INFERENCIA BAYESIANA, VÍA ALGORITMO DE CADENAS INDEPENDIENTES, EN EL ANÁLISIS GENÉTICO DE ESPECIES PERENNES

Freddy Mora P.[1], Sandra Perret D.[2], María Paz Molina B.[3], Carlos Alberto Scapim[4] y Elias Nunes Martins[5]

RESUMEN

Diversos estudios han mostrado las ventajas de la utilización de la mejor predicción linear no sesgada (BLUP) en el análisis genético de especies forestales, así como la estimación de componentes de varianza vía máxima verosimilitud restringida (REML). Sin embargo, existen otras metodologías que permiten la obtención también robusta tanto de los componentes de varianza como los valores genéticos predichos, como por ejemplo la inferencia Bayesiana basada en algoritmos MCMC (Markov Chain Monte Carlo). Así, el presente estudio busca examinar la viabilidad de la obtención de parámetros genéticos y valores predichos en el mejoramiento genético de especies forestales, a través de la utilización de una variante de la metodología MCMC llamado algoritmo de Cadenas Independientes (CI).

Un ensayo de 14 procedencias Australianas de *Acacia saligna*, establecido en el norte de Chile, en el marco de un proyecto financiado por el Fondo de Desarrollo e Innovación (FDI) y el Instituto de Investigación Forestal (INFOR) fue utilizado para ilustrar el análisis genético Bayesiano. El diámetro medio de copa fue analizado en cada genotipo en un diseño de bloques completos al azar. Las estimaciones de componentes de varianza y la predicción de valores genéticos vía CI fueron comparadas con los procedimientos clásicos de estimación/predicción de modelos lineales mixtos REML/BLUP. El promedio de cada estimativa, la mediana y la moda de cada muestra a posteriori de tamaño 11000, fue utilizado dentro del algoritmo CI. Componentes de varianza y correlaciones intraclase fueron muy similares entre ambas metodologías. Tal correlación tuvo un valor de 27,33% y de 26,99%-28,48%, para los procedimientos REML y Bayesiano, respectivamente. BLUP y algoritmo CI, entregan valores genéticos similares, no evidenciándose cambios en el ordenamiento de los genotipos (ranking) entre un procedimiento y otro.

Ambas metodologías apuntan como las mejores procedencias a Greenough River, Lancelin y Ravensthorpe 2, indistintamente. Nosotros recomendamos ampliamente la utilización de inferencia Bayesiana, vía algoritmo de cadenas independientes, como variante de los métodos de Monte Carlo (MCMC), en el análisis genético de especies perennes, del mismo modo que el análisis clásico de modelos lineales mixtos vía REML/BLUP.

Palabras clave: Cadenas Independientes, Bayes, REML/BLUP, modelos mixtos.

[1] Departamento de Agronomia, Centro de Ciências Agrárias, Universidade Estadual de Maringá-Paraná, Brasil. fmora@universiabrasil.net
[2] Instituto de Investigación Forestal, INFOR. sperret@infor.cl
[3] Instituto de Investigación Forestal, INFOR. mmolina@infor.cl
[4] Departamento de Agronomia, Centro de Ciências Agrárias, Universidade Estadual de Maringá-Paraná, Brasil. cascapim@uem.br
[5] Departamento de Zootecnia, Centro de Ciências Agrárias, Universidade Estadual de Maringá-Paraná, Brasil. enmartins@uem.br

INTRODUCCIÓN

Para la obtención del mejor estimador lineal no segado (BLUE), que es la estimación de los efectos fijos, y de la mejor predicción lineal no sesgada (BLUP), particularmente relacionados al valor genético, se deben conocer las valores de los componentes de varianza, los cuales deben ser estimados con la mejor precisión posible. Numerosos trabajos han mostrado que el método de la máxima verosimilitud restringida (REML), para la estimación de componentes de varianza, es actualmente el procedimiento idóneo para la obtención de BLUE y BLUP en modelos lineales mixtos, con datos de observaciones fenotípicas no balanceado, cuyas estimativas no son sesgadas.

Los componentes de varianza vía REML pueden ser obtenidos por medio de los procedimientos MIXED y VARCOMP de SAS® (Astorga y Mora, 2005), y las soluciones de los efectos aleatorios a través del procedimiento MIXED, que puede establecerse en un único proceso iterativo de estimación-predicción (Mora y Arnhlod, 2006).

Tanto los componentes de varianza como la predicción de valores genéticos pueden ser obtenidos a través de métodos de inferencia Bayesiana. El análisis Bayesiano se basa en el conocimiento de la distribución *a posteriori* de los parámetros genéticos y viabiliza la conformación de regiones de credibilidad para las estimativas de tales parámetros (Gianola y Fernando, 1986).

