Multivariate dynamic model for ordinal outcomes

Florence Chaubert, Frédéric Mortier, Laurent Saint André

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Multivariate dynamic model for ordinal outcomes

Florence Chaubert — Frédéric Mortier — Laurent Saint-André

N° 5999

Septembre 2006

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--5999--FR+ENG

INRIA
SOPHIA ANTIPOLIS

Multivariate dynamic model for ordinal outcomes

Florence Chaubert , Frédéric Mortier * , Laurent Saint-André

Thème BIO — Systèmes biologiques
Projet VirtualPlants

Rapport de recherche n° 5999 — Septembre 2006 — 24 pages

Abstract: Individual or stand-level biomass is not easy to measure. The current methods employed, based on cutting down a representative sample of plantations, make it possible to assess the biomasses for various compartments (bark, dead branches, leaves, ...). However, this felling makes individual longitudinal follow-up impossible. In this context, we propose a method to evaluate individual biomasses by compartments when these biomasses are taken as ordinals. Biomass is measured visually and observations are therefore not destructive. The technique is based on a probit model redefined in terms of latent variables. A generalization of the univariate case to the multivariate case is then natural and takes into account the dependency between compartment biomasses. These models are then extended to the longitudinal case by developing a Dynamic Multivariate Ordinal Probit Model. The performance of the MCMC algorithm used for the estimation is illustrated by means of simulations built from known biomass models. The quality of the estimates and the impact of certain parameters, are then discussed.

Key-words: Biomass, multivariate longitudinal data, ordinal variable, latent variable, probit link, MCMC.

* E-mail: fmortier@cirad.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 —Télécopie : +33 4 92 38 77 65

Modèle multivarié dynamique pour des réponses ordinales

Résumé : La biomasse d'un individu ou d'un peuplement est difficilement mesurable. Les méthodes de mesures actuelles, basées sur l'abattage d'un échantillon représentatif du peuplement, permettent d'évaluer les biomasses pour différents compartiments (tronc, branches mortes, feuilles). Cependant, cet abattage rend impossible un suivi longitudinal des individus. Dans ce contexte, nous proposons une méthode, permettant d'estimer les biomasses par compartiment d'un individu lorsque celles-ci sont classées de façon discrètes ordinales. Les mesures seront obtenues de visu et par conséquent non déstructives. Fondée sur les Modèles Probit redéfinis en terme de variables latentes gaussiennes, une généralisation du cas univarié au cas multivarié à un temps donné est naturelle. Ces modèles seront étendus au cas longitudinal en développant un Modèle Probit Multivarié Ordinal Dynamique. Les performances des algorithmes MCMC seront illustrées sur des simulations construites à partir de modèles de biomasse. On discutera de la qualité des estimations et de l'impact de certains paramètres.

Mots-clés : Biomasse, données longitudinales multivariées, variable ordinale, variable latente, lien probit, MCMC.

1 Introduction

The capacity to predict longitudinal quantitative traits is of major importance in certain fields such as economies [34], genetic breeding [2], carbon sequestration [36] and psychometrics or educational sciences [26]. Special attention has been paid to univariate or multivariate quantitative cases [15]. Models, often based on generalized linear mixed models (GLMM) [28], have recently been proposed in the context of univariate or bivariate binary variables [38, 33]. Multivariate polytomous traits are now regularly available. However, their time-course remains difficult to model and predict. This paper deals with the time prediction of random ordinal variables taking account of the dependencies between variables and time correlations.

In the univariate case, autocorrelations are treated in a variety of ways. Zhang [39] used a random intercept depending on time. Others, such as Carlin et al. [8], and Fahrmeir [17], put forward the use of dynamic or random regression models to take time dependencies into account. These models, which were first introduced for the analysis of times series, form a very large class of models. They are conventionally constructed in a hierarchical manner [4]: at the first level, given random parameters, observations are assumed to be independent and to follow a given distribution; at the second level, parameters are modeled by a discrete or continuous time process.

The construction of an appropriate dependency structure between polytomous variables is still one of the major difficulties encountered when generalizing the univariate to the multivariate case. Indeed, there is no standard definition for the correlation between polytomous variables. In the univariate case, the probit model is now well known and frequently used to model ordinal variables. Based on the inverse Gaussian distribution, these can be redefined in terms of latent Gaussian variables as proposed for multivariate binary variables by Ashford and Swoden [3]. A generalization to the multivariate ordinal case has been proposed by Daganzo [13]. This approach has been widely used in different fields such as medicine [25] or for generalization of Euclidian distances [5, 29]. However, other links are frequently used such as the logit link. O'Brien and Dunson [30] have defined a multivariate logistic model based on an approximation by a multivariate Student distribution. But, as explained by Joe [22], an explicit form of correlation structure does not exist in this case.

Consequently, in this paper, we propose to take account of the dependency between ordinal variables and autocorrelations, using multivariate probit and dynamic models. We developed Dynamic Multivariate Probit Ordinal Model (DMPOM) as part of the family of Generalized Linear Multivariate Mixed Models (GLMMM).

But, this model remains difficult to estimate and the likelihood evaluation is intractable when more than 4 polytomous variables are used or when the structure of the correlation is too complex [24]. In the Multivariate Probit Ordinal Model (MPOM), De Leon [14] proposed the use of a pseudo-likelihood approach based on a pairwise likelihood. The pseudo-maximum likelihood has well-known asymptotic properties [12] but cannot be used in the longitudinal case because of time dependencies within statistical units. As proposed by Chib and Greenberg [10] in the case of a MPOM, we chose to use Monte Carlo Markov

Chains (MCMC). The proposed algorithm is based on a mixture of the Metropolis-Hastings algorithm [21] and Gibbs sampling [18].

This work was prompted by the problem of performing biomass estimations in forest ecosystems. The standard procedure, which has been applied in numerous studies [6, 31, 32, 37], consisted of (i) a stand inventory (all trees were measured in circumference at breast height, c1.30), (ii) destructive sampling of trees distributed over the entire spectrum of inventoried c1.30 classes, (iii) calculation of weighted allometric relationships fitted for each individual compartment, and (iv) quantification of the stand biomass and nutrient content on a per ha basis by applying the fitted equations to the stand inventory. Destructive sampling is a major impediment in such methods. Indeed, it has been shown recently that the allometric relationship between tree size (given by c1.30 and height) and the biomass of most tree compartments (bark, living branches, dead branches, leaves) varied with stand age [36]. As a consequence, destroying trees each time a sample is taken may introduce an error in identifying and quantifying this age effect. DMPOM was seen as a good opportunity to overcome this problem because stand inventories could include, a visual assessment and classification of the standing biomass into variables as the same time as c1.30 and height measurements. For example, leaf biomass could be visually evaluated tree by tree and classified into three to four classes. This classification would be used together with c130 and Height into DMPOM so as to get the actual biomass value. This is particularly important for compartments that are traditionnaly difficult to assess with only the tree quantitative traits (diameter and height): leaves, living branches and dead branches. Furthermore, this new method can be used to collect longitudinal data for a large number of trees, something that was impossible with the destructive sampling method. As trees are divided into several compartments, this application falls perfectly within the scope of DMPOM: (i) the multivariate ordinal data consist of the biomass of each compartment which are then cross-correlated (for example, living branches are negatively correlated to dead branches and positively correlated to leaf biomass), (ii) these multivariate ordinal data change with stand age as the tree grows.

