Linear Mixed Modelling of Cone Production for Stone Pine in Portugal

***Abel Rodrigues, **Giovani L. Silva, ***Miguel Casquilho, ****João Freire, *Isabel Carrasquinho e *****Margarida Tomé**

Abstract. This study aimed to model the cone production weight of stone pine (*Pinus pinea* L.) trees using two approaches: classic linear mixed modelling and Bayesian mixed modelling. The field data were collected in the Setúbal Peninsula, the main production area in Portugal, where 51 plots for monitoring cone production per tree during three production periods (2004–2005, 2005–2006 and 2006–2007) were set up. Linear mixed models with a random intercept term were fitted to the whole dataset (416 trees) for the three production periods. Clustered longitudinal mixed models, under a Bayesian approach with random terms related to intercept, time slope and nested effects, were fitted to a subset of the global dataset (9 plots and 76 trees), corresponding to the plots where production data for all trees in the three years were available. The selected models included, as independent variables, crown width, basal area per hectare, tree height and accumulated rainfall in the five-year periods, prior to cone collection. Despite the small period considered in this study, the Bayesian analysis proved to be useful for calculations

*Senior Researcher, Instituto Nacional de Investigação Agrária e Veterinária, IP. Unidade Estratégica de Investigação de Sistemas Agrários e Florestais e Sanidade Vegetal. Quinta do Marquês, 2780-159 OEIRAS.
E-mails: abel.rodrigues@iniav.pt; isabel.carrasquinho@iniav.pt;
**Professor, Centro de Estatística e Aplicações and Dept. Matemática, Instituto Superior Técnico, Universidade de Lisboa. Av. Rovisco Pais 1, 1049-001 LISBOA.
E-mail: giovanni.silva@tecnico.ulisboa.pt
***Professor, Centro de Processos Químicos, Instituto Superior Técnico, Universidade de Lisboa. Av. Rovisco Pais 1, 1049-001 LISBOA.
E-mail: mcasquilho@tecnico.ulisboa.pt
****Research grant-holder, *****Full Professor, Centro de Estudos Florestais. Instituto Superior de Agronomia. Universidade de Lisboa, Tapada da Ajuda, 1349-017 LISBOA.
E-mails: joaofreire71@gmail.com; magatome@isa.ulisboa.pt

with smaller samples, and it was indicative about the tree and plot biometric complex dynamics, subjacent to cone production.
Key words: stone pine, cone production, mixed models, Bayesian analysis, clustered longitudinal data

Modelação da produção de pinha em pinheiro-manso em Portugal recorrendo a modelos lineares mistos

Sumário. Este estudo visou a quantificação da produção de pinha em povoamentos de pinheiro manso (*Pinus Pinea* L.) segundo duas abordagens: utilização de modelos mistos clássicos e modelação Bayesiana longitudinal mista. Os dados de campo foram obtidos na Península de Setúbal, a principal zona produtora em Portugal, onde se instalaram 51 parcelas para quantificação da produção de pinha durante três períodos de produção (2004-2005, 2005-2006 e 2006-2007). Os modelos mistos lineares com um termo de interceção aleatório, correspondente a cada parcela, foram ajustados a um conjunto de 416 árvores, para os três períodos, independentemente do período de produção ou de não disponibilidade de alguns dados, devido a dificuldades logísticas no terreno. Os modelos mistos longitudinais, ajustados por metodologia Bayesiana, com termos aleatórios associados à interceção, fator temporal e efeitos encaixados foram aplicados a um subconjunto de 76 árvores em 9 parcelas, para as quais estavam disponíveis os dados de produção para os três períodos de produção. Os modelos selecionados incluíram como variáveis independentes a largura das copas, a área basal por hectare, a altura das árvores e a precipitação acumulada nos períodos de cinco anos, anteriores à recolha das pinhas. Apesar do curto período abrangido por este estudo, a análise Bayesiana revelou-se útil para cálculos com amostras e indicativa sobre a dinâmica complexa do sistema árvore-parcela, subjacente ao processo de produção de pinha.
Palavras-chave: pinheiro manso, produção de pinha, modelos mistos, análise Bayesiana, análise longitudinal agrupada

Modélisation de la production de cônes de pin parasol au Portugal avec des modèles mixtes linéaires

Résumé. Cette étude visait à quantifier le poids de la production de cônes en peuplements de pin parasol (*Pinus pinea* L.), moyennant deux approches: modélisation mixte linéaire et modélisation Bayésienne mixte longitudinale. Les données de champ ont été recueillies dans la péninsule de Setúbal, la principale zone de production au Portugal, y étant préparées 51 parcelles pour l'observation de la production de cônes pendant trois périodes productives (2004-2005, 2005-2006 et 2006-2007). Des modèles mixtes linéaires avec un terme d'interception aléatoire, correspondant à chaque parcelle, ont été ajustés à l'ensemble des données (avec 416 arbres) pour les trois périodes productives, sans rapport avec l'année de production ou l'absence de données, en raison de problèmes logistiques dans le champ. Des modèles mixtes longitudinaux, sous une approche

Bayésienne avec des termes aléatoires relatifs à l'ordonnée à l'origine, au facteur temps et aux effets imbriqués ont été ajustés à un sous-ensemble des données globales (9 parcelles et 76 arbres), correspondant aux parcelles où étaient disponibles des données productives pour tous les arbres dans les trois années. Les modèles sélectionnés ont inclus, comme variables indépendantes, la largeur de la cime, la surface basale par hectare, la hauteur de l'arbre et la précipitation accumulée dans la période de cinq ans antérieure à la récolte des cônes. Malgré la brièveté de la période ciblée dans cette étude, l'analyse Bayésienne s'est révélée utile pour les calculs avec des échantillons plus petits, et fut révélatrice quant à la complexe dynamique biométrique sur l'arbre et la parcelle, sous-jacente à la production de cônes de pin.

Mots-clés: pin parasol, cônes, modèles mixtes, analyse Bayésienne, données groupées longitudinales

Introduction

Stone pine (*Pinus pinea* L.) is a Mediterranean forest species with the highest rate of increase in area in Portugal since 1995, occupying approximately 176 thousand ha, about 6% of the total national forest area (ICNF, 2013). It represents the fifth largest forest area, mainly for kernel production. From the total forest area of this species, about 106,200 ha correspond to young pure plantations (ICNF, 2013). This large expansion is due to its good adaptability to high air temperatures and low rainfall in summer, typical of the Mediterranean climates and to the high economic value of the edible kernels. Spain is also another big stone pine producer with an area of about 60% of the world's stone pine woodland and pine nut production (CALAMA *et al*., 2008). Together the two countries account for about 75% of stone pine's world area. Portuguese average cone production was 193 kg ha^{-1} in 2006 (ICNF, 2013).

Stone pine stands are also less sensitive to diseases and pests comparatively to other Mediterranean pine species, particularly *Monochamus galloprovincialis* vector of the pine wood nematode (*Bursaphelenchus xylophylus*), the agent of pine wilt disease recorded in Portuguese maritime pine stands since 1999. This makes stone pine a potential candidate for the replacement of some attacked maritime pine forests.

Stone pine is a native species in Portugal with the Setúbal Peninsula, in the southern coast, being its main concentrated area with more than 50% of national cone production (ICNF, 2013). Stone pine provides wood and non-wood products. Its main value comes from the nuts which are the most important edible product in Mediterranean forests (CARRASQUINHO *et al*., 2010). Less important, but still relevant, are the wood and resin productions.

Cone production per tree is a complex process depending on intrinsic and extrinsic factors to trees (CANÃDAS, 2000). There are genetic, hormonal and nutritional factors associated with tree age and health status. Site, climate and disturbances such as forest fire or diseases may also affect stand productivity. All these factors interact in complex patterns and processes, e.g., flowering dynamics, easiness in cone detachment or yearly fluctuation of cone production per tree. In Portugal, long term empirical evidence suggests that cone yield follows a cycle of 7 or 8 years (CARRASQUINHO *et al*., 2010). Nut yield changes every year, and the annual number of flowers, related to the number of cone buds, depends on the winter conditions of the previous year. On the other hand, a plentiful flowering sinks organic and inorganic nutrients, allowing for less

available resources to the trees in the following vegetative cycles, with lower diameter and crown growth (CANÃDAS, 2000).

