Afrika Statistika

Afrika Statistika
Vol. 15 (2), 2020, pages 2279 - 2293.
DOI: http://dx.doi.org/10.16929/as/2020.2279.159

ISSN 2316-090X

Empirical performance of estimation methods in Beta mixed models with application to ecological data

Bruno Enagnon Lokonon [1], **Freedath Djibril Moussa**[2], **Saliou Diouf**[3,*] **and Romain Glèlè Kakaï**[1]

[1]Laboratoire de Biomathématiques et d'Estimations Forestières, Université d'Abomey-Calavi, 04 B.P. 1525 Cotonou, Bénin
[2] Départment of Mathématiques, Faculté des Sciences et Techniques, Université d'Abomey-Calavi
[3] UFR SEF and LERSTAD, University Gaston Berger of Saint-Louis, Senegal

Received on December 15, 2019; Accepted on June 30, 2020

Abstract. This study uses a Monte Carlo simulation design to assess the performance of Beta and linear mixed models on bounded response variables through comparison of four estimation methods. Four factors affecting the performance of the estimation methods were considered: the number of groups, the number of observations per group, the variance and distribution of the random effects. Our results showed that, for small number of groups (less than 30), the Beta mixed model outperformed the linear mixed model whatever the size of the groups. In the case of a large number of groups (superior or equal to 30), both approaches showed relatively close performance. The results from the simulation study have been illustrated with real life data.

Key words: Beta distribution; continuous proportion; transformations; hierarchical modelling; performance; application.
AMS 2010 Mathematics Subject Classification Objects : 82B80; 81T80; 92D40; 62P12.

[*]Corresponding author Saliou Diouf: saliou.diouf@ugb.edu.sn
Bruno Enagnon Lokonon: brunolokonon@gmail.com
Freedath Djibril-Moussa :freedath.djibrilmoussa@gmail.com
Romain Glèlè Kakaï : glele.romain@gmail.com

Résumé. (Abstract in French) Cette étude utilise une approche empirique pour évaluer les performances des modèles mixtes bêta et linéaire sur des variables réponses bornées en comparant quatre méthodes d'estimation. Quatre facteurs affectant la performance des méthodes d'estimation ont été pris en compte, notamment le nombre de groupes, le nombre d'observations par groupe, la variance et la distribution des effets aléatoires. Les résultats ont montré que, pour un petit nombre de groupes (moins de 30), le modèle mixte Beta surpassait le modèle linéaire mixte quelle que soit la taille des groupes. Dans le cas d'un grand nombre de groupes (supérieur ou égal à 30), les deux approches ont montré des performances relativement proches. Les résultats de l'étude de simulation ont été illustrés par des données réelles.

The authors.

Bruno Enagnon Lokonon, Ph.D., is a Research assistant at the Faculty of Agronomic Sciences, University of Abomey-Calavi, Benin.

Freedath Djibril Moussa, Ph.D., is a Senior lecturer at the Department of Mathematics, Faculty of Sciences and techniques University of Abomey-Calavi, Benin.

Saliou Diouf, Ph.D., is an Associate professor at the *UFR SEFS* of University Gaston Berger of Saint-Louis, and affiliated to *LERSTAD*, Senegal.

Romain Glèlè Kakaï, Ph.D., is a Full professor at the Faculty of Agronomic Sciences, University of Abomey-Calavi, Benin.

1. Introduction

In ecology and evolution, models in which the response variable takes values in the standard unit interval are common. This type of data is referred to as proportional data (Douma and Weedon, 2019). Warton and Hui (2011) showed that more than one-third of ecological studies involve such proportional data. Proportional data are derived from discrete counts, for instance, the count of successes and failures, where the successes are divided by the total counts. Such data are suitably analyzed with logistic regression (Douma and Weedon, 2019). Other proportional data are obtained by dividing a continuous variable by a given value. An example can be seen in Poorter *et al.* (2012), where percentages of biomass were allocated to different plant organs. Contrary to proportions derived from counts, appropriate techniques for analyzing continuous (also called non-count-based or non-binomial) proportions are less established since they violate two important assumptions of standard statistical techniques: normality of error term and constant variance (Douma and Weedon, 2019). A common recommendation in such an instance is to apply a data transformation (Sokahl and Rohlf, 1995;

Quinn and Keough, 2002).

Over the last two decades, Paolino (2001), Kieschnick and McCullough (2003), and Ferrari and Cribari-Neto (2004) proposed Beta models as a method for modeling continuous proportional variables. However, all these studies used beta model in the context of generalized linear models (GLMs) when observations are independent and identically distributed (Bonat *et al.*, 2015a). Nevertheless, many experimental designs in ecology lead to correlated observations. Common examples are multiple observations within an experimental plot or repeated observations of the same experimental unit over time.