The paper is organized as follows. In section 2 we present the DMPOM used to model longitudinal multivariate ordinal variables. The relationship between covariates and observed variables in terms of latent Gaussian variables is first established at a given time. We describe the transition model specifying the evolution of the regression and the threshold parameters in time. Section 3 presents the *posterior* analysis. In section 4 we conduct simulations to assess DMPOM performance. The simulations are built from known biomass models evaluated of an eucalyptus plantation in Pointe-Noire (Congo).

2 Dynamic multivariate ordinal probit model

2.1 Multivariate probit ordinal model

Let $Y_i^t = (Y_{i1}^t, \ldots, Y_{iJ}^t)'$, $i = 1, \ldots, n_t$ be n_t independent Gaussian vectors of dimension $J \times 1$ observed on $t = 1, \ldots, T$ times (the number of observations may vary over time), such that

$$Y_i^t \sim \mathcal{N}_J\left(\mu^t + X_i^t \beta^t, R\right), \tag{1}$$

where $X_i^t = diag(X_{i1}^t, \ldots, X_{iJ}^t)$ is the $J \times PJ$ matrix of P covariates associated with individual i at times $t = 1 \ldots, T$, *diag* the block diagonal matrix, μ^t a time varying intercept, $(\beta^t)_{t\in 1,\ldots,T}$ an unknown vector of PJ parameters also varying over time and R an unknown correlation matrix assumed to be identical for all individuals. For identifiability reasons, we assume that R is a correlation matrix and not a covariance matrix [10]. Now let us assume that Y_j^t is not directly observed but measured via an ordinal variable Z_j^t with $c_j, j = 1, \ldots, J$ modalities defined as follows

$$Z_j^t = z_j \quad \Leftrightarrow \alpha_{j,z_j-1}^t \leq Y_j^t < \alpha_{j,z_j}^t; \;\; j = 1, \ldots, J; \;\text{ and }\; t = 1, \ldots, T,$$

where α_{j,z_j}^t are unknown thresholds with $-\infty = \alpha_{j,0}^t < \ldots < \alpha_{j,z_j}^t < \ldots < \alpha_{j,c_j}^t = +\infty$ which are different for each time. We assume that the number of modalities remains constant over time. For a given time t, the Multivariate Probit Ordinal Model (MPOM) is:

$$\mathrm{P}\left(Z^t = z \mid \mu^t, \beta^t, R, \alpha^t\right) = \mathrm{P}\left(Y^t \in A^t \mid \mu^t, \beta^t, R, \alpha^t\right), \tag{2}$$

with $A^t \subseteq \Re^J$ defined as follows $A^t = \left[\alpha_{1,z_1-1}^t, \alpha_{1,z_1}^t\right) \times \ldots \times \left[\alpha_{J,z_J-1}^t, \alpha_{J,z_J}^t\right)$.

2.2 Transition model

In this section, our objective is to take time dependency into account. In our model, intercept μ^t, regression parameters β^t and thresholds α^t are now assumed to be discrete time random processes. As the intercept and thresholds statutes are similar, we treat them globally.

A time-dependent structure of unknown parameters $(\beta^t;\ t = 1, \ldots, T)$ can be modeled by a general multivariate autoregressive model (MAR). MAR is a generalization of the state space approach for a time series [8] and for a univariate categorical time series [7]. A time autoregressive model of order 1 (AR(1)) was chosen to model the dynamic of the unknown time-dependent parameters β^t:

$$\beta^t = F\beta^{t-1} + v^t, \;\; v^t \sim \mathcal{N}_{PJ}(0, \Sigma_\beta), \tag{3}$$

where F is a $PJ \times PJ$ autocorrelation matrix and Σ_β a $PJ \times PJ$ covariance matrix. In our study, we first assume that the regression parameters β_j^t and $\beta_{j'}^t$ are independent

for $j \neq j' \in \{1, \ldots, J\}$ and that β_{jp}^t and $\beta_{jp'}^t$ are independent for $p \neq p' \in \{1, \ldots, P\}$. Therefore, the model for regression parameters is:

$$\beta_j^t = f_j \beta_j^{t-1} + v_j^t, \;\; v_j^t \sim \mathcal{N}_P(0, \Sigma_{\beta_j}), \;\; \Sigma_{\beta_j} = diag(\sigma_{jp}^2)_{p=1,\ldots,P}, \;\; j = 1, \ldots, J;$$

$$|f_{jp}| < 1, \;\text{ and }\; f_j = diag(f_{jp})_{p=1,\ldots,P} \;\; .$$

Different parametrizations of the Probit models are available for identifiability reasons. Either the intercept is assume to be zero and all thresholds are treated as unknown quantities, or the thresholds $\alpha_{j,1}$ for all j are null and the intercept is not zero. In order to generalize a Probit model to a time-dependent Probit model, the second parametrization seems to be simpler than the first. Indeed, unlike the regression parameters β^t, no multivariate autoregressive model for α's could be conceived to comply with the order of the thresholds. In the same manner as [23], we propose the following transition model:

$$\begin{aligned} \mu^t &= (\mu_1^t, \ldots, \mu_J^t)' = \gamma_\mu \mu^{t-1} + \varepsilon_\mu; \;\; \varepsilon_\mu \sim \mathcal{N}_J(0, \Sigma_\mu) \\ \alpha_{j,k}^t &= -\mu_j^t + \alpha_{j,k}; \;\; k = 2, \ldots, c_j - 1; \;\; j = 1, \ldots, J \;, \end{aligned} \tag{4}$$

where γ_μ is an unknown parameter. Thus, to model the dynamics of the time-dependent intercept μ^t, and then the time-dependent thresholds α^t, we used either a general multivariate autoregressive or an independent time varying coefficient. Finally, the dynamic multivariate ordinal probit model is given by the following definition:

Definition 1 *The dynamic multivariate ordinal probit model (DMOPM) is defined by the latent equation 1:*

$$Y_i^t \sim \mathcal{N}_J\left(\mu^t + X_i^t \beta^t, R\right), \;\; i = 1, \ldots, n_t, \;\; t = 1, \ldots, T;$$

the measure equation 2 :

$$P\left(Z_i^t = z \mid \mu^t, \beta^t, R, \alpha^t\right) = P\left(Y_i^t \in A^t \mid \mu^t, \beta^t, R, \alpha^t\right)$$

and by the transition equations 3 and 4 given by:

$$\begin{aligned} \eta'^t &= \left[\left(\mu^t\right)', \left(\beta^t\right)'\right]' = F_\eta \eta'^{t-1} + \varepsilon; \;\; \varepsilon \sim \mathcal{N}_{J(P+1)}(0, \Sigma_\eta), \\ \Sigma_\eta &= diag\left(\sigma_{\eta_{jp}}^2\right), \;\; F_\eta = diag(f_{\eta_{jp}}), \;\; j = 1, \ldots, J, \;\; p = 1, \ldots, P+1 \\ \alpha_{j,k}^t &= -\mu_j^t + \alpha_{j,k}; \;\; k = 2, \ldots, c_j - 1 \;\; . \end{aligned}$$

Given random processes $\eta^t = (\mu^t, \beta^t)$ and α^t, latent vector Y^t is independent of $Y^{t'}$ for $t \neq t'$. This property of conditional independence simplifies the estimation of the parameters.

