauls when companies share their logistics operations (Figure 5). Thus, after completing its route, a

vehicle may finish in a depot different from the initial one. That is the case studied in Li (2013), who showed that load factors could reach 92% by using such a collaborative strategy. Likewise, Bailey et al. (2011) investigated possible reductions in empty backhauls by considering customer requests from partners.

5. Environmental issues in horizontal collaboration

As noticed by Allen et al. (2017), one of the main advantages of HC practices is the reduction of the externalities associated with freight transportation. According to Belien et al. (2017), the main HC benefits include: *(i)* a 20-25% diminution in CO_2 emissions; *(ii)* a 10% improvement in transport reliability; and *(iii)* a 10-15% reduction in transportation cost. Following Demir et al.(2015), it is possible to classify these externalities into seven dimensions or impact groups: air pollution, greenhouse gas emissions, noise pollution, water pollution, traffic congestion, traffic accidents, and use of land by transport infrastructure. Despite all of these groups are relevant, air pollution and greenhouse gas emissions are likely to be the externalities that cause a higher social alarm. Green or sustainable HC refers to the use of HC practices that, by making a more efficient use of resources, contribute to reduce the environmental impact of L&T activities.

In effect, freight transport logistics generates emissions of greenhouse gases: carbon dioxide (CO_2), nitrous oxide, and methane. CO_2 is the dominant greenhouse gas, and the remaining gases can be expressed as CO_2 equivalents (Lera-López et al., 2014). Road transportation, as the primary mode of freight movement, is the largest source of freight-related CO_2 emissions in most developed countries. International agreements, such as the Kyoto Protocol and the Doha Amendment to Kyoto Protocol are pushing developed countries to accomplish a reduction in gas emissions. National policies have a great influence on transportation companies, which start to promote internal policies towards the development of environmentally-friendly supply chains. Aiming at reducing CO_2 emissions, countries such as UK have implemented strict government regulations (Ramanathan, Bentley and Pang, 2014). According to the International Energy Agency, worldwide CO_2 emissions due to fuel consumption raised a 56.4% from 1971 to 2013, whereas in OECD countries it raised just a 9.4% in the same time period (IEA, 2015). CO_2 emissions in the transport sector, and their contribution to climate change, represent one of the main issues in the sustainable management of logistics activities.

HC practices contribute to make the transportation sector more sustainable by means of the following policies: *(i)* design of conjoint routes in freight delivery, which leads to shorter distribution networks; *(ii)* sharing of responsibilities during the last-mile distribution, which allows to achieve 'greener' routes and to reduce the logistics activities in city centres; and *(iii)* construction of large-scale logistics scenarios, which benefit from a reduction in uncertainty –thus generating solutions involving less vehicles and routes.

As previously highlighted, the design of conjoint routes emerges as the primary source of reducing gas emissions. Insights on this topic are presented in Danloup et

Table 10: *Summary of recent works on Green HC.*

Level	Reference	Impact on CO_2
Tactical (conjoint routes)	Soysal et al. (2016)	−29%
	Danloup et al. (2015)	−26%
	Andriolo et al. (2015)	−27% to −50% (1)
	Perez-Bernabeu et al. (2015)	−5% to −92% (2)
	Özener (2014)	−5%
Strategic (consolidation centres)	van Lier, Caris and Macharis (2016)	−7%
	Pan et al. (2014)	−19%
	Pan, Ballot and Fontane (2013)	−14%
	Ballot and Fontane (2010)	−25%
Operational (load factors)	Basu, Bai and Palaniappan (2015)	−66%
	Pradenas et al. (2013)	−30%
	Juan et al. (2014)	−27%
	Lin and Ng (2012)	−3% to −20% (3)

(1): depending on the lot sizing policy applied
(2): −5% in a clustered topology and −92% in scattered topology
(3): depending on purchasing-of-carbon rights

al. (2015). These authors analysed how it was possible to reduce CO_2 emissions by simply increasing the loading factor of the trucks. In a similar way, Özener (2014) tested an extensive set of instances to assess CO_2 reduction. Freight consolidation is also another driver to reach environment-friendly logistics management. As described in Ballot and Fontane (2010), warehouses and distribution centres can be shared to consolidate production from several manufactures, thus reducing the number of deliveries required. Through a case study run in France, these authors showed that freight consolidation could achieve a significant reduction of CO_2 emissions. Another case study in France was conducted by Pan et al. (2014), where three different scenarios were compared to the original one in terms of CO_2 emissions. Internal collaboration is explored in van Lier et al. (2016). A summary of green HC references is displayed in Table 10. Again, a high variability occurs due to factors such as the distribution network topology, the degree of cooperation, and the specific cooperative mechanism adopted.