In conformity to Beta models used for modeling continuous proportional variables in case of independent observations, Beta mixed models should be used for correlated observations (Bonat *et al.*, 2015a). However, references for Beta mixed models are scarce (Bonat *et al.*, 2015a). Besides, estimation of parameters in Beta mixed models is not straightforward (Bonat *et al.*, 2015a) since the likelihood function is analytically intractable (Verkuilen and Smithson, 2012). Therefore, estimation has to be performed by using approximation methods (Callens and Croux, 2005; Bonat *et al.*, 2015b). Moreover, in the literature one of the most widely used and recommended approaches is to apply a transformation in order to meet the requirements of the statistical model and then perform linear model (Warton and Hui, 2011; Douma and Weedon, 2019). No result has been found in the literature comparing this approach (transformation) to Beta mixed models. Understanding the performance of these models on simulated datasets is important to serve as a guide to applied researchers in choosing appropriate models when working with real-life data. The aim of this study is to fill this gap by comparing these two approaches. Integrated nested Laplace approximation (INLA) and Laplace approximation (LA) were used as Beta mixed model estimators. These estimators were compared to adaptative Gauss-Hermite quadrature (AGHQ) and restricted maximum likelihood (REML) estimators which were respectively applied after a logit and a log transformation of the response variable.

One important assumption of Beta mixed models is that random effects must be normally distributed (Hernandez and Giampaoli, 2018). Indeed, for computational convenience, the random effects are assumed to be normally distributed. However, since they are not observed, the validity of this assumption is difficult to verify (McCulloch and Neuhaus, 2011). A natural concern is related to the impact of misspecification of the random effects distribution on the estimators used. Previous studies investigating the impact of the misspecification focused on linear, logistic, Poisson and Weibull mixed models (Litière *et al.*, 2008; Hernandez *et al.*, 2014; Hernandez and Giampaoli, 2018). But no significant studies have considered Beta mixed models. Therefore, this study also investigated the impact of misspecification of the random effects distribution on the estimators considered.

2. Material and methods

2.1. Parameters estimation in Beta mixed models

Let Y be the dependent variable measured continuously on the standard unit interval, i.e. $0 < Y < 1$ and y an observation. Suppose that $Y \sim \text{Beta}(\alpha, \beta) \in (0, 1)$, where $\alpha, \beta > 0$ are the shape parameters. The probability density function of Y (Ferrari and Cribari-Neto, 2004) is:

$$f(y, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1 - y)^{\beta-1}, \tag{1}$$

where $\Gamma(.)$ is the complete gamma function. The mean and the variance of y are:

$$\text{E}(Y) = \frac{\alpha}{\alpha + \beta}, \tag{2}$$

and

$$\text{var}(Y) = \frac{\text{E}(Y)(1 - \text{E}(Y))}{\alpha + \beta + 1}. \tag{3}$$

For models analysis, it is more convenient to use the mean of the response variable (Ferrari and Cribari-Neto, 2004). In order to obtain a model structure for the mean, a new parametrization is used for the Beta density. Let $\mu = \text{E}(Y)$ and precision parameter $\phi = \alpha+\beta$, which can be inverted to show $\alpha = \phi\mu$ and $\beta = \phi(1-\mu)$. The density of Y can be written in the new parametrisation (Verkuilen and Smithson, 2012) as:

$$f(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu)\Gamma(\phi(1 - \mu))} y^{\phi\mu-1}(1 - y)^{\phi(1-\mu)-1}, \tag{4}$$

where $0 < \mu < 1$ and $\phi > 0$. The dependent variable Y is now Beta distributed as $\text{Beta}(\phi\mu, \phi(1 - \mu))$.

Generalized linear mixed modeling using grouped data structure is a parsimonious strategy to take dependent observations into account (Bonat *et al.*, 2015a). Therefore, let Y_{ij} be the response variable of observations i=1, ..., n (n is number of observations) within group j and $j = 1, \ldots, N$ (N is number of groups). A Beta mixed model can be written as follows (Figueroa *et al.*, 2013; Bonat *et al.*, 2015a):

$$\text{G}(\text{E}(Y_{ij}|b_j)) = X_{ij}\beta + Z_{ij}b_j \tag{5}$$

where G(.) is a link function linking the conditional mean response vector $\text{E}(Y_{ij}|b_j)$ with the linear predictor $\gamma_{ij} = X_{ij}\beta + Z_{ij}b_j$, for which X_{ij} is the design matrix of dimension $n \times p$ corresponding to the vector $\beta = (\beta_1, \cdots, \beta_p)$ of regression coefficients (the fixed effects) and Z_{ij} is the design matrix of dimension $n \times q$ associated with the vector $b_j = (b_1, \cdots, b_q)$ (the random effects). The elements of b_j are assumed to be independent and normally distributed, $b_j \sim \text{N}(0, \Sigma)$.

For the logit link function, the most common used in Beta model (Ferrari and Cribari-Neto, 2004), the i-th component of the Equation 5 is:

$$\ln\{\frac{\mu_{ij}}{1 - \mu_{ij}}\} = \gamma_{ij} = X_{ij}\beta + Z_{ij}b_j, \tag{6}$$

where $\mu_{ij} = \mathrm{E}(Y_{ij}|b_j)$.