A linear mixed model (LMM) is a parametric linear model that includes a continuous response variable and also random and fixed effects as independent variables, allowing the clustering of available data. The parameter estimation is based on maximum likelihood (ML) or restricted maximum likelihood (REML) theory (e.g., SEARLE *et al*., 2006), allowing to simultaneous estimation of fixed and random effects, under a proposed matrix structure for random variables and error terms. LMMs are primary fundamental tools with an untapped potential for forest modelling at large. LMMs allow to quantify the linear relationships (e.g., CANÃDAS, 2000; CALAMA, 2004; CALAMA and MONTERO, 2005; CALAMA *et al.*, 2007) between stone pine cone production and tree and plot biometrical and environmental covariables. These models, by dividing the residual variance into different components, facilitate a discrimination of different sources of stochastic variability, related to possible levels of clusters, repeated measures and longitudinal analysis. Examples of these covariables are: diameter at breast height, crown diameter and height, tree height, average distance between trees, total number of trees per hectare or basal area by hectare. Linear mixed models can consider both classic (or frequentist) and the Bayesian statistical approaches. The Bayesian approach relies on the prior information about the model parameters, for which is considered a probability distribution, beyond the data information present on the likelihood, allowing to obtain the posterior distribution in order to make inference on the model parameters (e.g., PAULINO *et al*., 2003; GELMAN *et al*., 2006). We believe that the use of the Bayesian analysis for modelling cone production in stone pine is innovative compared to previous related studies.

Significant research work has been made on stone pine in the last 15 years concerning items as diameter increment (CALAMA and MONTERO, 2005; FREIRE, 2009), regional growth models (GARCÍA GÜEMES, 1999; CALAMA, 2004), variability of cone and seed production with geographical location and time by use of mixed models (CALAMA and MONTERO, 2005; CALAMA, 2004; FREIRE, 2009), empirical ecological type models for predicting stone pine cone production (CALAMA *et al*., 2008), stem form and tree volume (ALPUÍM *et al*., 2000; CALAMA and MONTERO, 2006) or effects of fertilization in stone cone production (CALAMA *et al*., 2007). This recent work has been carried out mainly in Spain, whereas for Portugal few generalized production models have been found in literature (FREIRE, 2009; CARRASQUINHO *et al*., 2010).

Regarding meteorological conditions, the most cone productive areas within a region of natural distribution are these in lower elevation, with warmer temperatures, free of extreme winter conditions and without lack of available water (CALAMA *et al*., 2008). Winter rainfall also explains temporal and spatial variability of flowering and fruiting processes in Mediterranean forests (KOENIG *et al*., 1996; MUTKE *et al*., 2005; CALAMA *et al*., 2008). Stone pine has a strong epinastic control based on two climatic years regarding the formation of buds and the primary growth, respectively (SUROVÝ *et al*., 2011). According to GONÇALVES and POMMERENING (2012), long term climate and production data would be necessary for each stone pine stand, to evaluate the influence on cone production by intrinsic and extrinsic factors.

Individual cone yield is linked to variables related to tree biometry, such as diameter at breast height (*dbh*) and crown diameter, up to advanced stages of cone production (CAÑADAS, 2000; MUTKE *et al*., 2000). Evidence indicates that trunk section just below the crown is well correlated with photosynthetic activity inducing leaf and cone production (CAÑADAS, 2000). For Spain, GARCÍA GÜEMES (1999), MUTKE *et al*. (2000) and CALAMA *et al*. (2008) pointed out that higher average cone yield both at stand or tree level was related to higher site index and also to covariables such as a lower number of trees or a higher basal area per hectare. The stand basal area is strictly associated to plot density and free space available for trees. CAÑADAS (2000) suggests dominant height of the hundred trees with the higher *dbh* in dominant strata per ha on regular stands as a parameter to evaluate site index in stone pine forests. GONÇALVES and POMMERENING (2012) for west-southern Portugal, grossly the same area of this study, point out that tree spacing and crown size are factors positively correlated with stone pine cone production. The tree spacing influence is probably associated with the light demanding nature of this species and with the competition for water and nutrients especially in conditions of low summer rainfall and low soil water and nutrient availability.

The present study is a sequence of the work made by CARRASQUINHO *et al.* (2010), on stone pine stands in Setúbal Peninsula. Our main objective is to obtain two independent kinds of linear mixed models, based on classic LMMs and Bayesian approaches, aimed to quantify the cone production per stone pine tree for the period of 2004-2007.

Materials and methods

Datasets and exploratory analysis

Under a joint initiative of the Portuguese "National Institute of Agrarian and Veterinarian Research" (INIAV), the Institute of Agronomy, the Portuguese Forest Service and local associations of forest owners, 51 stone pine field plots (Table 1) for measurement of cone production and several biometric variables were set up in 2004 and 2005. These plots are representative of the main stone pine Portuguese provenance region in the Setúbal Peninsula. The field work took place from 2004 to 2007. The plots had a circular shape, were typically even-aged and natural-regenerated. Plots with a maximum of 25 trees were 20 m radius and 30 m otherwise. The 14 variables measured or calculated were tree height (h), diameter at breast height (dbh), crown width (cw), basal area (g), crown height (ch), base crown height (bch), total number of trees per hectare (N_T), total basal area per hectare (G_T), dominant height (HD), dominant diameter (DD), crown area (CA), number of stone pine trees per hectare (N_{SP}), stone pine basal area per hectare (G_{SP}), and annual rainfall (R).

Five groups of trees were defined (CARRASQUINHO *et al*., 2010) depending on diameter at breast height: i) juvenile (dbh < 11 cm); ii) dominant vegetative growth (11 cm < dbh < 28 cm); iii) cone production phase I (28 cm < dbh < 46 cm); iv) cone production phase II (46 cm < dbh < 62 cm), and v) cone production phase III (62cm < dbh < 82 cm). Average yearly cone weight production by tree j in plot i (CWP_{ij}) was measured in trees belonging to the three productive phases. Unfortunately, due to practical reasons in the field, it was not possible to record cone production data for the three production years to all plots, and thus two distinct datasets were established, included in Table 1 and Table 2. Cone weight production was evaluated from datasets of 416 trees (Table 1) and 76 trees (Table 2) in the productive stages, belonging to 51 and 9 plots, respectively. The dataset 1, summarized in Table 1, corresponds to all trees included in the groups of the productive stages, with cone production observed at least in one campaign, whereas the dataset 2 in Table 2 is a subset of the dataset of Table 1. This subset corresponds to 9 plots with cone production recorded in each one of the three production periods (plots 9, 10, 12, 14, 16, 43, 45, 46 and 47).

For dataset 1 the dependent variable in the modelling is the global cone weight production per tree for the three production periods. For dataset 2 the dependent variable is the cone weight production per tree for each production

period. Tree and plot covariables were used as independent variables. The two kinds of models can integrate intercept, slope, and nested random effects.

Table 1 - *Average and range for some variables of the dataset 1 for cone production (416 trees belonging to 51 plots)*

	Mean	Range
dbh (cm)	54	13–98
h (m)	14	8–20
ch (m)	8	3–14
bch (m)	6	0.4–14
cw (m)	12	8–23
G_T (m^2ha^{-1})	10	3–25
N_T (ha^{-1})	89	21–428
R (mm)	7777	6013-8452
CWP_1 (kg)	59	2–391
CWP_2 (kg)	18	1–110
CWP_3 (kg)	31	1–226
CWP_{to} (kg)	61	2–576

Table 2 - *Average and range for some variables of the dataset 2 for cone production (76 trees belonging to 9 plots)*

	Mean	Range
dbh (cm)	55	13–87
h (m)	15	9–20
ch (m)	8.6	6–13
bch (m)	6	2–11
cw (m)	14	9–23
G_T (m^2ha^{-1})	12	4–16
N_T (ha^{-1})	66	39–92
R (mm)	7483	6013-8452
CWP_1 (kg)	60	5-391
CWP_2 (kg)	21	2-84
CWP_3 (kg)	37	1-181
CTP_{to} (kg)	117	9-485

In Tables 1 and 2 the variables CWP_1, CWP_2, CWP_3 and CWP_{to} correspond to the average production per tree of the global production for each of the three production periods (1:2004-2005, 2: 2004-2005 and 3:2004-2005). The variable CWP_{to} corresponds to the average production per tree in the three periods. Data from local automatic weather stations (Portuguese Meteorological Institute), available online (http://snirh.apambiente.pt) allowed computing accumulated rainfall (*AR*) for five-year periods precedent to each of the three production periods.