Remark: In our application, $X_i^t = r_i^{2t} h_i^t$, will be the product of the radius at breast height (r_i^{2t}) and the total three height (h_i^t) for individuals $i = 1, \ldots, n_t$ at time t. Z_j will be ordinal variables associated with the leaf, dead branch, living branch and bark biomass ($J = 4$).

3 Posterior analysis

While a conventional approach by means of a maximum-likelihood is difficult [24], the use of conditional independence and the introduction of the underlying latent variable Y, greatly simplifies the evaluation of the *posterior* distribution [10, 9]. In the following, the set of parameters will be denoted by $\theta = (Y, \eta, F_\eta, \Sigma_\eta, \alpha, R)$.
Given a random sample of Z, the *prior* density $\pi()$ on the parameters and the DMOPM definition, the *posterior* distribution (eq. 5) of θ given observation Z is proportional to:

$$\pi(\theta|Z) \propto \mathrm{P}[Z|Y,\alpha] f(Y|\eta,R) f(\eta|F_\eta,\Sigma_\eta) \pi\left(F_\eta\right) \pi\left(\Sigma_\eta\right) \pi\left(\alpha\right) \pi\left(R\right) \tag{5}$$

where

$$\mathrm{P}[Z^t|Y^t,\alpha^t] f(Y^t|\eta^t,R) = \phi_J\left(y^t|\eta^t,R\right) \mathbb{1}_{Y^t \in A^t}$$

is a multivariate truncated Gaussian distribution. Using a mixture of Gibbs and Metropolis sampling and according to DMOPM definition (see definition 1), we propose the following algorithm.

Latent variables: *posterior* distribution of the latent variables Y, $\phi_J(Y^t|Z^t,\eta^t,R)\mathbb{1}_{Y^t \in A^t}$, is a truncated multivariate normal distribution with mean $\mu^t + X^t\beta^t$ and correlation matrix R. To sample this distribution, we use the method described by Geweke [20] which is a modified Gibbs sampling.

Regression parameters: as η^t has independent distribution (η_j^t and $\eta_{j'}^t$ independent for $j \neq j'$; η_{jp}^t and $\eta_{jp'}^t$ independent for $p \neq p'$) and is equal to a general autoregressive process:

$$f(\eta^t|\eta^{t-1},\Sigma_\eta,F_\eta) = \begin{cases} \phi_{J(P+1)}(\eta^0|0,\Sigma_0), & \text{if } \ t=0, \\ \phi_{J(P+1)}(\eta^t|F_\eta,\Sigma_\eta), & \text{if } \ t>0; \end{cases}$$

where Σ_0 is the variance of the initial state and its mean is equal to zero. Then, the *posterior* distribution is equal to

$$\eta^t|F_\eta,\Sigma_\eta,\eta^{s\neq t} \sim \mathcal{N}_{J(P+1)}(b^t,B^t),$$

with $V_i^t = diag((1, X_{i1}^t), \ldots, (1, X_{iJ}^t))$

$$\begin{aligned} B^0 &= \left(\Sigma_0^{-1} + F_\eta' \Sigma_\eta^{-1} F_\eta\right)^{-1}, \quad t = 0 \\ B^t &= \left(\Sigma_\eta^{-1} + F_\eta' \Sigma_\eta^{-1} F_\eta + \sum_{i=1}^{n_t} V_i'^t R^{-1} V_i^t\right)^{-1}, \quad t = 1, \ldots, T-1 \\ B^T &= \left(\Sigma_\eta^{-1} + \sum_{i=1}^{n_t} V_i'^t R^{-1} V_i^t\right)^{-1}, \quad t = T \end{aligned}$$

and

$$\begin{aligned} b^0 &= B^t\left(F_\eta' \Sigma_\eta^{-1} \eta^{t+1}\right), \quad t = 0 \\ b^t &= B^t\left(\Sigma_\eta^{-1} F_\eta \eta^{t-1} + F_\eta' \Sigma_\eta^{-1} \eta^{t+1} + \sum_{i=1}^{n_t} V_i'^t R^{-1} Z_i^t\right), \quad t = 1, \ldots, T-1 \\ b^T &= B^t\left(\Sigma_\eta^{-1} F_\eta \eta^{t-1} + \sum_{i=1}^{n_t} V_i'^t R^{-1} Z_i^t\right), \quad t = T \end{aligned}$$

and can be simulated in a straightforward way by a Gibbs sampling.

Autocorrelations: in this part, for sake of simplicity we have omitted the indices and f, η will denote any $f_{\eta_{jp}}$, η_{jp}; $j = 1, \ldots, J$, $p = 1, \ldots, P+1$ (see definition 1). To sample autocorrelations f , we chose a Gaussian *prior* truncated on $]-1, 1[$:

$$f \sim \mathcal{N}\left(\mu_f, \sigma_f^2\right) \mathbb{1}_{\{f \in]-1;1[\}},$$

where μ_f and σ_f^2 are known and fixed. Then, the *posterior* distribution is equal to:

$$f|\eta, \sigma_\eta^2 \sim \mathcal{N}(\hat{f}, \lambda_f^2) \mathbb{1}_{\{f \in]-1;1[\}}$$

with

$$\lambda_f^2 = \left(\frac{1}{\sigma_f^2} + \frac{\sum_{t=1}^{T} (\eta^{t-1})^2}{\sigma_\eta^2}\right)^{-1} \qquad \hat{f} = \lambda_f^2 \left(\frac{\mu_f}{\sigma_f^2} + \frac{\sum_{t=1}^{T} \eta^t \eta^{t-1}}{\sigma_\eta^2}\right).$$

Simulation of f is straightforward using Gibbs sampling.