The model parameters can be estimated by maximising the marginal likelihood obtained by integrating the joint distribution of (Y_{ij}, b_j) over the random effects. The contribution to the likelihood for the jth group (Bonat *et al.*, 2015a) is:

$$f_j\left(y_j|\beta;\Sigma;\phi\right) = \int \prod_{i=1}^{n} f_{ij}\left(y_{ij}|b_j;\beta;\phi\right) f\left(b_j|\Sigma\right) db_j \tag{7}$$

Assuming independence among the N groups, the full likelihood is given by:

$$\mathrm{L}\left(\beta;\Sigma;\phi\right) = \prod_{j=1}^{N} f_j\left(y_j|\beta;\Sigma;\phi\right) \tag{8}$$

Evaluation of Equation 8 requires computing N integrals. Moreover, the distributions of both random effects b_j and response variable (Y_{ij}) differ. Thus, for Equation 7, it is quite difficult to calculate this integral and maximize it (Casals *et al.*, 2015). As a result, approximation methods have been developed with different degrees of accuracy (Capanu *et al.*, 2013). Two of these methods were considered in this study to solve the integral: the Laplace approximation (LA) and the integrated nested Laplace approximation (INLA). These methods were chosen since each one uses a different approach to compute the integral. The LA uses a Taylor series expansion to approximate the integrand by a function analytically tractable while the INLA uses a Bayesian framework. These methods were selected due to their robustness, their recent improvement in R software and their accessibility for applied researchers (Casals *et al.*, 2015). The choice is also based on the fact that they are widely used in ecology (Lokonon *et al.*, 2019). A short description of these methods is presented below.

2.1.1. Laplace approximation

LA has been designed to approximate the integral in the Equation 7 in the following form:

$$\int_{\mathcal{R}} \exp\{h(x)\}dx \approx (2\pi)^{\frac{1}{2}} \exp\{h(x_{max})\}|\frac{\partial^2 h(x)}{\partial x^2}|_{x=x_{max}}|^{-\frac{1}{2}} \tag{9}$$

where h is a sufficiently smooth (twice differentiable) and integrable function on $\mathbb{R}$ and x_{max} satisfies $h^{'}(x_{max}) = 0$ and $h^{''}(x_{max}) < 0$.
In order to estimate the parameters with LA, one needs to set in Equation 9, $h(x) = ln[f(y_{ij}|u_j)f(u_j]$ following Equation 7. This method is implemented in the R package glmmTMB (Douma and Weedon, 2019).

2.1.2. Integrated nested Laplace approximation

The Bayesian approach is attractive but requires specification of prior distributions (Bonat *et al.*, 2015b). The model specification is completed assuming prior distributions for all following parameters in the model: β, ϕ and Σ. Prior distribution for

these parameters have been defined in this study following Bonat *et al.* (2015b). The posterior density is given by:

$$\pi\left(\beta,\Sigma,\phi\right) \propto \pi(\beta)\pi(\phi)\prod_{i=1}^{n} f_{ij}\left(y_{ij}|\Sigma;\beta;\phi\right) \tag{10}$$

where, i=1,...,n (number of observations).
INLA provides a good approximation while reducing computational costs substantially comparative to Markov chain Monte Carlo (Bonat *et al.*, 2015b; Rue *et al.*, 2017). In R, it is implemented in the function *inla* from the package *R-inla* (Rue *et al.*, 2017).

Apart from these two approximation methods, two transformation approaches were used. Firstly, logit transformation was apply on the response variable and AGHQ was used to estimate the parameters through glmer function from the package lme4. Secondly, log transformation was applied on the response variable and REML was used to estimate the parameters through lmer function from the package lme4.

2.2. Simulation study design

The most used form of the model (Equation 6) in simulation study is the random intercept model (Hernandez *et al.*, 2014; Hernandez and Giampaoli, 2018) with the following notation:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2ij} + b_j \tag{11}$$

Considering the model (Equation 11), five factors affecting its accuracy varied in the study: number of groups, number of observations per group, variance and distribution of the random effects and different estimation methods used. The random effect b_j has zero mean and variance σ^2. The fixed effects were set from previous studies (Bauer and Sterba, 2011; Hernandez *et al.*, 2014; Ali *et al.*, 2016; Hernandez and Giampaoli, 2018): β_0=1, β_1=2, β_2=1 and ϕ=1 to keep the model simple. The between-cluster (x_1) and the within-cluster (x_2) were generated from standard Normal distribution $N(0,1)$. The number of clusters and the number of observations per cluster were set respectively as N= 5, 10, 30, 50 and n= 5, 10, 30, 50 in order to obtain various sample sizes from 25 to 2500. The variances of the random effect b_j were set at σ^2=0.4, 0.5, 1.1, 2, 3.4. Variances greater than 3.4 were not considered because they caused larger values of the random intercept (Hernandez and Giampaoli, 2018).