The average cone production per tree in the dataset 1 was 61 kg, respectively. The average N_T was 89 trees ha^{-1}. The higher tree cone yield (over 200 kg/tree) was shown in bigger trees (*dbh* greater than 60 cm, *cw* above 12 m and *h* above 14 m). The higher cone production (in the range 3000 - 7000 kg ha^{-1}) occurred in stone pine plots with a G_T of 5 - 21 m^2 $ha^{-1}$and 40 - 176 trees ha^{-1}.

A logarithmic transformation (natural log) of cone weight production per tree allowed obtaining a Gaussian distribution (graph not shown) for the transformed data, more appropriate to fit linear (Gaussian) mixed models. Table 3 shows the sample correlation matrix of these data revealing that, as expected, some independent variables are strongly correlated, with correlation measures

greater than 0.7, reflecting multicollinearity between variables. These high correlations implicate a selection of the independent variables to fit with the linear mixed models for quantifying cone production.

The higher correlations, highlighted in Table 3, are: i) *h* with both *HD* and *bch* (0.81 and 0.77, respectively), ii) *c* with both *g* and *CA* (0.86 and 0.99, respectively), iii) *Dh* with G_T, G_{SP} and *bch* (0.87, 0.83 and 0.70, respectively), iv) *N* with N_T (0.91), v) G_T with both G_{SP} and *bch* (0.98 and 0.72, respectively, vi) *g* with *Ca* (0.86), vii) G_{SP} with *bch* (0.73).

.

Table 3 - *Correlation (symmetric) matrix for 13 covariables plus correlation of these covariables with the cone weight production (CWP)*

	dbh	***h***	***cw***	***HD***	***DD***	***N***	***G_T***	***g***	***CA***	***G_{SP}***	***N_T***	***bch***	***ch***
dbh	1	0.33	0.52	0.44	0.58	-0.56	0.16	0.47	0.50	0.09	-0.59	0.01	0.46
h		1	0.35	**0.81**	0.39	-0.08	0.66	0.50	0.33	0.61	-0.18	**0.77**	0.28
cw			1	0.24	0.65	-0.62	-0.07	**0.86**	**0.99**	-0.13	-0.59	-0.10	0.66
HD				1	0.56	-0.08	**0.87**	0.37	0.21	**0.83**	-0.17	**0.70**	0.10
DD					1	-0.63	0.34	0.67	0.62	0.28	-0.61	0.07	0.46
N						1	0.24	-0.52	-0.58	0.34	**0.91**	0.31	-0.57
G_T							1	0.10	-0.10	**0.98**	0.16	**0.72**	-0.14
g								1	**0.86**	0.05	-0.54	0.14	0.52
CA									1	-0.16	-0.56	-0.11	0.65
G_{SP}										1	0.21	**0.73**	-0.22
N_T											1	0.11	-0.43
bch												1	-0.39
ch													1
CWP_{ij}	0.45	0.31	0.71	0.18	0.25	-0.40	-0.12	0.57	**0.71**	-0.19	-0.34	-0.06	0.55

For the modelling of cone production with the dataset 1 (Table 1), we firstly made a screening analysis of independent biometrical variables, by least squares regression of two ensembles of 13 and 7 covariables and of 14 ensembles of four, three, two and one covariables for detecting the ensembles of independent covariables with lower collinearity (low variance inflation factor), and higher coefficient of determination between them. Table 4 displays collinearity and *AIC* statistics for the 14 ensembles of independent variables. The two ensembles with higher number of variables showed, as expected, high level of collinearity between covariables and were discarded for posterior analysis. To a subsequent selection, the 14 ensembles of variables were submitted to a LMM fitting process, under a two level clustered data structure, with an intercept term, using the SAS PROC MIXED (Version 9.1.3). This selection was based on the information criteria

(*-2log likelihood* and *AIC*) and the fixed and random parameters significance and on a general residual diagnostic (INFLUENCE option of PROC MIXED) as well.

Table 4 - *Variance Inflation Factor (VIF) and Akaike Information Criterion (AIC) statistics for dataset 1*

Model Covariables	VIF	AIC
dbh, *cw*, *h*, G_T	1.26	881
HD, N_{SP}, *g*, G_{SP}	2.07	929
dbh, *ch*, *CA*, *h*	1.37	911
db, *ch*, *bch*, *cw*	1.70	894
g, *ch*, *DD*, G_T	1.47	925
h, cw, G_T	1.18	873
db, *ch*, N_T	1.18	1010
h, G_{SP}, *ch*	1.35	978
cw, N_T	1.07	880
HD, G_T	1.12	1022
cw, G_T	1	879
cw	-	884
ch	-	1003
h	-	980

Modelling

Classic linear mixed models

For modelling cone production per tree with the first dataset we used classic LMMs. The methodology was the restricted maximum likelihood estimation (REML), summarized below (eqs. 1-10), implemented in SAS Proc MIXED (Version 9.1.3). LMMs, relying on asymptotic normality, are attractive due to their ability to deal with multiple levels of data clustering. As mentioned above, LMMs have become invaluable tools in the analysis of experimental and observational data of hierarchical, clustered, longitudinal or spatial kinds, allowing for more than one term to be submitted to random variation.

The general matrix equation for LMMs (SEARLE, 1971) can be written as:

$$y_i = X_i\beta + Z_i u_i + \varepsilon_i \ , \qquad (1)$$

where y_i represents the continuous response vector of n_i known observations (trees) for the i-th sampling unit (plot), X_i is the $(n_i \times p)$ design matrix of

known elements of observations related to p independent variables or covariables, β is the unknown vector of p regression coefficients (fixed effect parameters), Z_i is a $(n_i \times q)$ design matrix associated to random effects, whose terms may be ones, zeros or measured values for the q predictor variables for the i-th plot, u_i is a vector of q random effects and ε_i is a vector of n_i residuals for the i-th sampling unit. The number of observations for each plot, n_i, is not necessarily equal, $i = 1,...,51$.

Random vectors u_i and ε_i are assumed to have multivariate normal distribution:

$$u_i \sim N(0, D), \qquad \varepsilon_i \sim N(0, R_i) \tag{2}$$

where 0 is a null vector, D is the $(q \times q)$ symmetric positive-definite variance-covariance matrix of random effects, and R_i is the $(n_i \times n_i)$ symmetric positive-definite variance-covariance matrix for the residual matrix in the i-th subject (plot).