6. Dynamism and uncertainty in real-life HC practices

The existing body of research on HC optimization mainly assumes deterministic and static models to describe freight transport systems. However, real-life optimization problems in the area of horizontal collaboration are usually characterized by properties such as large-scale dimension, dynamic conditions, and stochastic elements. In effect, since HC practices imply the aggregation of different distribution companies and their asso-

ciated customers, the size of the resulting problems tends to be much larger than the one associated with any individual partner. Since several combinatorial problems in the L&T area are NP-hard in nature, the use of metaheuristic algorithms is usually required to cope with these large-scale instances. Moreover, since HC optimization problems typically consider heterogeneous enterprises and their customers, they usually offer a high degree of dynamism and uncertainty: the working conditions (and their related constraints) might be different from one company to another, the availability of shared resources might depend upon changing environmental conditions, the customers' demands might vary according to the assigned distributor or distribution time, etc. Fortunately, different hybrid algorithms can be utilized to solve rich and real-life optimization HC challenges in L&T. Prominent examples are matheuristics that arise from integration of exact and metaheuristic methods (Doerner and Schmid, 2010), or simheuristics (Juan et al., 2015) that result from combination of simulation with metaheuristics. Different works discuss how metaheuristics can be employed to solve optimization problems under uncertainty scenarios (Bianchi et al., 2009). In particular, simheuristics allow to integrate real-life uncertainty both as part of the objective function and as probabilistic constraints in the optimization problems. Recent examples on the application of simheuristics to deal with horizontal collaboration problems under uncertainty can be found in the literature. Thus, Gruler et al. (2017) propose a simheuristic approach to optimize a waste collection problem in clustered urban areas where horizontal collaboration strategies are considered by different city managers. Likewise, Quintero-Araujo et al. (2017) propose the use of simheuristics to promote HC practices in city logistics under uncertainty conditions. Finally, de Armas et al. (2017) propose a simheuristic approach to solve large-scale facility location problems with stochastic demands –notice that this problem is strongly related to the use of consolidation centres in HC practices. In a similar way, by combining metaheuristics with statistical-learning techniques, learnheuristics allow to efficiently deal with the high level of dynamism around modern freight transport systems (Calvet et al., 2017, 2016b). Thus, for instance, in Calvet et al. (2016a) the authors propose the integration of statistical learning inside a metaheuristic framework to deal with a multi-depot distribution problem with dynamic users' demands. The ensuing models represent more accurately real-world freight transport scenarios. Among other strengths, these hybrid methods accommodate elements of uncertainty (stochastic factors) and dynamism (evolving environmental conditions). As solution methods and techniques grow rapidly in complexity, scale, and scope, and they can easier find their way in solving more practical instances across a number of fields, a further emergence of sustainable and green HC problems considering complex multi-objective functions and probabilistic constraints is warranted.

7. Conclusions

As analysed in this paper, horizontal collaboration (HC) practices represent an efficient way of reducing costs in freight transport logistics and promote environmentally-friendly policies. For that reason, the analysis of 'green' or sustainable HC practices is gaining importance in the recent literature. By using sustainable HC in freight transport logistics, small-size carriers may not only achieve greater economies of scale –thus increasing their competitiveness levels in a global market–, but also contribute to minimize the environmental impact of their business activities. Trust-related issues among companies, as well as difficulty to allocate costs and profits among partners are the main barriers to implement HC practices in real-life scenarios.

In this paper, a classification of HC activities has been provided, as well as an analysis of the benefits and challenges that HC practices can provide at each decision-making level: strategic (consolidation centres), tactical (conjoint routes), and operations (load factors). Since these practices often imply solving combinatorial optimization problems characterized by a large-scale dimension and the existence of stochastic / dynamic conditions, the use of hybrid algorithms (e.g., simheuristics and learnheuristics) is proposed as one of the most efficient ways to cope with rich and real-life HC optimization problems.