The algorithm of the simulation is described as follows:

Step 1 : Set values for the regression coefficients β_0, β_1, β_2, for variance σ^2 and for ϕ;

Step 2 : Generate the covariates x_1 and x_2 from the Standard Normal distributions and the random effect from Normal distribution with mean=0 and variance=σ^2;

Step 3 : Set the coefficients and obtain logit such that:
$\texttt{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2ij} + b_j$;
Step 4 : Calculate the predicted means μ_{ij} such that:
$\mu_{ij} = \texttt{invlogit}(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2ij} + b_j)$;
Step 5 : Obtain the continuous proportion y_{ij} such that:
$y_{ij} = \texttt{rbeta}(n_{ij}, \mu_{ij}\phi, (1-\mu_{ij})\phi)$,
where the sample size n_{ij} is the combination of clusters N= 5, 10, 30, 50 and cluster sizes n= 5, 10, 30, 50. n_{ij} varies from 25 to 2500 ;
Step 6 : For each combination of n_{ij} and σ^2, run the model (Equation 11) using the following estimation methods *a)* LA; *b)* INLA; *c)* AGHQ after a logit transformation of y_{ij}; *d)* REML after a log transformation of y_{ij};
Step 7 : Repeat step 6 S times (S=500).

For the misspecification study, the simulation design is the same but the random effects were generated from five true distributions: uniform, exponential and log–normal, log-gamma, and symmetric mixture of two normal densities that were defined following Hernandez and Giampaoli (2018). However, the model (Equation 11) was fitted by assuming normality for the random effects.

2.3. Comparison criteria

For each simulation setting and estimation method, mean bias (B) and mean squared error (MSE) were calculated for the fixed effects and the random effects. B and MSE were computed as follows:

$$\text{B} = \frac{1}{S}\sum_{j=1}^{S}\left(\beta - \hat{\beta}_j\right); \qquad \text{MSE} = \frac{1}{S}\sum_{j=1}^{S}\left(\beta - \hat{\beta}_j\right)^2 \tag{12}$$

where $\hat{\beta}_j$ is the estimated parameter, β is the true value and $j = 1, \ldots, S$, S is the number of the simulations (S=500). Moreover, the computational times was recorded with R function system.time as well as the convergence rate. At each iteration, the estimation method showing the low values of B and MSE is the best. B and MSE were plotted and analyzed.

For the misspecification study, a vector of the parameters used in the simulation is defined by $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2)^T$ and a relative distance (RD) is computed as follows (Hernandez *et al.*, 2014):

$$\text{RD} = \frac{||\hat{\theta} - \theta||}{||\theta||} \tag{13}$$

where $\hat{\theta}$ is the estimated parameter vector, θ the true parameter vector and ($||\theta|| = \sqrt{\beta_0^2 + \beta_1^2 + \beta_2^2 + \sigma^4}$). RD is used to quantify the impact of the misspecification on the estimates. The smaller the values of RD, the lower is the impact and better is the estimation method used.

2.4. Illustration with real data

The data are extracted from the study of Douma and Weedon (2019) but originally collected by Andrew and Underwood (1993) who experimentally manipulated the density of the sea urchins *Centrosthepanus rodgersii* to investigate the effects of its grazing on the colonization of filamentous algae. Four different treatments were imposed: an undisturbed control, complete removal of the sea urchins, and two levels of partial removal of sea urchins such that 33 % or 66 % of the original sea urchins density remained. Treatments were replicated in four randomly located patches (3-4 m^2). Algae colonization was measured by percentage cover in five randomly located 0.25 m^2 quadrats within the patches. There were therefore 5 measurements in total within 16 patches, equally divided among four different treatments. Beta and linear mixed effect models were applied to the data and the parameters were estimated with the same estimation methods used in the simulation study. The function check_model in the package performance was used to check the distribution of the random effect (patch). Furthermore, various comparison criteria such as Akaike Information Criterion (AIC, W-AIC for INLA), Bayesian Information Criterion (BIC), Mean Square Error (MSE), deviance and the computational time were used to compare the estimation methods.

3. Results

3.1. Performance of the estimation methods in the simulation studies

3.1.1. Mean bias

Figure 1 represents boxplots of absolute values of the mean bias of the four estimation methods for fixed and random effects in function of the variance of the random effects. Overall, the mean bias is constant for each method when the variance increases considering the fixed effects. For the random effects, the mean bias of the estimation methods increases with the variance for REML and AGHQ. REML showed the lowest median values of mean bias in the case of fixed effects. LA and INLA showed the lowest median values of mean bias for all values of variances. Considering the mean bias values, REML outperforms the other methods for estimating fixed effects while LA and INLA outperform the other methods for estimating random effects. Moreover, AGHQ is the worse estimation method for all values of variances and for all effects.