The REML estimation is based on an iterative optimization of the so called REML log-likelihood function (*e.g.*, WEST *et al*., 2007):

$$\begin{aligned} l_{REML}(\theta) = & -0.5(n-p)\ln(2\pi) - 0.5\sum \ln\det V_i - 0.5\sum r_i V_i^{-1} r_i \\ & - 0.5\sum \ln\det\left(X_i V_i^{-1} X_i\right) \end{aligned} \tag{3}$$

with:

$$r_i = y_i - X_i\left[\left(\sum X_i' V_i X_i\right)^{-1}\sum X_i' V_i Y_i\right] \tag{4}$$

where $n = \sum n_i$ is the total number of observations, θ is the vector of all model parameters and V_i is the $n_i \times n_i$ matrix corresponding to the implied marginal or population-averaged model for subject i:

$$Y_i = X_i\beta + \gamma_i \tag{5}$$

where the marginal residual vector is $\gamma_i \sim N(0, V_i)$, and the estimated form of variance-covariance matrix V_i $(n_i \times n_i)$ is given by:

$$\hat{V}_i = Z_i \hat{D} Z_i + \hat{R}_i \tag{6}$$

After obtaining an estimation of V_i, the following equations are used to estimate the fixed effects parameters and their standard errors:

$$\hat{\beta} = \left(\sum X_i' \hat{V}_i^{-1} X_i\right)^{-1}\sum X_i' \hat{V}_i^{-1} y_i \tag{7}$$

$$\mathrm{var}(\hat{\beta}) = \left(\sum X_i' \hat{V}_i^{-1} X_i\right)^{-1} \tag{8}$$

In this work, for dataset 1, a LMM was fitted considering the two-level hierarchy for independent variables, plot and tree, and the variance components and identity structures for covariance matrices for random effects and residuals, respectively.

The two criteria for model selection are the so-called information criteria that are mainly i) the –2 *res log-likelihood* criterion, $-2\mathrm{l}\left(\hat{\theta}\right)$, ii) the Akaike information criterion (*AIC*) (AKAIKE, 1974):

$$AIC = -2\mathrm{l}\left(\hat{\theta}\right) + 2\mathrm{p} \quad (9)$$

where l is the log-likelihood function, which is evaluated at the REML estimates of the model parameter vector ($\hat{\theta}$), and p is the total number of parameters, including the fixed and random effects. The best model is that for which calculations of the criteria give the smaller values. This methodology allows comparing any models fitted to the same data whether they are nested or not. The estimation of the random effects is given by the empirical best linear unbiased predictors, u_i (EBLUPS), (SEARLE, 1971):

$$\hat{u}_i \equiv \hat{E}\left(u_i \mid Y_i = y_i\right) = \hat{D}Z_i'\hat{V}\left(y_i - X_i\hat{\beta}\right) \quad (10)$$

Residual diagnostics in linear mixed models

Influence diagnostics of a fitted LMM allows verifying whether the distributional assumptions for the residuals are valid and whether the fitted model is sensitive to individual observations. The basic approach of this diagnostics methodology is to quantify the effects of an omission of a given subset of observations, in our case a plot (cluster unit), denoted by a subscript, *U*, on the global results of the analysis of the entire data. The model is refitted and statistics are computed based on the change between full data and reduced data estimation. The INFLUENCE option in SAS Proc Mixed (Version 9.1.3) allows the calculation of statistics for this kind of analysis and also the generation of panels of influence. This approach is similar to residual analysis applied to ordinary least squares analysis (*e.g.*, NETER *et al.*, 1996).

In this work, we used data and panels of unconditional and conditional studentized residuals, restricted likelihood distance, PRESS residuals, Cook's D, covratios, covtraces and MDFITTS for fixed and random effects.

Basically, unconditional and conditional residuals are defined respectively as:

$$\hat{\varepsilon}_m = y_i - X_i\hat{\beta} \quad (11)$$

$$\hat{\varepsilon}_c = y_i - X_i\hat{\beta} - Z_i\hat{u}_i \tag{12}$$

The residual studentization diminishes the impact of one problem of the raw residual, variance heterogeneity. The studentized form of these raw residuals is given by a scaling of the form (LITTELL *et al*., 2007):

$$\text{res}_i = \hat{e}_i / \sqrt{\hat{\sigma}^2(1-h_{ii})} \tag{13}$$

where an estimate of the variance of the residuals is placed in the denominator and h_{ii} is the i-th diagonal element of the leverage matrix (H):

$$H = X\left[X'V\left(\hat{\theta}\right)^{-1}X\right]^{-1}X'V\left(\hat{\theta}\right)^{-1} \tag{14}$$

The restricted likelihood distance (RLD) is given by:

$$RLD_{(U)} = 2\left|l_R\left(\hat{\psi}\right) - l_R\left(\hat{\psi}_{(U)}\right)\right| \tag{15}$$

where l_R is the restricted likelihood function, and ψ is an arbitrarily chosen vector of parameters, which can include parameters in β or θ . *RLD* reflects the change in REML log likelihood for all data, with ψ estimated for all data, $\hat{\psi}$, and reduced data, $\hat{\psi}_{(U)}$. To the definitions of PRESS residuals, Cook's D, covratios, covtraces and MDFITTS for fixed and random effects, refer to LITTELL *et al*., 2007.

Bayesian clustered longitudinal mixed models

For the second dataset, with cone production data available for the three production periods, three longitudinal models were fitted from a Bayesian perspective. Basically, the Bayesian analysis is based on a scaled product of a likelihood function $L(D\,|\,\theta)$, dependent on the available data D and the model parameter vector θ, by the prior distribution for θ, π(θ), representing the probability that θ takes any particular value based on prior information available. The result is the posterior distribution for θ, π(θ | D), expressing what is known about θ given the sample data.

For smaller sample sizes, Bayesian inference is especially advantageous over likelihood-based inference because the latter is related to asymptotic theory. Thus, Bayesian analysis can provide more reliable information than classic theory for LMMs, namely estimating the variance component of the random effects (PAULINO *et al*., 2003, and GELMAN *et al*., 2006).

The Bayesian clustered longitudinal mixed models, considered for the second dataset, were fitted taking on account random effects for clustered data with three levels of hierarchy: plot as level 3, tree as level 2, and production

period (for simplification designated as time) as level 1 (Figure 1). The difference between these models and the previous ones consisted in the incorporation of two additional random effect vectors, associated with trees (sampling units) nested within a plot (cluster), Eq. (16), and with plots sharing time slope random effects. Therefore, the longitudinal clustered mixed models are expressed as:

$$y_{ti} = X_{ti}\beta + u_{0i} + u_{1i}\,t + u_{2(i)} + \varepsilon_{ti} \qquad (16)$$

where y_{ti} represents the response vector of n_i observations (trees) for the *i-th* sampling unit (plot) at the *th*-production campaign, X_{ti} is the design matrix related to p possibly time-dependent covariables, u_{0i} and u_{1i} are random (plot) effect vectors associated with the intercept and time slope, respectively, $u_{2(i)}$ is the random effect vector associated with trees nested within a plot, and ε_{ti} is a vector of n_i residuals for the *i-th* sampling unit at the *th*-production campaign, $i = 1,...,9$, $t = 1,...,3$. The nested plot effects correspond to random effects associated to the plots, reflecting environmental or genetic variables not measured in our experimental design. Notice that, i) the random effects, u_{0i} and u_{1i}, were representative of variability inherent to the plot and to the production period, respectively, and ii) two sub-models from Eq. (16) are designed hereafter as (16*) and (16**) after removing the time slope (u_{1i}) and nested ($u_{2(i)}$) random effects and only the nested ($u_{2(i)}$) random effects, respectively.