The emergence of new optimization methods, as well as the continuous increase in computational power, allow to consider several research lines for the future, including: *(i)* the inclusion of multiple goals (e.g., monetary, environmental, sustainability indexes, etc.) in the function to be optimized; and *(ii)* the modeling and solving of realistic freight transport logistics scenarios including time-evolving and stochastic inputs (e.g., dynamic availability of shared resources, variable customers' demands depending on the assigned carrier, etc.).

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness & FEDER (TRA2013-48180-C3-P, TRA2015-71883-REDT), The Erasmus+ program (2016-1-ES01-KA108-023465), the Ibero American Program of Science and Technology for Development (CYTED2014-515RT-0489), the CAN Foundation (CAN2014-3758, CAN2015-70473), and the Spanish Ministry of Education (FPU-14/00024).

References

Adenso-Díaz, B., Lozano, S. and Moreno, P. (2014). Analysis of the synergies of merging multi-company transportation needs. *Transportmetrica A: Transport Science*, 10, 533–547.

Allen, J., Bektas, T., Cherret, T., Friday, A., McLeod, F., Piecyk, M., Piotrowska, M. and Zaltz-Austwick, M. (2017). Enabling a freight traffic controller for collaborative multi-drop urban logistics: Practical and theoretical challenges. *Transportation Research Record: Journal of the Transportation Research Board*, 2609, 77–84.

Álvarez-SanJaime, O., Cantos-Sánchez, P., Moner-Colonques, R. and Sempere-Monerris, J. J. (2013a). Competition and horizontal integration in maritime freight transport. *Transportation Research Part E: Logistics and Transportation Review*, 51, 67–81.

Álvarez-SanJaime, O., Cantos-Sánchez, P., Moner-Colonques, R. and Sempere-Monerris, J. J. (2013b). Vertical integration and exclusivities in maritime freight transport. *Transportation Research Part E: Logistics and Transportation Review*, 51, 50–61.

Andriolo, A., Battini, D., Persona, A. and Sgarbossa, F. (2015). Haulage sharing aproach to achieve sustainability in material purchasing: new method and numerical applications. *International Journal of Production Economics*, 164, 308–318.

de Armas, J., Juan, A. A., Marquès, J. M. and Pedroso, J. P. (2017). Solving the deterministic and stochastic uncapacitated facility location problem: from a heuristic to a simheuristic. *Journal of the Operational Research Society*, (pp. 1–16).

Asawasakulsorn, A. (2015). Transportation collaboration: partner selection criteria and interorganizational system design issues for supporting trust. *International Journal of Business and Information*, 4.

Audy, J.-F., Lehoux, N., D'Amours, S. and Rönnqvist, M. (2012). A framework for an efficient implementation of logistics collaborations. *International transactions in operational research*, 19, 633–57.

Ayadi, O., Halouani, N. and Masmoudi, F. (2016). A fuzzy collaborative assessment methodology for partner trust evaluation. *International Journal of Intelligent Systems*, 31, 488–501.

Bahinipati, B.K. and Deshmukh, S. (2012). Vertical collaboration in the semiconductor industry: a decision framework for supply chain relationships. *Computers & Industrial Engineering*, 62, 504–526.

Bahinipati, B.K., Kanda, A. and Deshmukh, S. (2009). Horizontal collaboration in semiconductor manufacturing industry supply chain: an evaluation of collaboration intensity index. *Computers & Industrial Engineering*, 57, 880–895.

Bahrami, K. (2002). Improving supply chain productivity through horizontal cooperationthe case of consumer goods manufacturers. In *Cost Management in Supply Chains* (pp. 213–232). Springer.

Bailey, E., Unnikrishnan, A. and Lin, D.-Y. (2011). Models for minimizing backhaul costs through freight collaboration. *Transportation Research Record: Journal of the Transportation Research Board*, (pp. 51–60).

Ballot, E. and Fontane, F. (2010). Reducing transportation CO_2 emissions through pooling of supply networks: perspectives from a case study in french retail chains. *Production Planning & Control*, 21, 640–650.