3.1.2. Mean square error

Figure 2 shows the trend of MSE of the four estimation methods in function of N (number of observations per group) for the combination of all effects and n. This figure reveals the impact of the simulation factors on the performance of the four estimation methods. Overall, when N increases from 5 to 30, the MSE value of the four methods decreases. LA and INLA present the lowest MSE values and then outperform AGHQ and REML in this interval of N. For N greater or equal to 30, the MSE value is relatively close for all methods in the case of within effect. For the random and between effects, the MSE value of LA, INLA and and REML are

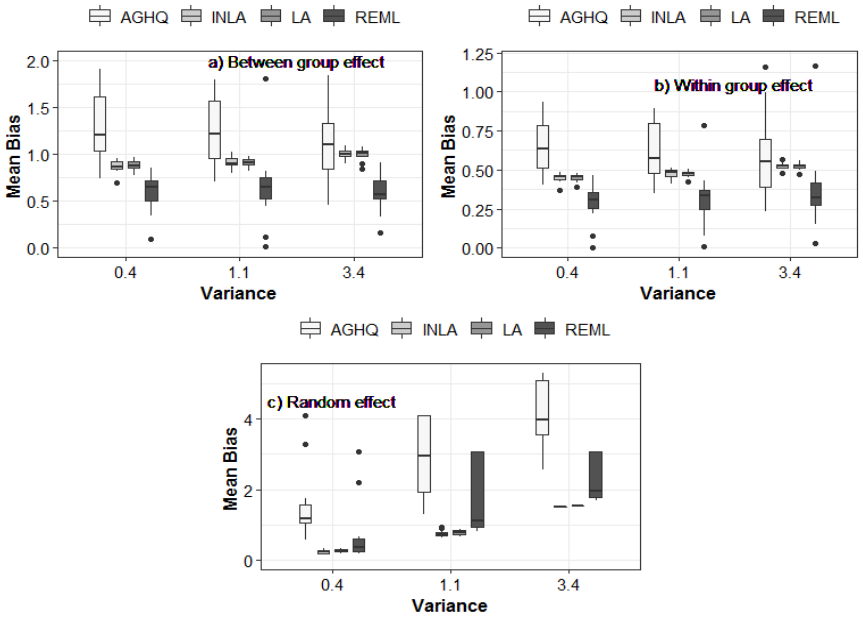

Fig. 1. Boxplots of mean bias for LA, INLA, AGHQ and REML in function of variance of the random effect

relatively close for N greater or equal to 30 while AGHQ shows the larger values of MSE.

3.1.3. Effect of misspecification of the random effects distribution on the performance of the estimation methods

In figure 3, the relative distance between the estimated parameter vector and the true parameter vector for the estimation methods is presented according to the distribution of the random effects. In all situations, LA and INLA showed the lowest and very close relative distance. For AGHQ and REML, the RD decreases as N increases from 5 to 30 and remains relatively constant for N greater or equal to 30. It appears that LA and INLA outperform the other methods in all situations of misspecification.

3.1.4. convergence rate and Computing time

Convergence rate and computation time of the four methods are presented in Table 1. Overall, the convergence rates were high for the four methods, particularly for INLA and LA showing 100 %, except for smaller values of n and N. As far as computational time is concerned, REML and AGHQ method requires less times and

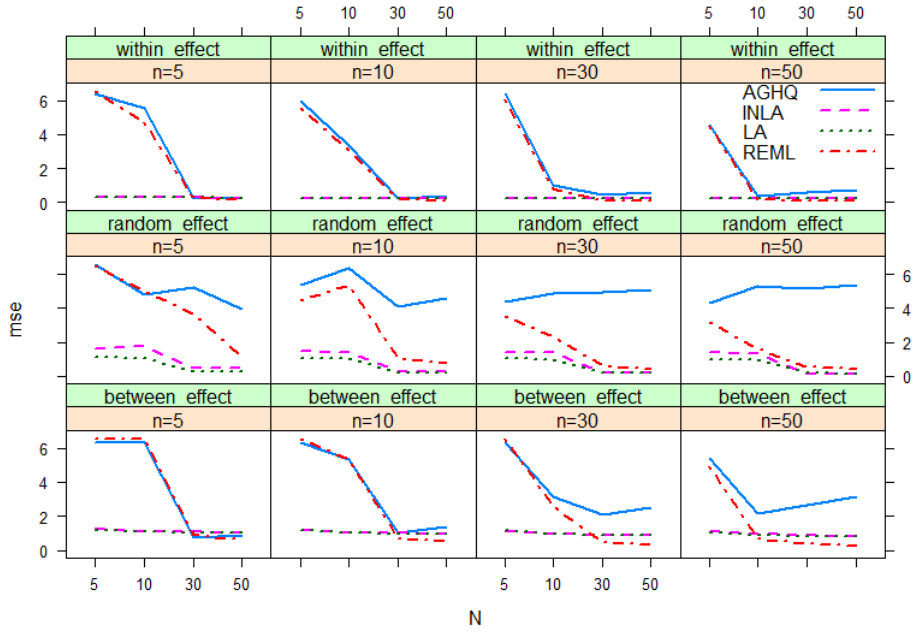

Fig. 2. Plot of MSE of LA, INLA, AGHQ and REML in function of N for the combination of effects (fixed and random) and n

REML is by far the fastest method. INLA and LA require more times, whereas the INLA is by far the slowest method.