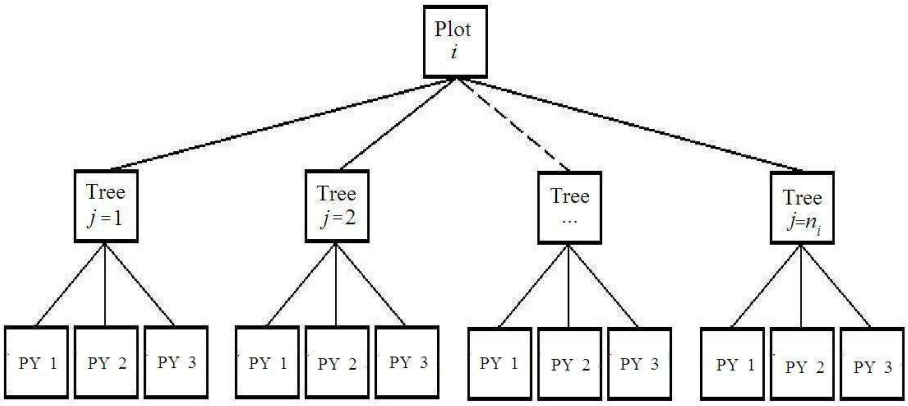

Figure 1 - *Structure of the clustered longitudinal data for the i-th plot in cone production data*

In this work, random vectors u_{0i}, u_{1i} and $u_{2(i)}$, and ε_{ti} are assumed as multivariate normally distributed:

$$u_{0i} \sim N(0, D_0),\ u_{1i} \sim N(0, D_1),\ u_{2(i)} \sim N(0, D_2),\ \varepsilon_{ti} \sim N(0, R_{ti}) \tag{17}$$

where D_0, D_1, and D_2 are the $(q \times q)$ symmetric positive-definite covariance matrix of the random effects u_{0i}, u_{1i} and $u_{2(i)}$, respectively, and R_{ti} is the symmetric positive-definite covariance matrix $(n_i \times n_i)$ for the residual matrix in the i-th subject (plot) at time point i, ($i = 1, \ldots, 9$, $t = 1, 2, 3$). These covariance matrices were considered:

$$R_{\mathrm{ti}} = \sigma^2 I, D_0 = \sigma_0^2 I, D_1 = \sigma_1^2 I, D_2 = \sigma_2^2 I \tag{18}$$

where I is the identity matrix, proportional to the appropriate Gaussian likelihood function times associated with prior distribution. We assumed also (independent) normal prior distributions for regression coefficients β with mean zero and variance v_k^2 (known), $k=0,\ldots,p$ (number of independent variables), and independent inverse gamma prior distributions for the variance components:

$$\sigma^2 \sim IG(a,b),\ \sigma_k^2 \sim IG(a_k, b_k),\ k=0,1,2 \tag{19}$$

Consequently, the joint posterior distribution of θ for longitudinal mixed model in Eq. (16) is expressed by

$$\pi(\theta \mid \mathrm{D}) \ \propto\ \mathrm{L}(\theta \mid D) \times \pi(\theta)$$

$$\propto \prod_{t=1}^{3}\prod_{i=1}^{9}\prod_{j=1}^{n_i} \sigma^{-1} e^{-\frac{(y_{tij}-\mu_{tij})^2}{2\sigma^2}} \times \prod_{k=0}^{1}\prod_{i=1}^{9} \sigma_k^{-1} e^{-\frac{u_{ki}^{\ 2}}{2\sigma_k^2}} \times \prod_{i=1}^{9}\prod_{j=1}^{n_i} \sigma_2^{-1} e^{-\frac{u_{2(i)}^{\ 2}}{2\sigma_2^2}} \times \tag{20}$$

$$\times \prod_{k=0}^{p} v_k^{-1} e^{-\frac{\beta_k^{\ 2}}{2v_k^2}} \times \sigma^{-2(a+1)} e^{-\frac{b}{\sigma^2}} \times \prod_{k=0}^{2} \sigma_k^{-2(a_k+1)} e^{-\frac{b_k}{\sigma_k^2}}$$

where $\theta = (\beta, \sigma^2, \sigma_0^2, \sigma_1^2, \sigma_2^2)$ and $\mu_{tij} = \mathrm{X_{ti}}\beta + \mathrm{u_{0i}} + \mathrm{u_{1i}}\,\mathrm{t} + \mathrm{u_{2(i)}}$ (response mean).

The posterior in Eq. (20) is awkward to work with, since the marginal posterior distributions of θ are not easy to obtain explicitly. Numerical methods can be employed in order to overcome that problem, *e.g.*, Markov chain Monte Carlo (MCMC) methods that are implemented in software such as OpenBugs (LUNN *et al.*, 2009). MCMC simulation involves simulating a sample from the joint posterior distribution of the model parameters using one of three main algorithms: Metropolis Hastings, Gibbs sampling and slice sampling. It is thereby possible to generate numerically, with a sufficient large sample, the posterior distribution from which inferences of interest can be made. For example, ZEGER and KARIN (1991) coped with generalized linear models under

a Bayesian framework using the Gibbs sampler to solve computational limitations, whereas FONG *et al*. (2010), confirmed the appealing Bayesian approach for LMMs by implementing prior distributions with variance components by the use of integrated nested Laplace approximations.

For the adequacy of fitness diagnostics of the Bayesian models, we used three criteria: i) the posterior mean of the deviance $\overline{D}$ (-2 log-likelihood function), ii) the deviance information criterion (*DIC*), which is a generalization of *AIC*, proposed by SPIEGELHALTER *et al*. (2002) and calculated as:

$$DIC = p_D + \overline{D} \tag{21}$$

where p_D is the effective number of the model parameters (a smaller DIC indicates a better model fitness), and iii) the conditional predictive ordinate (CPO, GELMAN *et al*., 2004), defined by the posterior predictive distribution:

$$p(y_i \mid y_{[i]}) = \int p(y_i \mid \theta) p(\theta \mid y_{[i]}) d\theta \tag{22}$$

where y_i is the observation i of the data y, and $y_{[i]}$ is y without the current observation i, being $p(y_i \mid \theta)$ a predictive distribution and $p(\theta \mid y_{[i]})$ a joint posterior distribution for the model parameter θ. The CPO is based on the leave-one-out cross-validation predictive density, and is easy to obtain with MCMC methods. It may be used to identify outliers, influential observations, and for hypothesis testing across different non-nested models.

Results and discussion

Cone production by LMMs

The selected two level clustered LMM model with an interception term (cone weight production by tree) was the following:

$$\ln CWP_{ij} = \beta_0 + \beta_1 cw_{ij} + \beta_2 G_{T_i} + \beta_3 h_{ij} + u_i + \varepsilon_{ij} \tag{23}$$

where $lnCWP_{ij}$ is the natural logarithm of the weight cone total production of the j_{th} tree belonging to the i_{th} plot (i=1,...,51) and the other tree and plot variables correspond to crown width, basal area by hectare and tree height, respectively. The terms u_i and ε_{ij} correspond respectively to random intercepts, specific to each plot, and residuals, defined above (Section 2.2.1). The estimated parameters for this model [Eq. (23)] are shown in Table 5.

Table 5 *- Interception linear mixed model (REML) for logarithm model (cone production (kg)). Fixed effects; Random effects; Information criteria*

	Parameter	Estimate (Sd. Error)	*p*-value
Fixed effects			
Intercept	β_0	1.42(0.38)	<0.0001
cw	β_1	0.16(0.015)	<0.0001
G_T	β_2	-0.09(0.021)	0.0007
h	β_3	0.08(0.02)	0.0006
Random effects			
Intercept	Plot	0.47(0.11)	<0.0001
Residual		0.33(0.03)	<0.0001
Information criteria			
–2 res log likelihood		869.4	
AIC (Akaike information criteria)		873.4	

The selected LMM [Eq. (23)] showed that two tree variables, the crown width (cw_{ij}) and the tree height (h_{ij}), and a plot variable, the total basal area by hectare (G_{Ti}), were highly influential variables to predict cone production. The dependent variables cw_{ij} and G_{Ti} are easy to obtain. Measurement of tree height, a variable associated to site index, is not so practical, leading to the need of modelling height-diameter relationship.

The average influence residual statistics for the selected LMM are shown in Table 6. These statistics are in reasonable agreement with reference values of zero for Cook'D, MDFFITS and Covtrace and PRESS residuals, and of one for Covratio. The restricted likelihood distance, with the average value of 0.25 and a standard deviation (Table 6) of 0.009 fits well with a reference value, corresponding to the 75th percentile for a chi-square distribution with six degrees of freedom (total number of fixed and random parameters in the interception model) of about 10. This means that the log likelihood estimation of the full data is not affected by data sampling reduction (WEST *et al*., 2007). Figure 2 depicts Covratio statistics for fixed effects and covariance parameters. Although the PRESS statistic shows deviation from zero, its graphical display (not shown) is in reasonable agreement with some references (*e.g.*, WEST *et al*., 2007; LITTELL *et al*., 2007). Conditional studentized residuals (Table 6 and Figure 3), follow the typical Gaussian distribution.