Variances of autoregressive process: The indices have been deleted for the same reasons as above. We chose to use an inverse gamma conjugate *prior* specification $\sigma_\eta^2 \sim \mathcal{IG}(a;b)$ where a and b are fixed and sufficiently small to be a non informative *prior* distribution. The *posterior* distribution of σ_η^2 is given by:

$$\sigma_\eta^2|f,\eta \sim \mathcal{IG}\left[a+\frac{T}{2};\left\{\frac{1}{b}+\frac{1}{2}\sum_{t=1}^{T}\left(\eta^t-f\eta^{t-1}\right)^2\right\}^{-1}\right].$$

Thresholds: we now consider the sampling of the thresholds $\alpha_{j,k}; k=2,\ldots,c_j-1$ and $j=1,\ldots,J$ which are modeled in the dynamic multivariate ordinal probit model by $\alpha_j^t = -\mu_j^t + \alpha_j$. We assume that the *prior* distribution of the thresholds is the order distribution of c_j-2 uniform random variable defined as follows:

$$\left(\alpha_{j,2},\ldots,\alpha_{j,c_j-1}\right) \sim (c_j-2)!U[-u,u]^{\otimes c_j-2};\ \ j=1,\ldots,J\ ,$$

with fixed values of u. Cowles and Carlin [11] pointed out that the Gibbs sampling was not well suited to simulate these parameters and can converge very slowly. Cowles and Carlin [11] proposed a Metropolis-Hastings algorithm to simulate α_j according to its complete conditional distribution by using a truncated Gaussian conditional density $q()$ which improves the convergence of the Gibbs sampling employed by [1]. Accordingly, we used a Metropolis-Hastings algorithm to simulate the *posterior* distribution of α_j. Therefore, the proposal is accepted according to the usual Metropolis-Hastings (M-H) acceptance probability:

$$\varrho(\alpha_j,\alpha_j^\star) = \min\left(\frac{\mathrm{P}(Z_j|Y_j,Z_{l\neq j},\eta_j,\alpha_j^\star)q(\alpha_j|\alpha_j^\star)}{\mathrm{P}(Z_j|Y_j,Z_{l\neq j},\eta_j,\alpha_j)q(\alpha_j^\star|\alpha_j)},1\right)$$

where $\alpha_j^\star = \alpha_j^{\star t} + \mu_j^t$. The proposal distribution is given by:

$$q(\alpha_{j,z_j}|\alpha_{j,z_j}^\star,\alpha_{j,z_j-1}^\star,\alpha_{j,z_j+1}) = \mathcal{N}(\alpha_{j,z_j}^\star,\omega^2)\mathbb{1}_{\{\alpha_{j,z_j}\in]\alpha_{j,z_j-1}^\star;\alpha_{j,z_j+1}[\}},$$

where ω^2 a fixed parameter and

$$\mathrm{P}(Z_j|Y_j,Z_{l\neq j},\eta_j,\alpha_j) = \prod_{t=1}^{T}\prod_{i=1}^{n_t}\int_{\alpha_{j,z_j-1}}^{\alpha_{j,z_j}} f(Y_{ij}^t|Y_{i,l\neq j}^t,\eta^t,R)dY_{ij}^t\ .$$

Dependencies between latent variables: as proposed by Chib and Greenberg [10], we used a random-walk Metropolis-Hastings algorithm to simulate the *posterior* distribution of the correlation matrix R. Let $\rho = (\rho_{1,2},\ldots,\rho_{J-1,J})$ denote the vector of the $M = [J(J-1)]/2$ correlations with $\rho \in [-1;1]^M$. We used a uniform distribution on $[-1,1]^M$ as *prior* distribution. The proposal distribution is given by:

$$q(\rho^\star|\rho) = \rho + \iota$$

where ι is an independent symmetric random disturbance. This proposal is therefore accepted according to the usual Metropolis-Hastings (M-H) acceptance probability

$$\varrho(\rho, \rho^\star) = \min\left(\frac{f(Y|\eta, R^\star)\mathbb{1}_{\{\rho^\star \in [-1;1]^M\}}}{f(Y|\eta, R)\mathbb{1}_{\{\rho \in [-1;1]^M\}}}, 1\right)$$

where $f(Y|\eta, R^\star)/f(Y|\eta, R)$ is the likelihood ratio. Chib and Greenberg [10] discussed the choice of the conditional density in relation with the number of ordinal variables.

4 Simulations

Saint-André et al. [36] developed biomass equations for each compartment, including leaves, bark, dead branches and so on. Our approach assesses the biomass of the above compartments. In this section we use Saint-André et al. [36] equations to investigate the quality of the estimations for the different parameters and the biomass. We focus on the number of observations (in terms of trees) and on the time points at which the measurements are to be taken for each unit. The aim of these simulations is to indicate which protocol (number of trees, number of modalities for each biomass, how many measurements) should be envisaged. Mortier et al. [29] showed that when variables were binary, the number of observations must be greater than 500 to obtain an accurate estimation of the correlation matrix and Euclidean distances. This number is not realistic when estimating biomass. On the other hand, if dependent variables are ordinal with three modalities and if the number of observations is sufficiently large, the approach proposed by Mortier et al. [29] gives accurate results. This is why we opted for this approach and assumed that the dependent variables were ordinal with 3 modalities. In practice, this option is applicable in the field. We set the number of ordinal dependent variables to $J = 4$, corresponding to bark (Y_1), living branches (Y_2), dead branches (Y_3), and leaves (Y_4). Saint-André et al. [36] equations for these four compartments are

$$\begin{aligned} Y_{1i} &= 9.08 * X_i^{0.72} + \varepsilon_1, \\ Y_{2i} &= (7.78 + 1224.1e^{-0.18age})X_i + \varepsilon_2, \\ Y_{3i} &= (11.67 - 0.084age)X_i + \varepsilon_3, \\ Y_{4i} &= (5.26 - 0.024age + 565.1e^{-0.15age})X_i + \varepsilon_4, \end{aligned} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix} \sim \mathcal{N}_4(0, R) \quad (6)$$

where X_i is the product of the squared radius at breast height (r_i^{2t}) and the total tree height (h_i^t) for individuals $i = 1, \ldots, n_t$ at time t. The different parameters and the correlation matrix R were estimated using conventional destructive methods. Correlation R was

$$R = \begin{pmatrix} 1.0 & 0.22 & 0.15 & -0.16 \\ 0.22 & 1.0 & -0.22 & 0.10 \\ 0.15 & -0.22 & 1.0 & -0.23 \\ -0.16 & 0.10 & -0.23 & 1.0 \end{pmatrix}.$$

The covariable $X = r^2h$ was simulated using a specific eucalyptus growth model. A complete description of the chain of models Eucalypt-Dendro can be found in Saint-André et al. [35]. Simulations were performed for an average site index (fertility). Briefly, this growth and yield model use four main equations: (i) dominant height is modelled as a function of stand age, stand density and site index, (ii) the stand basal area increment was modelled as a function of the dominant height increment, (iii) the individual tree basal area growth was modelled as a function of the tree circumference at the previous time and the stand basal area increment, and (iv) the height of the trees was obtained from a height-girth relationship. As a result, both stand and tree traits were simulated monthly. These simulated data are in agreement with observed data on fields at Congo and have been simulated several times. Nevertheless, in this part, we present results for different scenarios concerning with only once simulated data (results based on the others are similar). Finally, the thresholds α were chosen on a variable-by-variable basis. They are based on quantiles of the simulated latent variables in such a manner to give an equilibrium distribution of the observations in each class. The consequences of this choice were investigated and were not found to have any noticeable impact on the results. Finally, for the different thresholds, latent variables Y were discretized to obtain ordinal variables Z.