Barratt, M. (2004). Understanding the meaning of collaboration in the supply chain. *Supply Chain Management: an International Journal*, 9, 30–42.

Basu, R.J., Bai, R. and Palaniappan, P.K. (2015). A strategic approach to improve sustainability in transportation service procurement. *Transportation Research Part E: Logistics and Transportation Review*, 74, 152–168.

Belien, J., Boute, R., Creemers, S., De Bruecker, P., Gijsbrechts, J., Padilla Tinoco, S.V. and Verheyen, W. (2017). Collaborative shipping: logistics in the sharing economy. *ORMS Today*, 44, 20–23.

Bengtsson, M. and Kock, S. (1999). Cooperation and competition in relationships between competitors in business networks. *Journal of Business & Industrial Marketing*, 14, 178–194.

Bianchi, L., Dorigo, M., Gambardella, L.M. and Gutjahr, W.J. (2009). A survey on metaheuristics for stochastic combinatorial optimization. *Natural Computing*, 8, 239–287.

Bottani, E., Rizzi, A. and Vignali, G. (2015). Improving logistics efficiency of industrial districts: a framework and case study in the food sector. *International Journal of Logistics Research and Applications*, 18, 402–423.

Calvet, L., Armas, J.d., Masip, D. and Juan, A.A. (2017). Learnheuristics: hybridizing metaheuristics with machine learning for optimization with dynamic inputs. *Open Mathematics*, 15, 261–280.

Calvet, L., Ferrer, A., Gomes, M.I., Juan, A.A. and Masip, D. (2016a). Combining statistical learning with metaheuristics for the multi-depot vehicle routing problem with market segmentation. *Computers & Industrial Engineering*, 94, 93–104.

Calvet, L., Juan, A.A., Serrat, C. and Ries, J. (2016b). A statistical learning based approach for parameter fine-tuning of metaheuristics. *SORT-Statistics and Operations Research Transactions*, 1, 201–224.

Caputo, M. and Mininno, V. (1996). Internal, vertical and horizontal logistics integration in italian grocery distribution. *International Journal of Physical Distribution & Logistics Management*, 26, 64–90.

Chen, S. (2005). Task partitioning in new product development teams: A knowledge and learning perspective. *Journal of Engineering and Technology Management*, 22, 291–314.

Chopra, S. and Meindl, P. (2007). Supply chain management. Strategy, planning & operation. *Das summa summarum des management*, (pp. 265–275).

Cruijssen, F., Borm, P., Dullaert, W. and Hamers, H. (2010). A versatile framework for cooperative hub network development. *European Journal of Industrial Engineering*, 4, 210–227.

Cruijssen, F., Bräysy, O., Dullaert, W., Fleuren, H. and Salomon, M. (2007a). Joint route planning under varying market conditions. *International Journal of Physical Distribution & Logistics Management*, 37, 287–304.

Cruijssen, F., Cools, M. and Dullaert, W. (2007b). Horizontal cooperation in logistics: opportunities and impediments. *Transportation Research Part E: Logistics and Transportation Review*, 43, 129–142.

Cruijssen, F., Dullaert, W. and Fleuren, H. (2007c). Horizontal cooperation in transport and logistics: a literature review. *Transportation Journal*, (pp. 22–39).

Cruijssen, F.C.A.M. (2006). *Horizontal Cooperation in Transport and Logistics*. Ph.D. thesis Universiteit van Tilburg. The Netherlands.

Czernek, K. (2013). Determinants of cooperation in a tourist region. *Annals of Tourism Research*, 40, 83–104.

Dahl, S. and Derigs, U. (2011). Cooperative planning in express carrier networks – An empirical study on the effectiveness of a real-time decision support system. *Decision Support Systems*, 51, 620–626.

Dai, B. and Chen, H. (2012). Profit allocation mechanisms for carrier collaboration in pickup and delivery service. *Computers & Industrial Engineering*, 62, 633–643.

Dai, B. and Chen, H. (2015). Proportional egalitarian core solution for profit allocation games with an application to collaborative transportation planning. *European Journal of Industrial Engineering*, 9, 53–76.