Table 1. Convergence rate in percentage (%) and computational time in second (in bracket) of the estimation methods with the package used in bracket

Samples	LA (glmmTMB)	INLA (INLA)	AGHQ (lme4)	REML (lme4)
n=5 N=5	97.67(2.10)	96.67(8.87)	77.50(0.35)	99.83(0.16)
n=30 N=5	100(5.26)	100(17.26)	95(0.39)	99.67(0.17)
n=50 N=5	100(7.57)	100(24.51)	97.67(0.36)	99.33(0.18)
n=5 N=30	100(5.00)	100(16.17)	99.17(0.36)	99.83(0.17)
n=30 N=30	100(23.70)	100(73.57)	99.83(0.43)	99.83(0.22)
n=50 N=30	100(40.42)	100(108.25)	100(0.48)	99.67(0.25)
n=5 N=50	100(7.54)	100(25.40)	99.33(0.37)	99.83(0.18)
n=30 N=50	100(39.95)	100(108.08)	100(0.52)	99.83(0.26)
n=50 N=50	100(66.00)	100(183.94)	100(0.59)	99.83(0.32)

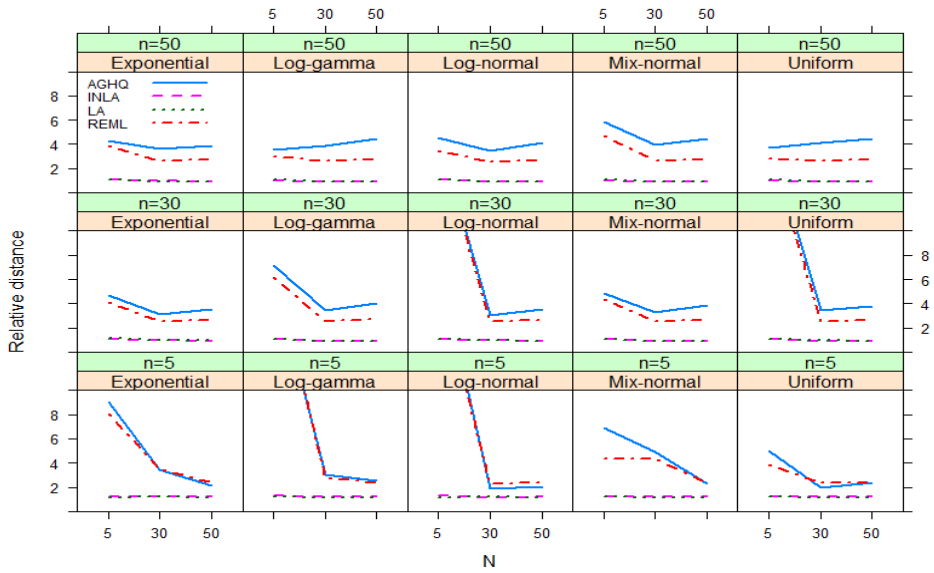

Fig. 3. Relative distance between $\hat{\theta}$ and θ for the estimation methods in function of N

3.2. Analysis of the real data

Checking the random effects distribution (Figure 4), it was noted that they are normally distributed meaning that the estimation methods can be used without care about misspecification. The residuals behaviour was also shown on the figure. The data have a hierarchical structure with N=16 groups (patches) and n=5 observations per group (measurements per patch). According to the findings of the simulation study (Figures 2 and 3) and these characteristics of the data, it is expected that LA and INLA outperform the other methods.
Table 2 shows the results from the four methods. Model selection based on AIC, deviance, BIC and MSE (goodness of fit) clearly favours LA and INLA since these methods present the smaller values of the comparison criteria. The computation time, as expected is smaller for AGHQ and REML. These results are in agreement with the outcome from the simulation study.