To study the quality of the biomass estimation based on ordinal variables, we considered the number of observation effects and the number of time points at which the measurements were taken for each unit. Our figures correspond to 50, 100 and 200 observations (trees) measured 7, 14 and 21 times. This choice corresponds to one, two or three measures per year until the eucalyptus is 7 years old, at which time the stand is harvested. Five simulations were performed for each situation giving $3 \times 3 \times 5 = 45$ simulations. Each simulation was based on 50000 MCMC runs with thinning of 3 and with a burning of 20000.

In the application we use non informative priors by setting $\Sigma_0 = diag(1000000)$, $a_j = 0.01$ and $b_j = 0.0001$; $j = 1, \ldots, 4, .$ Finally, we set $\mu_{f_j} = 0$ and $\sigma^2_{f_j} = 10$. Various MCMC runs have been conducted using different initial values. All MCMC estimations have given same results. Gelman et al. [19] statistics have been calculated and always tended to one (see figure 1).

4.1 Quality of parameter estimations

First, we focussed on the quality of the estimation for the correlation matrix R (table 1). Results have been obtained using same data where 50 and 100 observations have been randomly chosen among 200 simulated data.

These results underline that the number of measurements time points and the number of observations have a little impact on the quality of the estimation. This matches the conclusion drawn by Mortier et al. [29]. Indeed, if variables are ordinal with three modalities and if the number of observations is greater than 200, the correlation estimation is satisfactory. In our model, the correlation matrix is time-independent which leads to a homoscedasticity assumption. So 7 time points and 50 observations (trees) give 350 measures used to estimate the correlation matrix.

$\rho_{12} = 0.22$	$n = 50$	$n = 100$	$n = 200$
$T = 7$	0.213(0.086)	0.232(0.066)	0.230(0.047)
$T = 14$	0.199(0.059)	0.193(0.046)	0.210(0.038)
$T = 21$	0.210(0.050)	0.227(0.042)	0.229(0.035)
$\rho_{24} = 0.10$	$n = 50$	$n = 100$	$n = 200$
$T = 7$	0.114(0.072)	0.121(0.055)	0.096(0.039)
$T = 14$	0.123(0.054)	0.099(0.037)	0.109(0.025)
$T = 21$	0.124(0.039)	0.096(0.029)	0.110(0.020)

Table 1: *Posterior* mean and standard deviation for some correlations of R for different numbers of observations (n) and number of measures over time (T).

We then studied the impact of the number of time points and observations on parameter estimations. As no reference value was available, we choose 21 time points and 200 observations as reference values. Table 2 presents the estimated autocorrelations.

We observed that the level of correlation and the number of time points had a significant impact on the estimation results while the number of observations did not. Indeed, it would seem that when the autocorrelations values were between 0 and 0.6, the estimation were fairly exact irrespective of the number of time points or the number of observations: the standard deviation values were similar for all numbers of time points and observations. But, when the autocorrelation value was strong (greater than 0.6), the number of time points had a significant impact on the estimations. The bias is strong when 7 measurement time points were used. This bias was slightly lessened when the number of observations increased but even with 200 observations there was still a marked discrepancy between the

reference value: $f = 0.58$	$n = 50$	$n = 100$	$n = 200$
$T = 7$	0.561(0.045)	0.555(0.032)	0.560(0.028)
$T = 14$	0.574(0.036)	0.572(0.028)	0.564(0.026)
$T = 21$	0.597(0.037)	0.593(0.033)	0.579(0.022)
reference value: $f = 0.98$	$n = 50$	$n = 100$	$n = 200$
$T = 7$	0.334(0.366)	0.419(0.354)	0.612(0.310)
$T = 14$	0.848(0.063)	0.901(0.055)	0.957(0.032)
$T = 21$	0.940(0.055)	0.975(0.022)	0.982(0.017)

Table 2: Estimation and standard deviation for some autocorrelations of F for different numbers of observations (n) and number of measures over time (T).

estimated autocorrelation (0.612) and the reference autocorrelation (0.982). Estimations became satisfactory for 14 measurement time points (and more), irrespective of the number of observations. As a first conclusion, we can state that the number of measurement time points has a greater impact on parameter estimations than the number of observations.

In our approach, one of the major difficulties is to estimate the variance of auto-regressive processes accurately. Table 3 presents the estimation of these variance. We assumed again that the estimation based on the greatest number of time points and observations would be most appropriate as reference value.

$n = 200$	$T = 7$	$T = 14$	$T = 21$	$T = 21$	$n = 50$	$n = 100$	$n = 200$
Σ_{β_1}	50.52	0.19	0.14	Σ_{β_1}	0.90	0.37	0.14
Σ_{β_2}	58.08	17.48	16.19	Σ_{β_2}	16.61	18.42	16.19
Σ_{β_3}	15.85	0.14	0.10	Σ_{β_3}	0.49	0.20	0.10
Σ_{β_4}	40.96	12.48	4.11	Σ_{β_4}	11.48	4.54	4.11

Table 3: Estimation of Σ_β for different numbers of observations (n) and number of measures over time (T).

In the same manner as previously, the number of time points had a greater impact on the Σ_β estimations than the number of observations. When 200 individuals were measured, the variance of the auto-regressive process was closely related to the number of measurement time points: at least 14 time points may be recommended. On the other hand, when 21 measurements were made, the variances were roughly the same for 50, 100 and 200 observations even though at least 100 observations were recommended. Lastly, when we crossed the number of time points and the number of observations (trees), then 100 observations and 14 time points gave satisfactory results with respect to the reference value given by 200 observations and 21 time points.

To conclude with regard to the quality of the estimation for unknown parameters Σ_β and F, it may be stated that the effect of the number of observations is weak whereas that due to the number of time points is greater. If the number of observations is sufficiently large (100 observations) and if the number of measurement time point is 14, the approach proposed by the DMPOM (see definition) seems to give good results for parameter estimations.

4.2 Quality of biomass estimations

This section investigates the quality of biomass estimations which are of major importance for the forest manager or forest scientist. Figures 2, 3, 4, present the results obtained for the studied compartments.