Wolfinger y Kass (2000) examinaron regiones de credibilidad para funciones generales de componentes de varianza usando inferencia Bayesiana en un algoritmo de cadena independiente. Aquí, la distribución base para los efectos aleatorios es exacta, así que las muestras derivadas de los métodos de Monte Carlo, vía Cadenas de Marcov (MCMC) son rechazadas solamente si ellos no están en el espacio del parámetro, lo cual puede ocurrir por ejemplo, si un componente de varianza individual es estimado con valor negativo. Con ello el algoritmo de cadenas independientes es un caso especial de muestreo de rechazo (Tierney, 1994).

Siguiendo el trabajo de Mora y Arnhlod (2006) quienes compararon la inferencia Bayesiana con la metodología clásica de modelos lineares mixtos REML/BLUP, aplicados al mejoramiento del maíz, el presente trabajo busca examinar la viabilidad de la obtención de parámetros genéticos y valores predichos en el mejoramiento genético de especies perennes, a través de la utilización de una variante de la metodología MCMC llamado algoritmo de Cadenas Independientes (CI).

MATERIAL Y MÉTODOS

Ensayo y diseño experimental

Un ensayo de 14 procedencias Australianas de *Acacia saligna*, establecido en el norte de Chile, en el marco de un proyecto financiado por el Fondo de Desarrollo e Innovación (FDI) y el Instituto de Investigación Forestal (INFOR) fue utilizado para ilustrar el análisis genético Bayesiano.

El ensayo de procedencias fue establecido en la Comunidad Agrícola de Cuz–Cuz, en Illapel, centro norte de Chile (31° 63´ Lat S; 71° 22´ Long O; ubicación interior con 244 mm de precipitación media anual), en un diseño de bloques completos al azar, con parcelas de 20 árboles, 3 bloques, y todos los árboles fueron plantados a un distanciamiento de 2mx3m. Mediciones de Diámetro media de copa fue

considerado en el análisis genético de procedencias. El origen de cada procedencia es mostrado en el Cuadro 1.

Cuadro 1. Procedencias de *Acacia saligna* consideradas en el análisis Bayesiano y frecuentista.

Procedencia	Latitud S	Longitud O	Altitud (m)
Ravensthorpe 1	33° 35´	120° 08´	150
Kelmscott	32° 07´	116° 02´	150
Sanford River	27° 20´	116° 24´	320
Greenough River	28° 42´	115° 02´	175
Geraldton	28° 35´	114° 37´	200
Murchison River	27° 51´	114° 37´	180
Mingenew	29° 12´	115° 26´	180
Lake Indoon	29° 52´	115° 39´	100
Moora	30° 34´	116° 01´	200
Muntadgin Rock	31° 45´	118° 35´	320
Lake Muir	34° 26´	116° 40´	170
Boyatup Hill	33° 14´	123° 02´	183
Lancelin	31° 01´	115° 20´	10
Ravensthorpe 2	33° 35´	120° 03´	234

REML/BLUP: análisis de modelos mixtos clásico

Para la descripción de los valores observados se utilizó el siguiente modelo lineal estadístico, que incluye factores fijos y aleatorios, apropiados para la metodología de modelos mixtos:

$$y = X\beta + Z\gamma + \varepsilon$$

Donde *y* es el vector de las respuestas observadas, *X* es la matriz de diseño de los efectos fijos, *β* corresponde al vector de parámetros de efectos fijos, *Z* es la matriz de incidencia de los efectos aleatorios, *γ* es el vector de efectos aleatorios de procedencias y *ε* es el vector de residuos. Se asume que *γ* y *ε* son independientes y tienen distribución normal univariada con media 0 y matriz de varianza G y R, respectivamente.

Las ecuaciones de modelos mixtos (Henderson, 1984) son las siguientes:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Las que pueden ser escritos de la forma:

$$\hat{\beta} = (X'V^{-1}X)^{-}X'V^{-1}y$$
$$\hat{\gamma} = GZ'V^{-1}(y - X\beta)$$

Donde "−" es una inversa generalizada de la matriz correspondiente, V^{-1} es la inversa de la matriz de varianza de *y*, definida como $V = ZGZ'+R$.

La correlación intraclase fue estimada a partir de la siguiente expresión:

$$\Gamma^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\varepsilon^2}$$

Donde $\hat{\sigma}_p^2$ y σ_e^2 son los componentes de varianza del efecto de procedencias y residual, respectivamente, estimados vía máxima verosimilitud restringida (REML).

Las estimaciones de componentes de varianza y las soluciones de los efectos aleatorios fueron calculados en un único proceso iterativo de estimación-predicción, utilizando el procedimiento PROC MIXED de SAS ® 8.02 (SAS Institute, 1996).