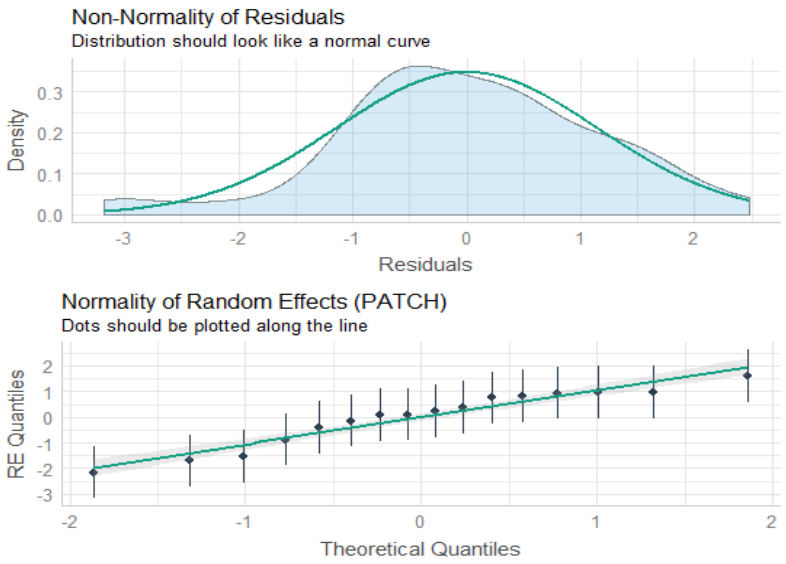

Fig. 4. Random effect distribution and residuals behaviour

Table 2. Estimates of the models explaining proportion algae cover within patches by treatment, goodness of fit and timing

Parameters	LA (glmmTMB)	INLA (INLA)	AGHQ (lme4)	REML (lme4)
Intercept	-2.50	-2.48	-4.57	-4.59
33 % removal	0.86	0.85	1.77	1.51
66 % removal	0.91	0.90	2.06	1.77
removal	1.94	1.92	3.57	2.97
σ^2	0.42	0.54	2.50	1.73
ϕ	3.77	3.90	-	-
Goodness of fit				
AIC	-167.55	-182.17	322.20	296.94
Deviance	-179.60	-194.91	310.20	284.94
BIC	-153.26	-	336.49	311.23
MSE	0.029	-	1.80	1.30
Timing				
Time (sec.)	0.44	7.10	0.25	0.06

4. Discussion

This study analyzes continuous proportional data using four estimation methods. The estimation of Beta mixed models involves solving an intractable integral when evaluating the likelihood function. Two numerical approaches to solve

such integral are applied, Laplace approximation and integrated nested Laplace approximation. Additionally, logit and log transformation were applied on the response variable and two other methods (AGHQ and REML) were used. Several studies recommend the transformation of continuous proportional data in order to apply a linear model. Methods for bias-reduction and bias-correction are an active research area in applied statistics, especially in Beta regression (Grün *et al.*, 2012; Douma and Weedon, 2019), hence this study aimed to contribute to such knowledge base.

The overall analysis based on the mean bias revealed the superiority of REML for estimating fixed effects. In contrast, LA and INLA outperformed the other methods for estimating random effects. With regards to AGHQ, the mean bias is generally greater than those of the other methods. In terms of accuracy of the estimation methods, the mean square error (MSE) values of LA and INLA are largely lower than those of REML and AGHQ for N less than 30. LA and INLA are more accurate than REML and AGHQ for N less than 30. However, when N is greater or equal to 30, the MSE is relatively constant and very close for all methods except AGHQ. Our results confirm previous studies showing that samples with a least N=30 groups are necessary to obtain lower biased estimates in hierarchical modeling (Dedrick *et al.*, 2009; Maas and Hox, 2005). Furthermore, our results imply that the log transformation applied on the response variable followed by parameter estimation using REML is applicable for N greater than 30 for Normal random effects. This result improves the statement of Figueroa *et al.* (2013) affirming that linear models are not appropriately applicable when the response variable has a doubly bounded interval.

Concerning the misspecification, several studies have showed that estimation methods are severely affected in situations where random effects distribution are misspecified (Agresti *et al.*, 2004; Litière *et al.*, 2008; Hernandez and Giampaoli, 2018). Our results are in accordance with these findings. In practice, we recommend checking of the random effects distribution using a diagnostic test (Drikvandi *et al.*, 2017) and then, if the random effect distribution is not normal, use LA or INLA to estimate the model parameters. Finally, we apply the estimation methods used in the simulation study to percent cover data. The two studies lead to the same results.

5. Concluding remarks

The purpose of this study was to compare the properties of linear mixed models and Beta mixed models when the response variable is bounded between 0 and 1. Several previous studies have concluded that response variables bounded between 0 and 1 should not be analyzed with linear models after transformation. This study shows that for hierarchical data, when the number of groups is smaller than 30, Beta mixed model is more accurate than a linear mixed model even when data are transformed. However, for groups greater or equal to 30, linear mixed model could be applied after a log transformation if the random effects are normally distributed.

We recommend the use of Beta mixed model when a diagnostic test shows that the random effect does not follow a Normal distribution.

Acknowledgment

This work was financially supported by African Centre of Excellence in Mathematics and Applications (CEA-SMA). We thank this institution and its donors. We also thank EMS-Simons for Africa for sponsoring the visit research carried out by the first author at UFR SEFS of the Université Gaston Berger - Saint Louis (Sénégal) during which this paper was written.

The authors are also grateful to the anonymous referee and to the associated editor for their helpful suggestions that lead to an improved paper.