First, regardless of the number of observations and the number of measurement time points, the young stages of the living branch and leaf biomasses were incorrectly estimated. This may stem from the difficulty encountered when estimating the initial value of a dynamic system. The biomass growth curve for these two compartments is bell-shaped as

shown by Saint-André et al. [35]. But, in our application, we did not simulate sufficient time points in the young stages to catch the entire curve. Therefore, the initial value (zero biomass at time 0) had a marked impact on the year 1 and 2 estimations, resulting in under-estimations for these two stages. Second, the number of observations had a marked impact on the estimations of both mean biomass and its confidence intervals. When 50 observations were used, some discrepancy between the simulated and the estimated biomass were observed (see bark and dead branch biomasses). These differences were slightly diminished when the number of measurement time points increased. Furthermore, the confidence intervals were 3 times broader for 50 observations than for 200 observations. In conclusion, when estimating biomass at least 14 measurement time points combined with at least 100 observations may be recommended.

5 Discussion and Conclusions

Destructive sampling is a major drawback in the current methodology used to estimate carbon stocks and biomasses because it renders impossible to gather longitudinal data and therefore to assess the effect of the environment, competition and age on the tree and on stand biomasses. DMPOM was seen as a good opportunity to overcome this obstacle and this paper proved that such models are able to provide accurate estimates of standing biomasses provided that the dataset is sufficiently large (number of trees x number of inventories > 1400; the recommendation being 100 trees measured 14 times). Trees cannot be measured more than once a year during usual forest management operations but for research purposes this condition could be easily fulfilled. Permanent plots often embed 50 to 100 trees that can be measured and visually assessed three or four times a year. The time elapsing between inventories can be shortened for the first years (e.g.: one measurement every 2-3 months) and enlarged when the stand reaches maturity (after 4 years, one inventory per year). The use of a continuous time process could be envisaged. This is a real innovation because, thanks to DMPOM and the combined measurement of height, diameter and visual assessment, standing biomass was estimated without felling the trees. This also provides the possibilities to follow the seasonal course of compartments with high rates of turn-over (such as leaves or bark). Furthermore, a great improvement can be expected for dead branch biomass estimations: conventional models are inappropriate because biological features are difficult to differentiate (death of the branch and abscission) from random events (wind, animals which may cause an artificial branch pruning). The visual tree assessment is a way to overcome this problem.

The only constraint for field applicability lies in the protocol to be applied. Visual assessment is basically "user-dependent". Each inventory should be prepared by performing a pre-assessment using a digital camera. About 20 to 40 contrasted trees should be photographed and then by ranking should be placed in 3 or 4 classes of biomass for each tree compartment (leaves, living branches, dead branches and bark; we make the assumption that for the trunk, diameter and height are sufficient to assess the stem biomass correctly). By selecting photos that are representative of each class of biomasses, the operator can

therefore rank all the trees in the permanent plot. The pre-assessment should be made *prior* to each inventory. In addition, the time required to perform these operations is far less than that required for tree felling, drying aliquots and weighting all the samples (almost 2 to 3 months for 12 to 24 felled trees). This protocol will soon be tested in connection with carbon research operations that started 4 years ago in eucalyptus plantations in Congo [35, 36, 16]. Because the clone of this study nowadays tends to be replaced by more productive ones, we will take advantage of new field campains that are planned to assess the biomass of these innovative clones (by way of traditionnal tree felling). It will therefore be possible to confront DMPOM estimates to actual biomass values and test our model accuracy to real data.

Finally, assuming a multivariate time process to take into account simultaneously the dependency between time regression parameters (η^t) could be envisaged. Nevertheless, it is time consuming. Moreover, homoscedasticity could be not relevant. Taking into account Heteroscedasticity would generalized MPMOD (see definition 1) but also Item Response Theory; see for exemple Lord [27].

References

[1] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[2] L.A. Apiolaza, A.R. Gilmour, and D.J. Garrik. Variance Modelling of Longitudinal Height Data from a Pinus Radiata Progeny Test. *Canadian Journal of Forest Research*, 30:645–654, 2000.

[3] S.A. Ashford and R. R. Swoden. Multivariate probit analysis. *Biometrics*, 26:535–546, 1970.

[4] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data.* Monographs on Statistics and Applied Probability 101. Boca Raton, FL: Chapman and Hall/CRC. xvii, 452 p., 2004.

[5] J.E. Bedrick, J. Lapidus, and F.J. Powell. Estimating the Mahalanobis distance from mixed continuous and discrete data. *Biometrics*, 56(2):394–401, 2000.

[6] F. Caillez. *Estimation des volumes et accroissements des peuplements forestiers avec référence particulière aux forêts tropicales. Vol .1 Estimation des volumes.* Etude FAO Forêt. N:22/1. FAO Rome. 99p, 1980.

[7] B.P. Carlin and N.G. Polsen. Monte Carlo bayesian methods for discrete regression models and categorical time series. *Bayesian Statistics 4, J. Bernardo et al., Eds., Oxford University Press, Oxford*, pages 577–586, 1993.

[8] B.P. Carlin, N.G. Polsen, and D.S. Stoffer. A Monte Carlo approach to nonnormal and nonlinear state-space-modelling. *Journal of the American Statistical Association*, 87:493–500, 1992.

[9] M.H Chen and Q.M. Shao. Properties of prior and posterior distribution for multivariate categorical response data model. *Journal of Multivariate Analysis*, 71:277–296, 1999.

[10] S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2): 347–361, 1998.

[11] M.K. Cowles and B.P. Carlin. Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association,*, 91:883–904, 1996.

[12] D.R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika,*, 91:729–737, 2004.

[13] C. Daganzo. *Multinomial Probit. The Theory and its Application to Demand Forecasting.* Economic Theory, Econometrics, and Mathematical Economics. New York etc.: Academic Press, A Subsidiary of Harcourt Brace Jovanovich, Publishers. XIV, 222 p., 1979.

[14] A. R. De Leon. Pairwise Likelihood approach to grouped continuous model and its extension. *Statistics and Probability letters,*, 2005.

[15] P. J. Diggle, P. J. Heagerty, K.Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data. 2nd ed.* Oxford Statistical Science Series. 25. Oxford: Oxford University Press. xv, 379 p., 2003.

[16] D. Epron, Y. Nouvellon, O. Roupsard, W. Mouvondy, A. Mabiala, L. Saint-André, R. Joffre, C. Jourdan, J-M. Bonnefond, P. Berbigier, and O. Hamel. Spatial and temporal variation of soil respiration in a Eucalyptus plantation in Congo. *Forest Ecology and Management*, 202:149–160, 2006.

[17] L. Fahrmeir. Posterior mode estimation by extended Kalman Filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87:501–509, 1992.

[18] Alan E. Gelfand and Adrian F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[19] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis. 2nd ed.* Boca Raton, FL: Chapman and Hall/CRC. xxv, 668 p., 2004.

[20] J. Geweke. Efficient Simulation from the Multivariate Normal and Student t Distributions Subject to Linear Constraints. *E. M. Keramidas (ed.),Computing Science and Statistics: Proceedings of the Twenty Third Symposium on the Interface*, pages 571–578, 1991.

