Pacote *pedigreemm* e sua utilização no melhoramento vegetal

Danielle Silva Pinto[1], Rafael Simões Tomaz[2], Jacqueline Siqueira Glasenapp[2], Regiane Abjaud Estopa[3], Marcos Deon Vilela de Resende[4], Cosme Damião Cruz[5].

Resumo

Com intuito de ajustar o modelo de análise genética de delineamento de progênies de meios irmãos com a utilização do pacote *pedigreemm* via software R, elaborou-se um script que permite ao pesquisador obter a estimação dos componentes de variâncias e predição de parâmetros genéticos via REML/BLUP. O experimento consistiu no delineamento de blocos ao acaso com uma única planta por parcela (Single Tree Plot- STP), com 30 repetições, 122 progênies de meios irmãos de *Eucalyptus urophylla*, no espaçamento 3,0 x 2,5 m. A variável estudada foi altura (m). A partir de modificações realizadas no arquivo de análise e utilização da equação de correção, aplicada na saída original do pacote, foi possível a obtenção do REML/BLUP corretos para a referida análise.

Introdução

No melhoramento florestal, em que a população de melhoramento possui longos ciclos de avaliação e a medição sempre ocorre mais de uma vez no mesmo indivíduo, além de ocorrer perdas nas parcelas por diversos motivos, a utilização dos modelos mistos através do REML/BLUP (máxima verossimilhança restrita e melhor predição linear não viesada) tem sido o procedimento mais adequado para a seleção de indivíduos superiores por levar em consideração essas peculiaridades.

Na modelagem genética - estatística softwares como o R (R Development Core Team, 2012) tem sido amplamente utilizado para o desenvolvimento de pacotes relacionados a essa modelagem. Diversos pacotes já foram criados e para a utilização de modelos lineares mistos voltados a estimação de componentes de variância e predição de valores genéticos, pode se citar o *pedigreemm* (Vazquez et al., 2009), que foi desenvolvido para o modelo animal utilizando a matriz de parentesco entre os indivíduos da população. Entretanto, a utilização do *pedigreemm* para o melhoramento vegetal depende de algumas modificações no arquivo de dados e correções no ajuste do modelo.

Com isso, objetivou-se a elaboração de um script no software R para a avaliação de famílias de meios irmãos de *Eucalyptus urophylla* S.T. Blake, através da metodologia REML/BLUP, utilizando-se o pacote *pedigreemm*.

Material e Métodos

Utilizou-se dados do programa de melhoramento florestal da empresa Klabin S.A. de teste de progênies de meios irmãos de *Eucalyptus urophylla*, plantado no ano de 2006, com medição de 4 anos de idade, instalado na fazenda da empresa em Telêmanco Borba, Paraná.

O experimento consistiu no delineamento de blocos ao acaso com uma única planta por parcela (Single Tree Plot- STP), com 30 repetições, 122 progênies testadas provenientes de uma área de produção de sementes, no espaçamento 3,0 x 2,5 m. A variável estudada foi altura (m).

Para a utilização do pacote *pedigreemm* com os dados de progênies de meios irmãos é necessária a criação de linhas adicionais no início do arquivo, que serão referentes aos genitores das progênies avaliadas. No final do arquivo deve-se duplicar a última linha.

[1] Doutoranda em Genética e Melhoramento - UFV. Av. Peter Henry Rolfs, s/n, Campus Universitário, 36570-000- Viçosa -MG. E-mail: daniamazon@gmail.com;

[2] Pós-Doutorando (a) em Genética e Melhoramento- UFV. Av. Peter Henry Rolfs, s/n, Campus Universitário, 36570-000- Viçosa –MG. E-mails: rafaelst@gmail.com; siqueiragaia@yahoo.com.br

[3] Melhorista da empresa Klabin S.A. E-mail: raestopa@gmail.com

[4] Pesquisador EMBRAPA Floresta. Prof. Credenciado da Universidade Federal de Viçosa – UFV. Av. Peter Henry Rolfs, s/n, Campus Universitário, 36570-000- Viçosa -MG. E-mail: marcos.deon@gmail.com

[5] Prof. Titular Departamento de Biologia Geral – UFV. Av. Peter Henry Rolfs, s/n, Campus Universitário, 36570-000- Viçosa -MG. E-mail: cdcruz@ufv.br

A estimação/predição foi obtida pelo procedimento ótimo de avaliação genética REML/BLUP. O modelo estatístico adotado:

Em que: ***y*** é o vetor de dados; é o vetor dos efeitos de repetição (fixos) somados a média geral, é o vetor dos efeitos genéticos aditivos individuais (aleatórios), ***e*** é o vetor de erros ou resíduos (aleatórios). As letras ***X*** e ***Z*** representam as matrizes de incidência para os referidos efeitos, respectivamente.