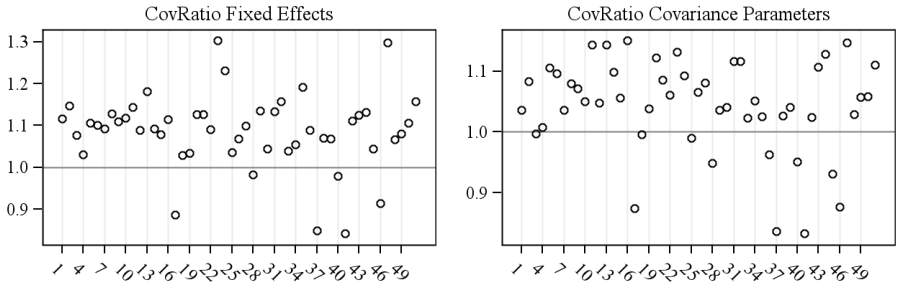

Figure 2 - *Influence statistics, interception model*

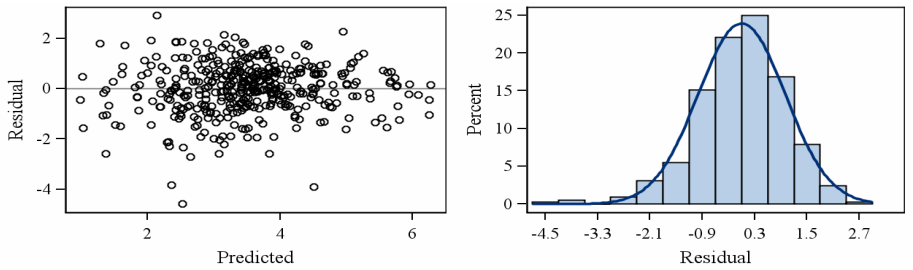

Figure 3 - *Conditional studentized residuals, interception model*

Table 6 - *Residual analysis of interception linear mixed model (REML) for ln (cone production (kg)) - Statistics, Studentized residuals, Conditional studentized residuals*

	Average	Standard error
Cook'D (fixed effects)	0.04	0.001
Cook'D (covariance parameters)	0.10	0.005
Covratio (fixed effects)	1.09	0.002
Covratio (covariance parameters)	1.04	0.001
MDFFITS (fixed effects)	0.03	0.001
MDFFITS (covariance parameters)	0.10	0.001
Covtrace (fixed effects)	0.10	0.070
Covtrace (covariance parameters)	0.07	0.007
Restricted likelihood distance	0.25	0.009
Press statistic	4.11	0.600

	Average	Range	Standard deviation
Studentized residuals	0.020	-3.46 – 2.36	0.97
Conditional studentized residuals	0.001	-2.52 – 1.57	0.55
	Mean error	RMSE	Model efficiency
Additional statistics	0.001	0.54	77 %

The positive values of β_1 (associated to c) and β_3 (associated to h) and the negative β_2 (associated to G_T) are indicative that in general terms, bigger trees with larger crowns and lower plot density and competition are the ones with greater cone production. This coincides with conclusions in previous works (*e.g.*, CAÑADAS, 2000; ALPUÍM *et al*., 2000; CALAMA, 2004, CALAMA *et al*., 2008; FREIRE, 2009). ALPUÍM *et al*. (2000), for Portuguese stone pine stands, noted that under similar ecological conditions stone pine fructification depends mainly on the crown diameter.

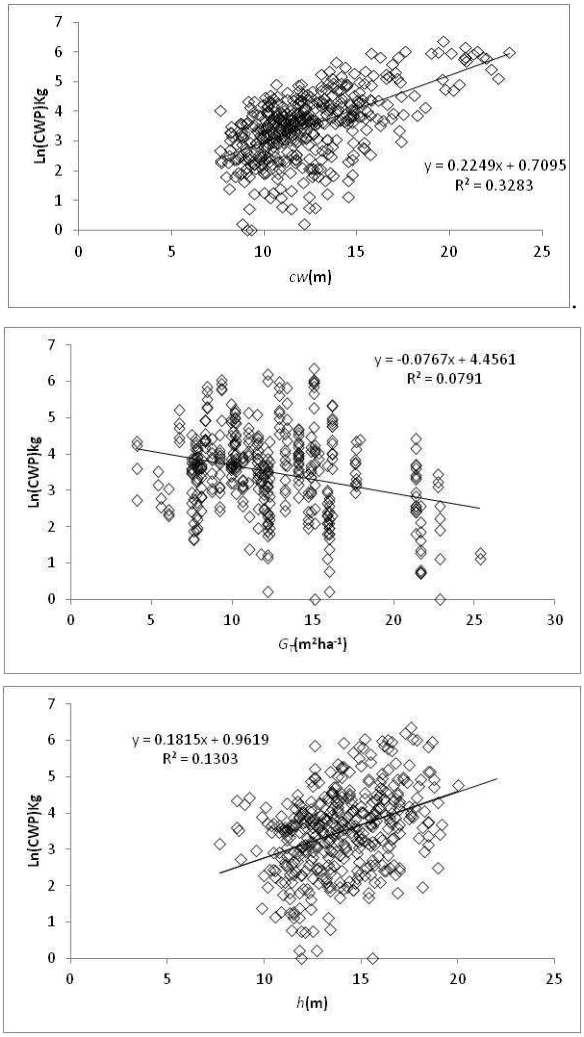

Figure 4 - *Variation of cone production with crown width, basal area and tree height*

The selected model (Eq. 23), reflected the cone production variability by single least squares regression with its dependent variables (Figure 4). Indeed, the more productive plots were originated from natural regeneration, managed to cone production since the beginning of productive cycle, with thinning operations to low tree densities, irregular distribution, allowing for crown development and diameter at breast height enlarging with the canopy area covering about 53 % to 73 % of forest surface. Plot randomness explained 58% of global variability (Eq. 25) and therefore the inclusion of random effects associated with plots, allowed to improve the quality of estimates. Given that the experimental plots were located in the same productive region (Setúbal Peninsula), thus minimizing regional effects, we think that factors such as distinct productive stages, environmental site conditions, distinct topography, allowing for water accumulation by gravity, same genetic variation due to predominant natural regeneration, and also the drought anomaly in 2004/2005 (RODRIGUES *et al*., 2011) explained most of the global heterogeneity in cone production. The precipitation accumulated in five-year periods, prior to each of the three production periods, evidenced no significant effect on global cone production, circumstance due, from our point of view, to the fact that the global dataset included all data irrespective of the production year. Eq. 23 can be used to obtain estimates of plot specific cone production in a three-year period, using the calculated EBLUPS per plot. The same equation can be used to obtain population-averaged production in the same period to all productive area, by omitting random effects.

Cone production by Bayesian clustered longitudinal analysis

For the second dataset (Table 2) related to three production periods, several Bayesian linear mixed models [in Eq. (16), including (16*) and (16**)] were fitted assuming flat prior distribution [Eq.(17)- (19)] with $a = a_k = b = b_k = 0.001$, $k=0,1,2, v_j^2 = 10^6$, $j=0,1,\ldots p$, since such values avoid an excess of information for a non-informative setting. Besides, the results of this Bayesian analysis were obtained through the software OpenBugs (LUNN *et al*., 2009) after 50000 iterations of simulation and 2000 iterations of burn-in.

Table 7 displays a model comparison based on DIC and Deviance posterior mean for clustered longitudinal data. The second column in Table 7 is related to no-time slope and no-nested clustered longitudinal model (16*), whereas the third column reports no-nested clustered longitudinal model (16**). In the last

two columns are indeed the fitting results of the comprehensive three-level clustered longitudinal models [Eq. (16)]. In that table, there are the better fitting longitudinal mixed models, i.e., models M_1, M_2, M_3 and M_4 that were defined from sets of the covariables: *dbh*, *h*, *cw*, *HD*, *DD*, *N*, G_T, *g*, *CA*, G_{SP}, N_T, *bch*, *ch*, *time* (production period) and *AR*. DIC and Deviance values indicate that linear model in Eq. (16) is better fitted, and the former values point out in favour of M_3 for both model comparison measures. Notice that DIC takes into account the complexity of the model, which is an important issue in mixed models, and the selected model M_3 [Eq. (16)] has better fitting even than parsimonious model M_4.