[21] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[22] H. Joe. *Multivariate Models and Dependence Concepts.* Monographs on Statistics and Applied Probability. 73. London: Chapman and Hall. xviii, 399 p., 1997.

[23] G. Kauermann. Modeling longitudinal data with ordinal response by varying coefficients. *Biometrics*, 56(3):692–698, 2000.

[24] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions. Vol. 1: Models and applications. 2nd ed.* New York, NY: Wiley. xxii, 722 p.,, 2000.

[25] E. Lesaffre and G. Molenberghs. Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medecine*, 10:1391–1403, 1991.

[26] L.C Liu and D. Hedeker. A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, 62:261–268, 2006.

[27] R. Lord. *Applications of Item Response Theory to Practical Testing Problems.* Erlbaum, Hillside, N.J., 1980.

[28] C. E. McCulloch and S. R. Searle. *Generalized, Linear and Mixed Models.* Wiley Series in Probability and Statistics, 358 p., 2001.

[29] F. Mortier, S. Robin, S. Lassalvy, Baril C.P., and A. Bar-Hen. Prediction of Euclidean distances with discrete and continuous outcomes. *Journal of Multivariate Ananlysis.*, In Press, 2005.

[30] S. M. O'Brien and D. B. Dunson. Bayesian Multivariate logistic regression. *Biometrics*, 60:739–746, 2004.

[31] J. Pardé. Forest Biomass. *Forestry Abstract Review Article*, 41(8):343–362, 1980.

[32] B.R. Parresol. Assessing tree and stand biomass: a review with examples and critical comparisons. *Forest Science*, 45(4), 1999.

[33] R. Rekayaa, D. Gianola, and G. Shookb. Longitudinal random effects models for genetic analysis of binary data with application to mastitis in dairy cattle. *Genet. Sel. Evol.*, 35:457–468, 2003.

[34] D. Rivers and Q. Vuong. Model selection tests for nonlinear dynamic models. *The Econometrics Journal*, 5:1–10, 2002.

[35] L. Saint-André, J-P. Laclau, J-P. Bouillet, P. Deleporte, A. Miabala, N. Ognouabi, H. Bailléres, and Y. Nouvellon. Integrative modelling approach to assess the sustainability of the eucalyptus plantations in Congo. In *In: Connection between Forest Resources and Wood Quality: Modelling Approaches and Simulation Software.IUFRO Working Party S5.01.04.*, pages 611–621, 2002.

[36] L. Saint-André, A. Thongo, A. Mabiala, W. Mouvondy, C. Jourdan, O. Roupsard, P. Deleporte, O. Hamel, and Y. Nouvellon. Age related equations for above and below biomass of a eucalyptus hybrid in Congo. *Forest Ecology and Management*, 205:199–214, 2005.

[37] C. Sicard, L. Saint-André, D. Gelhaye, and J. Ranger. Effect of initial fertilisation on biomass, nutrient content and nutrient-use efficiency of even-aged Norway spruce and Douglas-fir stands planted in the same ecological conditions. *Trees Structure and Function, In press*, 2006.

[38] Thomas R. Ten Have and Alfredo Morabia. Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics*, 55(1):85–93, 1999.

[39] D. Zhang. Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics*, 60(1):8–15, 2004.

Contents

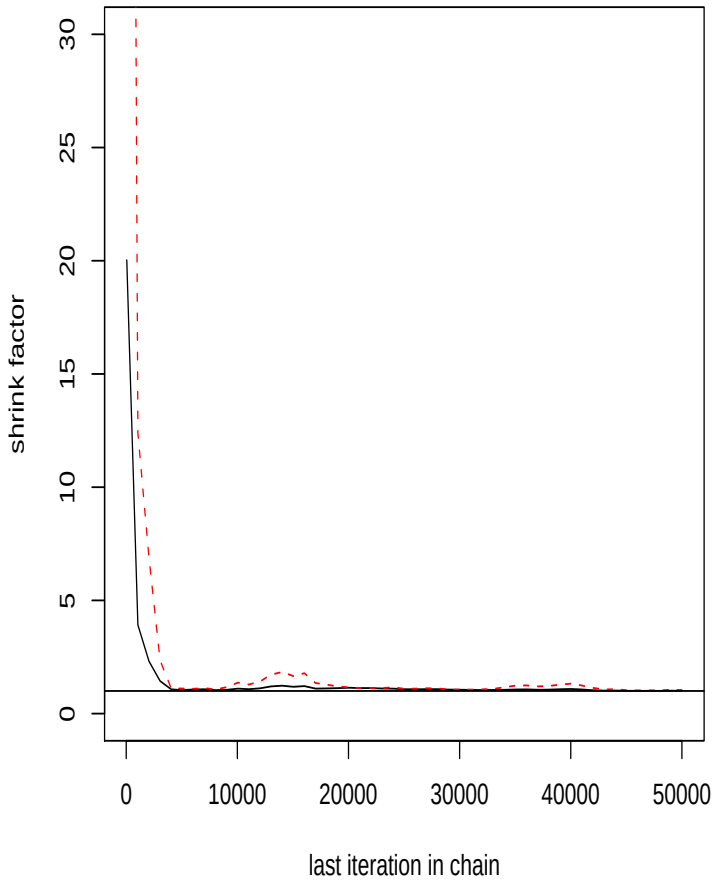

Figure 1: Gelman and Rubin statistics (red line is $97,5\%$ interval confidence and the black one is the estimated statistic)