Danloup, N., Mirzabeiki, V., Allaoui, H., Goncalves, G., Julien, D. and Mena, C. (2015). Reducing transportation greenhouse gas emissions with collaborative distribution: a case study. *Management Research Review*, 38, 1049–1067.

Dao, S.D., Abhary, K. and Marian, R. (2014). Optimisation of partner selection and collaborative transportation scheduling in virtual enterprises using ga. *Expert Systems with Applications*, 41, 6701–6717.

Defryn, C., Sörensen, K. and Cornelissens, T. (2016). The selective vehicle routing problem in a collaborative environment. *European Journal of Operational Research*, 250, 400–411.

Demir, E., Huang, Y., Scholts, S. and Van Woensel, T. (2015). A selected review on the negative externalities of the freight transportation: modeling and pricing. *Transportation research part E: Logistics and Transportation Review*, 77, 95–114.

Doerner, K.F. and Schmid, V. (2010). Survey: matheuristics for rich vehicle routing problems. *Hybrid Metaheuristics*, 6373, 206–221.

Fernández, E., Fontana, D. and Speranza, M.G. (2016). On the collaboration uncapacitated arc routing problem. *Computers & Operations Research*, 67, 120–131.

Frisk, M., Göthe-Lundgren, M., Jörnsten, K. and Rönnqvist, M. (2010). Cost allocation in collaborative forest transportation. *European Journal of Operational Research*, 205, 448–458.

Fu, X., Homsombat, W. and Oum, T.H. (2011). Airport-airline vertical relationships, their effects and regulatory policy implications. *Journal of Air Transport Management*, 17, 347–353.

Ghaderi, H., Dullaert, W. and Amstel, W.P.V. (2016). Reducing lead-times and lead-time variance in cooperative distribution networks. *International Journal of Shipping and Transport Logistics*, 8, 51–65.

Golicic, S.L., Foggin, J.H. and Mentzer, J.T. (2003). Relationship magnitude and its role in interorganizational relationship structure. *Journal of Business Logistics*, 24, 57–75.

Gou, Q., Zhang, J., Liang, L., Huang, Z. and Ashley, A. (2014). Horizontal cooperative programmes and cooperative advertising. *International Journal of Production Research*, 52, 691–712.

Groothedde, B., Ruijgrok, C. and Tavasszy, L. (2005). Towards collaborative, intermodal hub networks: a case study in the fast moving consumer goods market. *Transportation Research Part E: Logistics and Transportation Review*, 41, 567–583.

Gruler, A., Fikar, C., Juan, A.A., Hirsch, P. and Contreras-Bolton, C. (2017). Supporting multi-depot and stochastic waste collection management in clustered urban areas via simulation-optimization. *Journal of Simulation*, 11, 11–19.

Guajardo, M. and Rönnqvist, M. (2015). Operations research models for coalition structure in collaborative logistics. *European Journal of Operational Research*, 240, 147–159.

Guajardo, M. and Rönnqvist, M. (2016). A review on cost allocation methods in collaborative transportation. *International Transactions in Operational Research*, 23, 371–392.

Hageback, C. and Segerstedt, A. (2004). The need for co-distribution in rural areas – A study of pajala in Sweden. *International Journal of Production Economics*, 89, 153–163.

Hammant, J. (2011). Collaboration: heard it all before? *Logistics & Transport Focus*, 13, 32–34.

Hernández, S. and Peeta, S. (2014). A carrier collaboration problem for less-than-truckload carriers: characteristics and carrier collaboration model. *Transportmetrica A: Transport Science*, 10, 327–349.

Hsu, P.-H. and Wee, H.M. (2005). Horizontal suppliers coordination with uncertain suppliers deliveries. *International Journal of Operations Research*, 2, 17–30.

IEA (2015). CO_2 emissions from fuel combustion-highlights. *International Energy Agency (IEA)*, Paris http://www.iea.org/co2highlights/co2highlights.pdf, cited 3 July 2017.

Juan, A.A., Faulin, J., Grasman, S.E., Rabe, M. and Figueira, G. (2015). A review of simheuristics: extending metaheuristics to deal with stochastic combinatorial optimization problems. *Operations Research Perspectives*, 2, 62–72.