Table 7 - *Model comparison based on DIC and Deviance for clustered longitudinal data*

LMM	DIC (16*)	Deviance (16*)	DIC (16**)	Deviance (16**)	DIC (16)	Deviance (16)
M_1: *cw*, *time*, *AR*	516.45	504.37	474.75	455.48	445.01	392.76
M_2: *cw*, *time*	514.56	503.37	482.62	464.65	463.66	417.22
M_3: *cw*, *AR*	529.88	518.51	471.65	452.34	441.08	388.22
M_4: *cw*	535.39	525.31	483.28	464.88	464.53	417.82

The selected model, representative of average annual production in the *t-th* campaign (time) of the *j-th* tree belonging to the *i-th* plot, (i = 1, …, 9, j=1,…, 76, t = 1, 2, 3) is expressed by

$$lnCWP_{tij} = \beta_0 + \beta_1 cw_{ij} + \beta_2 AR_{i_t} + u_{0i} + u_{1i}t + u_{2(i)j} + \varepsilon_{tij} \quad (24)$$

where $lnCWP_{tij}$ is the natural logarithm of the weight cone production for tree j in plot i at production period t, and the other variables correspond to crown width (cw_{ij}), five -year accumulated rainfall (AR_t), time t (for convenience, assuming values -1, 0 and 1), and u_{0i}, u_{1i}, and $u_{2(i)j}$ are the random intercept, time slope and nested (plot) effects, and ε_{tij} is the residual term, i=1,…, 9, j=1,…, 76, t=1,2,3. This model also showed the complexity of the cone production process, by including random factors associated to the production period, related to the variability between production years, and nested plot effects, related to the local variables of stand, environmental or genetic type associated to the plots themselves, not measured under our experimental design. Posterior estimates based on the selected clustered longitudinal model [Eq. (24)] are shown in Table 8: posterior mean, standard deviation (s.d.), median and 95% credible interval of some random parameters of interest. The selected covariables, crown width (cw_{ij}) and five-year accumulated rainfall (AR_i), have a positive influence on the

weight cone production, indicating that a greater discrepancy in cw_{ij} and AR_i is associated with a higher mean value of $lnCWP_{tij}$. The variable cw_{ij} is common to the two selected models in Eq.(23) and Eq.(24), with a coefficient of the same order of magnitude, reflecting the relevance of crown width in cone production. The rainfall was relevant on longitudinal data of cone production per production period, reflecting the fact that globally, weight cone production in all plots encompassed by this study, was higher in 2004/05 (4578 kg), decreasing substantially in 2005/06 (1601.8 kg), and recovering partially in 2006/07 (2821 kg). The reduction in 2005/06 may reflect the aforementioned drastic diminution in rainfall occurred in the period 2004/05.

In Table 8 variance component estimates were obtained for the selected model [Eq. (31)]. These include the variance of the intercept (σ_0^2), time slope (σ_1^2) and nested (σ_2^2) random effects, apart from the variance of the residual terms (σ^2). There is some unobserved heterogeneity at the three levels of data in decreasing order (Table 8) by overall plot effect (0.917), time plot effect (0.52) and individual tree within plots effect (0.321), and some residual heterogeneity among trees (0.568). The intraclass correlation coefficient (ICC) (WEST *et al*., 2007) is calculated from the several estimated "total" variance components as follows:

$$\sigma^2_{Tot\_012}=\frac{\sigma_0^2+\sigma_1^2+\sigma_2^2}{\sigma^2+\sigma_0^2+\sigma_1^2+\sigma_2^2},\ \sigma^2_{Tot\_0}=\frac{\sigma_0^2}{\sigma_0^2+\sigma_1^2+\sigma_2^2},\ \sigma^2_{Tot\_1}=\frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2+\sigma_2^2}, \quad (25)$$

Table 8 - *Posterior estimates for the selected clustered longitudinal model [Eq. (24)] mean, standard deviation (s.d.), median and 95% credible interval (CI)*

	M_5	**mean**	**s.d.**	**median**	**2.5% CI**	**97.5% CI**
intercept	β_0	-2.717	1.3330	-2.678	-5. 278	-0.131
cw	β_1	0.166	0.0234	0.166	0.121	0.213
AR	β_2	0.0014	0.0005	0.0014	0.0005	0.0024
residual	σ^2	0.568	0.0360	0.566	0.499	0.639
intercept	σ_0^2	0.917	0.3340	0.857	0.386	1.584
slope	σ_1^2	0.520	0.1630	0.491	0.262	0.844
nested	σ_2^2	0.321	0.0680	0.323	0.187	0.456
ICC_1	$\sigma^2_{Tot\_012}$	0.772	0.0893	0.576	0.597	0.936
ICC_2	$\sigma^2_{Tot\_0}$	0.659	0.1523	0.317	0.350	0.920
ICC_3	$\sigma^2_{Tot\_1}$	0.244	0.1303	0.060	0.037	0.507

Thereby it can be inferred that most of the total variation in the weight cone production per tree is dominated by the random plot effects (0.772), comparing with 0.7 from LMM in eq. 23, and this random variation is due to overall among-plot (65.9%) and time slope among-plot (24.4%) effects.

For assessment of mixed model in Eq. (24), CPO [Eq. (22)] estimates were plotted in Figure 5 (left side), ranging from 0.000001 to 0.2215, depicting no outliers and influential values for all three production periods. In addition, conditional raw residuals, Eq. (11), were obtained and plotted in Figure 6 suggesting no specific model fitting problem both from its histogram (left side) and from its comparison with corresponding predictive values (right side). In Figure 5 (right side) there is a graphical representation of time slope random effect estimates and 95% credible intervals pointing out that plots 10, 16, 45, 46, and 47 have significant random effects over time. In these plots, an average N_T of 91 ha^{-1}, above a global average of 66 ha^{-1}, indicates an occupation of mainly younger trees, and thereby the occurrence of significant time random effects reflects a higher longitudinal variability in younger stands.

On the other hand, Eq.(24) includes nested plot effects, associated to the plots, not measured and probably associated to micro-environmental and genetic factors, reflecting their relevance to the cone production process

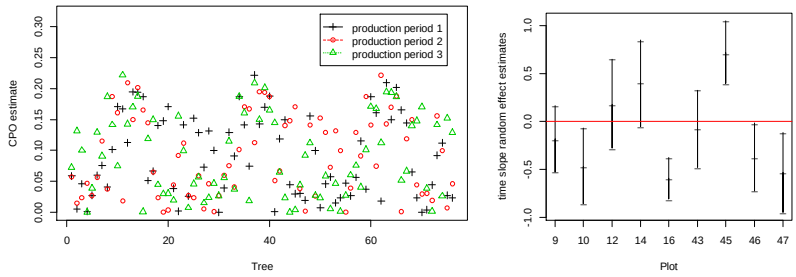

Figure 5 - *Conditional Predictive Ordinate estimates and 95% credible intervals of the time slope random effects (u_{1i}) for the selected clustered longitudinal model [Eq.(24)]*

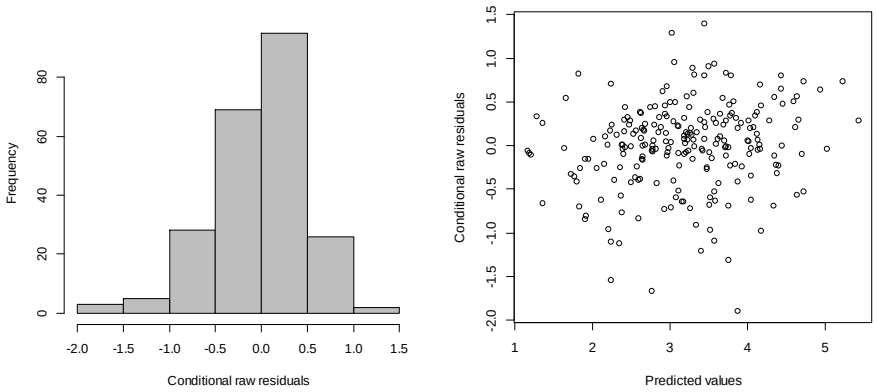

Figure 6 - *Conditional raw residuals histogram (left) and those residuals versus predicted values (right) for the selected clustered longitudinal model [Eq. (24)]*