As equações de modelos mistos são:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda_1 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Em que $\lambda_1 = \frac{1-h_a^2}{h_a^2}$; h_a^2: $\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$; σ_a^2: variância aditiva; σ_e^2: estimativa da variância residual

As distribuições e estruturas de médias e variâncias são: $y|b, V \sim N(Xb, V)$; $a|A, \sigma_a^2 \sim N(0, A\sigma_a^2)$ e $|\sigma_e^2 \sim N(0, I\sigma_e^2)$.

Resende et al. (2012) desenvolveram as correções para os valores genéticos preditos dos indivíduos gerados pelo *pedigreemm*, para progênies de meios irmãos, conforme a seguir:

$$\hat{a}_i = \hat{a}_{iR} + \hat{a}_{gen-incorreto} \frac{(3/4)h_a^2}{(4-h_a^2)} + \hat{a}_{gen-correto} \frac{(1/2)(1-h_a^2)}{(1/4)(4-h_a^2)}$$

Em que:

$\hat{a}_i$: vetor de valores genéticos corretos preditos dos indivíduos;

$\hat{a}_{iR}$: vetor de valores genéticos incorretos dos indivíduos preditos pelo *pedigreemm*;

$\hat{a}_{gen-correto}$: vetor de valores genéticos corretos dos genitores;

$\hat{a}_{gen-incorreto}$: vetor de valores genéticos incorretos dos genitores preditos pelo *pedigreemm*.

Resultados e Discussão

Com as devidas modificações propostas no arquivo de dados para a utilização do pacote *pedigreemm* foi possível à obtenção dos componentes de variância e dos BLUP's de cada indivíduo. A estimação dos componentes de variância via REML pode ser visualizado a seguir (Tabela 1).

Tabela 1: Estimação de componentes de variância via REML

Linear mixed model fit by REML				
Formula:	alt ~ -1 + Bloc + (1|Id)			
Data:	dados			
AIC	BIC	logLik	deviance	REMLdev
28315	29288	-14002	28827	28005
Random	effects:			
Groups	Name	Variance	Std.Dev.	
Id	(Intercept)	8.8087	2.9624	
Residual		84.7192	9.2043	

As estimações dos parâmetros genéticos e ambientais foram realizadas conforme Resende (2007b), levando-se em consideração 100% sobrevivência (Tabela 2).

Tabela 2: Estimativas dos parâmetros genéticos e ambientais da análise de teste de progênies de meios irmãos de *Eucalyptus urophylla*

Estimativas dos parâmetros variável Altura	
var.fen	93.4947
h2.ind	0.0942
h2.mp	0.4197
acurácia	0.6478
PEV	1.2778
SEP	1.1304
CV.gen.ind	20.6501
CV.gen.prog	10.3250
CV.res	66.4912
Média geral	14.3724

var.fen: variância fenotípica; h2.ind: herdabilidade individual; h2.mp: herdabilidade média de progênies; h2.ajust: herdabilidade ajustada; PEV: variância do erro de predição; SEP: desvio padrão do erro de predição; CV.gen.ind: coeficiente de variação genético individual ;CV.gen.prog: coef. de variação genético de progênie; CV.res: coef. de variação residual e média geral do experimento para altura.

O vetor de efeitos fixos é constituído por valores genéticos de cada genitor, seguidas pelas médias dos blocos. Isso é decorrente da utilização das linhas iniciais, que representam as informações dos genitores de cada progênie. Os blocos individuais por genitor (associada a cada linha criada) tem a função de fornecer os devidos efeitos de cada genitor e também serão utilizados na equação de correção para a predição dos valores genéticos corretos por indivíduo através de .

As predições geradas pelo pacote ($\hat{a}_{gen-incorreto}$) e a correção de cada predição por indivíduo ($\hat{a}_{gen-correto}$) podem ser visualizadas conforme Tabela 3.

Tabela 3: Predição do valor genético aditivo dos 20 primeiros genitores, após correção do resultado do pacote *pedigreemm*.