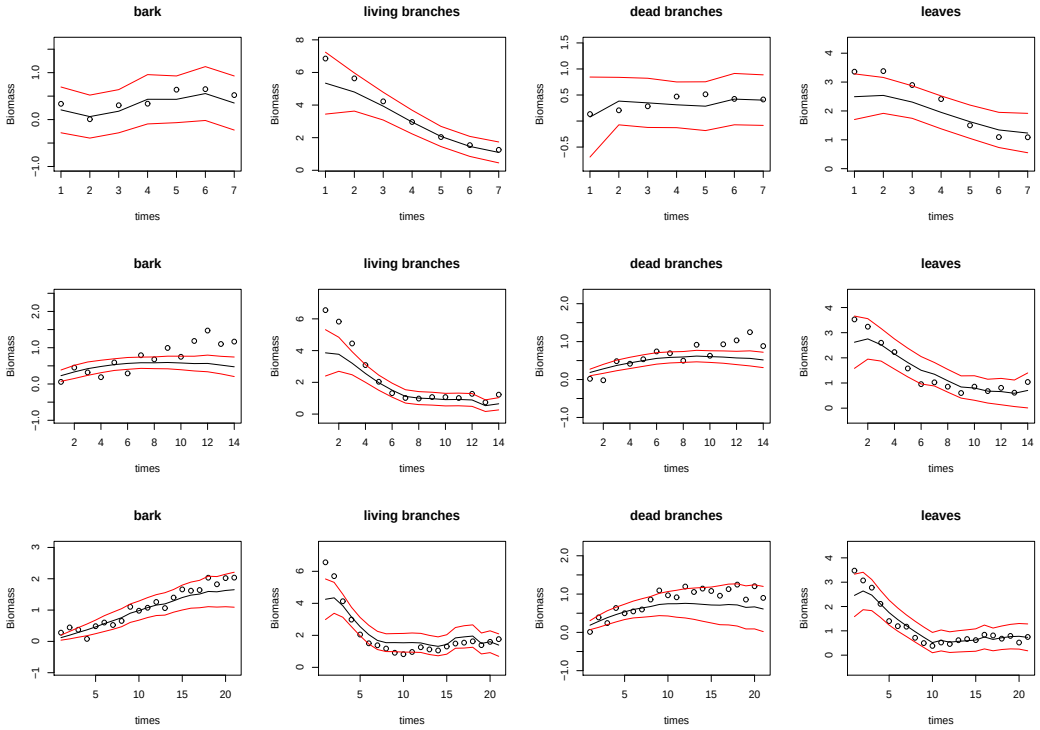

Figure 2: Mean of each biomass by compartment for 50 observations, T=7,14,21 measures over time. True values (points), estimated values (solid line) and confidence intervals (dashed line)

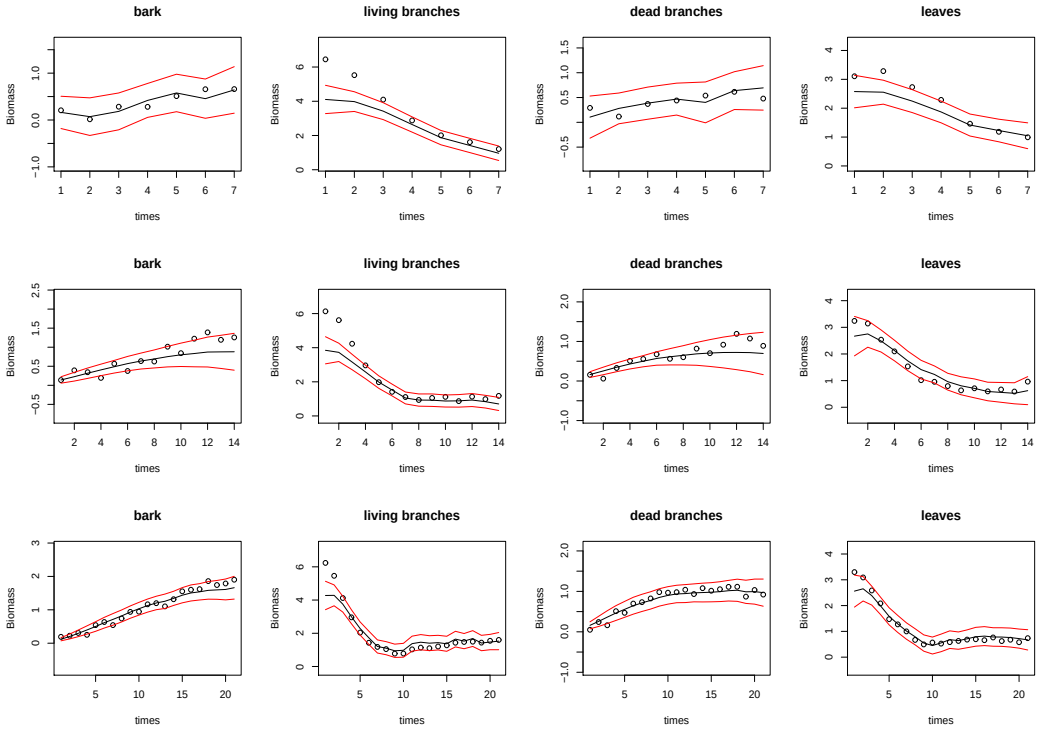

Figure 3: Mean of each biomass by compartment for 100 observations, T=7,14,21 measures over time. True values (points), estimated values (solid line) and confidence intervals (dashed line)

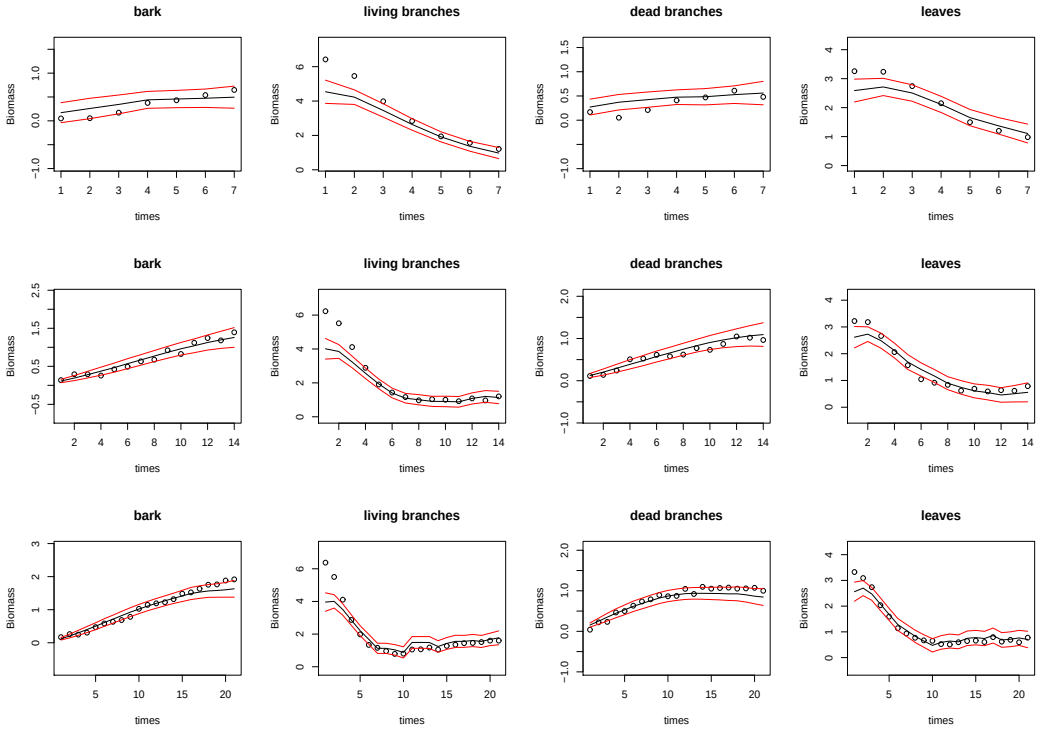

Figure 4: Mean of each biomass by compartment for 200 observations, T=7,14,21 measures over time. True values (points), estimated values (solid line) and confidence intervals (dashed line)

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399