Conclusions

This work aimed to develop linear mixed models under classic and Bayesian approaches for cone pine production in the Portuguese Setúbal Peninsula, based on cone weight production data in stone pine plots during three production periods. The two-level clustered linear model was estimated from all cone production data available for three consecutive years, including plots from which the production data collection was not possible for every year of the period considered. This model included crown width, total basal area per hectare and tree height as dependent variables determinant for cone production. The Bayesian three-level clustered longitudinal linear model was obtained from a data subset, referring to plots from which production data collection was possible in all the three productive periods. In this model, crown width and five-year accumulated rainfall, prior to each production period were included as the independent variables determinant for cone production. The complexity of the cone production process in the field was reflected by the inclusion in the Bayesian equations of random factors associated to the production period, and to nested plot effects, related to the local environmental or genetic variables of the plots. These variables should be considered under a throughout experimental design. The variables selected for both models agree with results from other studies on stone pine cone production. In both models the intra-

cluster variance, expressing the randomness associated to plot effects, significantly contributes to more than 58% of the global variability and thereby justifies an experimental design based on data clustering in plots and production years.

With a dataset limited to only three production periods, the aimed objectives had only a preliminary significance, given the fact that long term empirical evidence suggested that cone production follows a 7–8 year campaign cycle. Despite the small period of data collection the results obtained are indicative about the tree and stand variables adequate for proper forest management aiming at cone production. The results obtained are still conjectural for the Portuguese stone pine forest, and an extension both of measurement period and other productive zones are obviously needed for a more thorough and representative analysis.

Acknowledgements

This study was supported by the AGRO 451 project "Stone Pine Stand Improved and Management Optimization for Cone and Seed Production". Giovani L. Silva was partially supported by Pest-OE/MAT/UI0006/2011.

References

AKAIKE, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6): 716–723.

ALPUÍM, M., BAETA, J., CARNEIRO, M.M., CARVALHO, M.A., ROCA, M.E., PESSOA, J., 2000. *Classification of stone pine selected for pine kernels production, IUFRO International Meeting*. May, Sevilla, Spain.

CALAMA, R., 2004. *Modelo Interregional de Silvicultura para Pinus pinea L. Aproximación mediante funciones con componentes aleatorios*. Tesis Doctoral, Escola Técnica Superior de Ingenieros de Montes, Universidade Politécnica de Madrid 307pp.

CALAMA, R., MONTERO, G., 2005. Cone and seed production from stone pine (Pinus pinea L.) stands in Central Range (Spain). *Eur. J. Forest Res.* **126**(1): 23–35.

CALAMA, R., MONTERO, G., 2006. Stand and three-level variability on stem form and tree volume in Pinus pinea L.: a multilevel random components approach. Invest. Agrar: Sist. Recur. For. 15(1), 24–41.

CALAMA, R., MADRIGAL, G., CANDELA, J.A., MONTERO, G., 2007. Effects of fertilization on the production of an edible forest fruit: stone pine (*Pinus pinea* L.) nuts in south-west Andalusia. Invest. *Agrar: Sist. Recur. For.* **16**(3): 1–12.

CALAMA, R., GORDO, F.J., MUTKE, S., MONTERO, G., 2008. An empirical ecological-type model for predicting stone pine (Pinus pinea L.) cone production in the Northern Plateau (Spain). *Forest Ecology and Management 255*: 660–673.

CAÑADAS, M.N.C., 2000. *Pinus pinea L. en el Sistema Central (valles del Tiétar y del Alberche): desarrollo de un modelo de crecimiento y producción de piña*. Ph. Dr. Thesis. Universidad Politécnica de Madrid, 356pp.

CARRASQUINHO, I., FREIRE, J., RODRIGUES, A., TOMÉ, M., 2010. Selection of *Pinus pinea* L. plus tree candidates for cone production. *Ann. For. Sci.* **67**: 814. DOI:10.1051/forest /2010050.

FONG, Y., RUE, H., WAKEFIELD, J., 2010. Bayesian inference for generalized linear mixed models. *Biostatistics* **11**(3): 397–412.

FREIRE, J., 2009. *Modelação do crescimento e da produção de pinha no pinheiro manso*. PhD Thesis, Instituto Superior de Agronomia, Universidade Técnica de Lisboa.

GELMAN, A., CARLIN, J., STERN, H., RUBIN, D., 2004. *Bayesian Data Analysis*. 2nd edition. Chapman & Hall, Boca Raton, FL.

GONÇALVES, A.C., POMMERENING, A., 2012. Spatial dynamics of cone production in Mediterranean climates: A case study of *Pinus pinea* L. in Portugal. *Forest Ecology and Management* **266**: 83–93.

GARCÍA GÜEMES, C., 1999. *Modelo de simulación silvicola para Pinus pinea L. en la província de Vallalodid*. PhD Thesis. Escuela Técnica Superior de Ingenieros e Montes. UPM. Madrid.

ICNF, 2013. Inventário Florestal Nacional (Portugal Continental) Lisboa. http://www. icnf.pt/ portal/florestas /ifn/ifn6.

KOENIG, W.D., KNOPS, J.M.H., CARMEN, W.J., STANBACK, M.T., MUMME, R.L., 1996. Acorn production by oaks in central California: influence of weather at three levels. *Can. J. For. Res.* **26**: 1677–1683.

LITTELL, C.R., MILLIKEN, G.A., STROUP, W.W.; WOLFINGER, D.W., SCABENBERGER, O., 2007. *SAS for Mixed Models*. Second Edition. SAS Institute Inc., 813pp.

LUNN, D., SPIEGELHALTER, D., THOMAS, A., BEST, N., 2009. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine* **28**: 3049–3067. OpenBugs version 3.2.1, 2011: http://www.openbugs.info/w/.

MUTKE, S., GORDO, F.J., GIL, L., 2000. Seleccion de indivíduos de *Pinus pinea* L. grandes productores de fruto en las masas de la meseta norte. In: *1er. Simposio del pino pinonero (Pinus pinea L.)*. 22- 24 de February 2000, Valladollid. Edited by Junta de Castilla Y León Tomo II, 85–93.

MUTKE, S., GORDO, F.J., GIL, L., 2005. Variability of Mediterranean stone pine cone production: yield loss as response to climatic change. *Agric. For. Met.* **132**: 263–272.

NETER, J., KUTNER, M.H., NACHTSHEIM, C.J., WASSERMAN, W., 1996. Applied linear statistical models,. WCB/McGraw Hill, 1408pp.

PAULINO, C.D., AMARAL TURKMAN, A., MURTEIRA, B., 2003. *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisbon, 446 pp.

SEARLE, S.R., 1971. *Linear Models*. John Wiley & Sons, 532 pp.

SEARLE, S.R., CASELLA, G., MCCULLOCH, 2006, *Variance components*. John Wiley & Sons, 501 pp.

SPIEGELHALTER, D.J., BEST, N.G., CARLIN, B.P., VAN DER LINDE, A., 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal *Statistical Society* B **64**(4): 583–639.

SUROVÝ, P, RIBEIRO, N., PEREIRA, J.S., 2011. Observations on 3-dimensional crown growth of stone pine. *Agroforestry Systems* **82**: 105-110.

WEST, B.T., WELCH, K.B., GALECKI A.T., 2007. *Linear mixed models. A practical guide using statistical software*. Taylor and Francis Group, 353pp.

ZEGER, S.L, KARIN, M.R., 1991. Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc*. **86**: 79-86.

Submitted for publication in May 2014
Accepted in June 2014