Juan, A.A., Faulin, J., Perez-Bernabeu, E. and Jozefowiez, N. (2014). Horizontal cooperation in vehicle routing problems with backhauling and environmental criteria. *Procedia-Social and Behavioral Sciences*, 111, 1133–1141.

Karsten, F., Slikker, M. and Van Houtum, G.-J. (2015). Resource pooling and cost allocation among independent service providers. *Operations Research*, 63, 476–488.

Kimms, A. and Kozeletskyi, I. (2016). Core-based cost allocation in the cooperative traveling salesman problem. *European Journal of Operational Research*, 248, 910–916.

Krajewska, M.A. and Kopfer, H. (2006). Collaborating freight forwarding enterprises. *OR Spectrum*, 28, 301–317.

Kuchinke, B.A. and Sickmann, J. (2005). The joint venture terminal 2 at Munich airport and the consequences: a competition economic analysis. *In 4th Conference on Applied Infrastructure Research, Berlin*. volume 8.

Lambert, D.M., Emmelhainz, M.A. and Gardner, J.T. (1996). Developing and implementing supply chain partnerships. *The international Journal of Logistics Management*, 7, 1–18.

Lambert, D.M., Emmelhainz, M.A. and Gardner, J.T. (1999). Building successful logistics partnerships. *Journal of Business Logistics*, 20, 165.

Lau, H.C. and Liu, Q.Z. (2000). Collaborative model and algorithms for supporting real-time distribution logistics systems. *Electronic Notes in Discrete Mathematics*, 4, 17–31.

Lehoux, N., Damours, S. and Langevin, A. (2014). Inter-firm collaborations and supply chain coordination: review of key elements and case study. *Production Planning & Control*, 25, 858–872.

Leitner, R., Meizer, F., Prochazka, M. and Sihn, W. (2011). Structural concepts for horizontal cooperation to increase efficiency in logistics. *CIRP Journal of Manufacturing Science and Technology*, 4, 332–337.

Lera-López, F., Faulin, J., Sánchez, M. and Serrano, A. (2014). Evaluating factors of the willingness to pay to mitigate the environmental effects of freight transportation crossing the pyrenees. *Transportation Research Procedia*, 3, 423–432.

Li, J. (2013). Model and simulation for collaborative VRPSPD. *Journal of Networks*, 8, 331–338.

Li, J., Xiong, N., Park, J.H., Liu, C., Shihua, M. and Cho, S. (2012). Intelligent model design of cluster supply chain with horizontal cooperation. *Journal of Intelligent Manufacturing*, 23, 917–931.

Lier, T. van, Caris, A. and Macharis, C. (2016). Sustainability SI: bundling of outbound reight flows: analysing the potential of internal horizontal collaboration to improve sustainability. *Networks and Spatial Economics*, 16, 277–302.

Limoubpratum, C., Shee, H. and Ahsan, K. (2015). Sustainable distribution through coopetition strategy. *International Journal of Logistics Research and Applications*, 18, 424–441.

Lin, D.-Y. and Ng, K.H. (2012). The impact of collaborative backhaul routing on carbon reduction in the freight industry. *Transportation Research Part D: Transport and Environment*, 17, 626–628.

Liu, R., Jiang, Z., Liu, X. and Chen, F. (2010). Task selection and routing problems in collaborative truckload transportation. *Transportation Research Part E: Logistics and Transportation Review*, 46, 1071–1085.

López-Ramos, F. (2014). Integrating network design and frequency setting in public transportation networks: a survey. *SORT-Statistics and Operations Research Transactions*, 38, 181–214.

Lozano, S., Moreno, P., Adenso-Díaz, B. and Algaba, E. (2013). Cooperative game theory approach to allocating benefits of horizontal cooperation. *European Journal of Operational Research*, 229, 444–452.

Massol, O. and Tchung-Ming, S. (2010). Cooperation among liquefied natural gas suppliers: is rationalization the sole objective? *Energy Economics*, 32, 933–947.

Mentzer, J.T., Foggin, J.H. and Golicic, S.L. (2000). Collaboration: the enablers, impediments, and benefits. *Supply Chain Management Review*, 4, 52–58.

Muñoz-Villamizar, A., Montoya-Torres, J.R. and Vega-Mejía, C.A. (2015). Non-collaborative versus collaborative last-mile delivery in urban systems with stochastic demands. *Procedia CIRP*, 30, 263–268.