Genitores	$\hat{a}_{gen-incorreto}$	$\hat{a}_{gen-correto}$
1	1.702941797	0.91222932
4	4.167844789	2.23262487
8	2.880675574	1.54311599
12	-4.91529668	-2.63301878
13	-1.038155399	-0.55611753
16	3.83432726	2.0539667
17	2.052092961	1.09926209
18	2.896309208	1.55149059
19	-1.236181432	-0.66219582
20	5.074595573	2.71835177
21	5.908389397	3.1649972
25	3.05264555	1.63523661
29	-0.381542762	-0.20438426
31	-0.949564805	-0.50866146
32	-3.904321668	-2.0914612
16	-2.278423712	-1.22050261
17	-2.486872168	-1.33216397
18	1.093230064	0.58561985
19	2.958843745	1.58498899
20	-2.580673973	-1.38241158

Script de Análise

```
rm(list=ls(all=TRUE))

library(pedigreemm)
setwd("C:\\Dani_Pinto\\FMI - 22.02.13")       # definição da trilha de dados
dados=read.table("MODELO2_FMI.txt",h=T        # 1 planta por parcela

# o arquivo deve ser construído da seguinte maneira:
# Id : Identificação do indivíduo - numérico, de 1 até o número de indivíduos.
# Prog :
num.individuos <- nrow(dados)
num.familias <- length(unique(dados$Prog))
num.blocos <- length(unique(dados$Bloc))
                  num.variaveis <- ncol(dados)-3
                  col.parc <- 0

## Organizando a tabela de dados
colnames(dados)[1:2]<-c("Id","G1") # nomeando as colunas
G2<- rep(NA,num.individuos)
dados<-cbind(dados,G2)
if (sum(colnames(dados)=="Parc")==TRUE) {
         dados<-(dados[,c(1:2,ncol(dados), (ncol(dados)-(2+num.variaveis)) : (ncol(dados)-1) )])
         } else {
                  dados<-(dados[,c(1:2,ncol(dados), (ncol(dados)-(1+num.variaveis)) : (ncol(dados)-1) )]) }

## Criação das linhas iniciais do arquivo de dados modificado
Id<-seq(1:num.familias); G1<-rep(NA,num.familias); G2<-rep(NA,num.familias)
Bloc<-seq(1:num.familias);

variaveis <- matrix(data= (0), nrow = num.familias, ncol = num.variaveis)
colnames(variaveis) <- colnames(dados[(5+col.parc):ncol(dados)]) # pega os nomes das colunas de variáveis
Cabecalho<-(cbind(Id,G1,G2,Bloc,variaveis))
dados$Id <- dados$Id + num.familias
dados$Bloc <- dados$Bloc + num.familias
dados<-rbind(Cabecalho, dados)

# Duplicação da última informação
dados<- rbind(dados,dados[nrow(dados),])
dados[nrow(dados),4]<-99999

##Construção da genealogia
dados$Bloc <- as.factor(dados$Bloc)
genealogia <- dados[1:(nrow(dados)-1),1:3]              # Retira-se o indivíduo repetido do fim do arq. de dados modificado
G1 = genealogia[,2];        G2 = genealogia[,3];        Ind = genealogia[,1]
genealogia = pedigree(sire= as.integer(G1), dam=as.integer(G2), label=as.character(Ind))

##Modelo para predição de efeitos genéticos de indivíduos
for (n in 1:num.variaveis){
#Modelo para predição de efeitos genéticos de indivíduos
         ajuste= pedigreemm(dados[,4+n] ~ -1 + Bloc + (1|Id) ,
                                          data = dados,
                                          pedigree = list(Id = genealogia))
```


Agradecimentos

Os autores agradecem a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo apoio financeiro, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão da bolsa de estudo de doutorado e pós-doutorado, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudo de mestrado e a empresa Klabin S.A. pela concessão dos dados.

Referências

Resende, MDV (2007b) **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 362p.

Resende, MDV, FF Silva, PS Lopes, C F Azevedo (2012) **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/ Departamento de Estatística. 291p. URL http://www.det.ufv.br/ppestbio/corpo_docente. php

R Development Core Team (2012) R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org.

Vazquez AI, DM Bates, GJM Rosa, D Gianola and KA Weigel (2009) Technical note: An R package for fitting generalized linear mixed models in animal breeding. **Journal of Animal** Science.88:497 -504. Doi:10.2527/jas.2009-1952.