References

Agresti A., Caffo, B. and Ohman-Strickland P. 2004. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Comput. Stat. Data Anal.* 47, 639-653

Ali S., Ali A., Khan S. A. and Hussain S. 2016. Sufficient Sample Size and Power in Multilevel Ordinal Logistic Regression Models. *Comput. Math. Method. M.*. http://dx.doi.org/10.1155/2016/7329158

Andrew N., and Underwood A. 1993. Density-dependent foraging in the sea urchin centrostephanus rodgersii on shallow subtidal reefs in new south wales, australia. *Mar Ecol Press Series* 99, 89-98.

Bauer D. J. and Sterba, S. K. 2011. Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychol. Methods* 16, 373–390

Bonat W.H., Ribeiro Jr, P.J. and Zeviani W.M. 2015a. Likelihood analysis for a class of beta mixed models. *J Appl. Stat.* 42, 252-266.

Bonat W.H., Ribeiro Jr, P.J. and Shimakura S.E. 2015b. Bayesian analysis for a class of beta mixed models. *Chil. J. Stat.* 6, 3-13.

Callens M. and Croux, C. 2005. Performance of likelihood-based estimation methods for multilevel binary regression models. *J. Stat. Comput. Sim.* 75, 1003-1017

Capanu M., Gönen M. and Begg C. B. 2013. An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Stat. Med.* 32, 4550-4566

Casals M., Langohr K., Carrasco J. L. and Rönnegård L. 2015. Parameter estimation of Poisson generalized linear mixed models based on three different statistical principles: a simulation study. *Sort* 39 (2), 281-308

Dedrick R. F., Ferron J. M., Hess M. R., Hogarty K. Y., Kromrey J. D. and Lang, T. R., ... Lee, R. S. 2009. Multilevel modeling: A review of methodological issues and applications. *Rev. Educ. Res.* 79 (1), 69-102.

Douma J. C. and Weedon J. T. 2009. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods Ecol Evol.*, 00:1–19.

Drikvandi R.; Verbeke G. and Molenberghs G. 2017. Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics*, 73, 63–71.

Ferrari S. and Cribari-Neto F. 2004. Beta regression for modelling rates and proportions. *J. Appl. Stat.* 31(7), 799-815.

Figueroa-Zuniga J. I., Arellano-Valle R. B. and Ferrari S. L. 2013. Mixed beta regression: A Bayesian perspective. *Comput Stat Data Anal.* 61(0), 137-147.

Grün B., Kosmidis I. and Zeileis A. 2012. Extended beta regression in R: shaken, stirred, mixed, and partitioned. *J. Stat. Softw.* 48(11), 1-25.

Hernandez F., Usuga O. and Giampaoli V. 2014. A misspecification simulation study in Poisson mixed model, In: Kneib T, Sobotka F, Fahrenholz J and Irmer H, *Proceedings of the 29th International Workshop on Statistical Modelling, (207-212)* Gottingen, Germany, 14-18 July 2014.

Hernandez F. and Giampaoli V. 2018. The Impact of Misspecified Random Effect Distribution in a Weibull Regression Mixed Model. *stats* 1, 48-76.

Kieschnick R. and McCullough B. D., 2003. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Stat. Model.* 3(3), 193-213.

Litière S., Alonso A. and Molenberghs G. 2008. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Stat. Med.* 27, 3125-3144.

Lokonon B E., Beh Mba R., Gbeha M. and Glèlè Kakaï R. 2019. Parameters Estimation Methods in Generalized Linear Mixed Models Applied in Ecology: A Critical Review. *International Journal of Engineering and Future Technology* 16(3), 12-27.

Maas C. J. M. and Hox J. J. 2005. Sufficient Sample Sizes for Multilevel Modeling. Methodology. Europ. *J. Res. Meth. Behav. Soc. Sc.* 1, 85-91.

McCulloch C.E. and Neuhaus J.M. 2011. Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Stat. Sci.* 26, 388-402.

Paolino P. 2001. Maximum likelihood estimation of models with beta-distributed dependent variables. *Pol. Anal.* 9(4), 325-346.

Poorter H., Niklas K. J., Reich P. B., Oleksyn J., Poot P., and Mommer L. 2012. Biomass allocation to leaves, stems and roots: Meta-analyses of interspecific variation and environmental control. *New. Phytol.* 193(1), 30–50.

Quinn G., and Keough M. 2002. Experimental design and data analysis for biologists. Cambridge: Cambridge University Press.

Rue H., Riebler A., Sorbye S. H., Illian J. B., Simpson D. P. and Lindgren, F. K. 2017. Bayesian computing with INLA: A review. *Ann. Rev. Stat. Appl.* 4, 395-421.

Sokahl R. and Rohlf F. 1995. Biometry (3rd ed.). New York: W. H. Freeman.

Verkuilen J. and Smithson M. 2012. Mixed and mixture regression models for continuous bounded responses using the beta distribution. *J. Educ. Behav. Stat.* 37(1), 82-113.

Warton D. I. and Hui F. K. C. 2011. The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, 92(1), 3-10.