Nadarajah, S. and Bookbinder, J.H. (2013). Less-than-truckload carrier collaboration problem: modeling framework and solution approach. *Journal of Heuristics*, 19, 917–942.

Özener, O.Ö. (2014). Developing a collaborative planning framework for sustainable transportation. *Mathematical Problems in Engineering*, 2014.

Özener, O.Ö., Ergun, Ö and Savelsbergh, M. (2011). Lane-exchange mechanisms for truckload carrier collaboration. *Transportation Science*, 45, 1–17.

Pan, S., Ballot, E. and Fontane, F. (2013). The reduction of greenhouse gas emissions from freight transport by pooling supply chains. *International Journal of Production Economics*, 143, 86–94.

Pan, S., Ballot, E., Fontane, F. and Hakimi, D. (2014). Environmental and economic issues arising from the pooling of SMEs supply chains: case study of the food industry in Western France. *Flexible Services and Manufacturing Journal*, 26 , 92–118.

Peetijade, C. and Bangviwat, A. (2012). Empty trucks run reduction in Bangkok area towards sustainable transportation. *International Journal of Trade, Economics and Finance*, 3, 91.

Perez-Bernabeu, E., Juan, A.A., Faulin, J. and Barrios, B.B. (2015). Horizontal cooperation in road transportation: a case illustrating savings in distances and greenhouse gas emissions. *International Transactions in Operational Research*, 22, 585–606.

Petersen, K.J., Handfield, R.B. and Ragatz, G.L. (2005). Supplier integration into new product development: coordinating product, process and supply chain design. *Journal of Operations Management*, 23, 371–388.

Pomponi, F., Fratocchi, L., Tafuri, S.R. and Palumbo, M. (2013). Horizontal collaboration in logistics: a comprehensive framework. *Research in Logistics & Production*, 3.

Pradenas, L., Oportus, B. and Parada, V. (2013). Mitigation of greenhouse gas emissions in vehicle routing problems with backhauling. *Expert Systems with Applications*, 40, 2985–2991.

Quintero-Araujo, C.L., Gruler, A., Juan, A.A., de Armas, J. and Ramalhinho, H. 2017). Using simheuristics to promote horizontal collaboration in stochastic city logistics. *Progress in Artificial Intelligence*, (pp. 1–10).

Ramanathan, U., Bentley, Y. and Pang, G. (2014). The role of collaboration in the UK green supply chains: an exploratory study of the perspectives of suppliers, logistics and retailers. *Journal of Cleaner Production*, 70, 231–241.

Raue, J.S. and Wallenburg, C.M. (2013). Alike or not? Partner similarity and its outcome in horizontal cooperations between logistics service providers. *Logistics Research*, 6, 217–230.

Raue, J.S. and Wieland, A. (2015). The interplay of different types of governance in horizontal cooperations: a view on logistics service providers. *The International Journal of Logistics Management*, 26, 401–423.

Roijakkers, N. and Hagedoorn, J. (2006). Inter-firm r&d partnering in pharmaceutical biotechnology since 1975: trends, patterns, and networks. *Research Policy*, 35, 431–446.

Rokkan, A.I., Heide, J.B. and Wathne, K.H. (2003). Specific investments in marketing relationships: expropriation and bonding effects. *Journal of Marketing Research*, 40, 210–224.

Saeed, N. (2013). Cooperation among freight forwarders: mode choice and intermodal reight transport. *Research in Transportation Economics*, 42, 77–86.

Schmoltzi, C. and Wallenburg, C.M. (2011). Horizontal cooperations between logistics service providers: motives, structure, performance. *International Journal of Physical Distribution & Logistics Management*, 41, 552–575.

Schmoltzi, C. and Wallenburg, C.M. (2012). Operational governance in horizontal cooperations of logistics service providers: performance effects and the moderating role of cooperation complexity. *Journal of Supply Chain Management*, 48, 53–74.

Schulz, S.F. and Blecken, A. (2010). Horizontal cooperation in disaster relief logistics: benefits and impediments. *International Journal of Physical Distribution & Logistics Management*, 40, 636–656.

Soosay, C.A. and Hyland, P. (2015). A decade of supply chain collaboration and directions for future research. *Supply Chain Management: An International Journal*, 20, 613–630.

Soysal, M., Bloemhof-Ruwaard, J.M., Haijema, R. and van der Vorst, J.G. (2016). Modeling a green inventory routing problem for perishable products with horizontal collaboration. *Computers & Operations Research*, in press.

Sprenger, R. and Mönch, L. (2012). A methodology to solve large-scale cooperative transportation planning problems. *European Journal of Operational Research*, 223, 626–636.

Stojanović, D.M. and Aas, B.Ø. (2015). Transport outsourcing and transport collaboration relationship: the risk hedging perspective. *Serbian Journal of Management*, 10, 33–49.

Teirlinck, P. and Spithoven, A. (2013). Research collaboration and R&D outsourcing: Different r&d personnel requirements in smes. *Technovation*, 33, 142–153.

Vanovermeire, C., Sörensen, K., Van Breedam, A., Vannieuwenhuyse, B. and Verstrepen, S. (2014). Horizontal logistics collaboration: decreasing costs through flexibility and an adequate cost allocation strategy. *International Journal of Logistics Research and Applications*, 17, 339–355.

Verdonck, L., Beullens, P., Caris, A., Ramaekers, K. and Janssens, G.K. (2016). Analysis of collaborative savings and cost allocation techniques for the cooperative carrier facility location problem. *Journal of the Operational Research Society*, 67, 853–871.

Verdonck, L., Caris, A., Ramaekers, K. and Janssens, G.K. (2013). Collaborative logistics from the perspective of road transportation companies. *Transport Reviews*, 33, 700–719.

Vornhusen, B., Wang, X. and Kopfer, H. (2014). Vehicle routing under consideration of transhipment in horizontal coalitions of freight carriers. *Procedia CIRP*, 19, 117–122.

Wallenburg, C.M. and Raue, J.S. (2011). Conflict and its governance in horizontal cooperations of logistics service providers. *International Journal of Physical Distribution & Logistics Management*, 41, 385–400.

Wang, H., Unnikrishnan, A., Hernández, S. and Cheu, R. (2014a). Capacitated centralized carrier collaboration multihub location problem: tabu search approach. *Transportation Research Record: Journal of the Transportation Research Board*, (pp. 22–30).

Wang, X. and Kopfer, H. (2014). Collaborative transportation planning of less-than-truckload freight. *OR Spectrum*, 36, 357–380.

Wang, X., Kopfer, H. and Gendreau, M. (2014b). Operational transportation planning of freight forwarding companies in horizontal coalitions. *European Journal of Operational Research*, 237, 1133–1141.

Wei, J., Zhao, J. and Li, Y. (2015). Price and warranty period decisions for complementary products with horizontal firms' cooperation/noncooperation strategies. *Journal of Cleaner Production*, 105, 86–102.

Wilhelm, M.M. (2011). Managing coopetition through horizontal supply chain relations: linking dyadic and network levels of analysis. *Journal of Operations Management*, 29, 663–676.

Xu, N., Yu, C., Zhang, L. and Liu, P. (2009). Profit allocation in collaborative less-than-truckload carrier alliance. In *Automation and Logistics, 2009. ICAL'09. IEEE International Conference on* (pp. 258–263). IEEE.

Yam, R.C. and Chan, C. (2015). Knowledge sharing, commitment and opportunism in new product development. *International Journal of Operations & Production Management*, 35, 1056–1074.

Yengin, D. (2012). Characterizing the Shapley value in fixed-route traveling salesman problems with appointments. *International Journal of Game Theory*, 41, 271–299.

Zeng, F., Chen, Y., Dong, M. C. and Zheng, J. (2015). Understanding distributor opportunism in a horizontal network. *Industrial Marketing Management*, 46, 171–182.

Zhang, J., Gou, Q., Liang, L. and Huang, Z. (2013). Supply chain coordination through cooperative advertising with reference price effect. *Omega*, 41, 345–353.

Zinn, W. and Parasuraman, A. (1997). Scope and intensity of logistics-based strategic alliances: a conceptual classification and managerial implications. *Industrial Marketing Management*, 26, 137–147.