Inferencia Bayesiana

En inferencia Bayesiana, todos los parámetros del modelo son considerados como variables aleatorias (Mora y Arnhold, 2006). Los componentes de varianza, heredabilidad y valores genéticos predichos obtenidos por el método Bayesiano se basaron en el promedio, la mediana y la moda de cada muestra *a posteriori* de tamaño 11000, generada por el algoritmo de cadena independiente, variante de los métodos MCMC (Markov Chain Monte Carlo). El procedimiento PROC MIXED con opción PRIOR se utilizó para este análisis, construyéndose un conjunto de datos con las informaciones *a priori* de los parámetros (densidad base).

Definimos θ como el vector de componentes de varianza. La función de densidad conjunta de (β, γ, θ) es:

$$f(\beta, \gamma, \theta \mid y) = f(\beta, \gamma \mid \theta, y)\, f(\theta \mid y)$$

La especificación de cada distribución a priori de los elementos del modelo es:

$$f(y \mid \beta, \gamma, \sigma_\varepsilon^2, \sigma_p^2) \sim N(X\beta + Z\gamma, R)$$
$$f(\beta) \sim Uniforme$$
$$f(\gamma \mid \sigma_p^2) \sim N(0, G)$$
$$f(\sigma_p^2) \sim IG(a, b)$$
$$f(\sigma_\varepsilon^2) \sim IG(c, d)$$

IG corresponde a distribución gamma invertida, con parámetros *a priori* obtenidos a través del cuadro de densidades base de un análisis preliminar, utilizando distribución no informativa de Jeffreys. La función de densidad a priori del parámetro de efecto fijo es asumida como uniforme (opción FLAT).

La distribución a posteriori conjunta es definida como:

$$f(\beta, \gamma, \sigma_\varepsilon^2, \sigma_p^2 \mid y) \propto f(\beta) f(\gamma \mid \sigma_p^2) f(\sigma_p^2) f(\sigma_\varepsilon^2) f(y \mid \beta, \gamma, \sigma_\varepsilon^2, \sigma_p^2)$$

Las muestras a posteriori fueron analizadas utilizando los procedimientos PROC AUTOREG y UNIVARIATE (SAS Institute, 1996). La moda de cada muestra fue determinada de acuerdo al procedimiento de densidad de Kernel.

RESULTADOS Y DISCUSIÓN

Los componentes de varianza y la correlación genética intraclase vía máxima verosimilitud restringida (REML) e inferencia Bayesiana, basada en el algoritmo de cadenas independientes, son mostrados en el Cuadro 2. Cabe destacar la inexistencia de período de descarte (burn-in) en el algoritmo de cadenas independientes, siendo una ventaja del método al compararlo con los métodos tradicionales de Gibbs (Wolfinger y Kass, 2000).

Los componentes de varianza y correlaciones intraclase fueron muy similares entre ambas metodologías. Tal correlación tuvo un valor de 27,33% y de 26,99%-28,48%, para los procedimientos REML y Bayesiano, respectivamente.

BLUP y algoritmo CI, entregan valores genéticos similares, no evidenciándose cambios en el ordenamiento de los genotipos (ranking) entre un procedimiento y otro (Cuadro 3). Esto fue confdirmado a través de los valores de correlación de Spearman, entre BLUP y cada variable considerada en la metodología Bayesiana, los cuales fueron altos y altamente significativos ($p<0,01$) (datos no mostrados). Nogueira *et al*. (2003), en un análisis comparativo entre inferencia Bayesiana y Frecuentista de un modelo lineal mixto, trabajando con animales encontró valores divergentes entre los componentes de varianza residuales, con altas desviaciones estándar de las varianzas y valores genéticos, utilizando el algoritmo de muestreo de Gibbs, y observó una divergencia importante en el ordenamiento de los genotipos.

Ambas metodologías apuntan como las mejores procedencias a Greenough River, Lancelin y Ravensthorpe 2, indistintamente.

Cuadro 2. Componentes de varianza y correlación intraclase vía máxima verosimilitud restringida (REML) e inferencia Bayesiana (algoritmo de cadenas independientes).

Estimativa	REML	BAYESIANO (Cadenas Independientes)			
		Media	Mediana	Moda	DE[1]
Varianza de procedencias	438,34	480,2657	451,5012	414,9731	168,4781
Varianza Residual	1165,35	1171,553	1166,944	1153,097	83,18331
Correlación Intraclase	0,2733	0,2848	0,2786	0,2699	0,06893

[1]DE: Desviación estándar

Cuadro 3. Valores genéticos del efecto de procedencias de Acacia, considerando la mejor predicción linear no sesgada (BLUP) y la inferencia Bayesiana, vía algoritmo de cadenas independientes.

Procedencia	BLUP	BAYESIANO (Cadenas Independientes)			
		Media	Mediana	Moda	DE[1]
Ravensthorpe 1	-6,436	-6,545	-6,630	-6,533	9,9
Kelmscott	-7,768	-7,697	-7,805	-8,481	9,4
Sanford River	0,444	0,580	0,515	0,680	9,6
Greenough River	14,789	14,828	14,670	13,107	9,6
Geraldton	-14,739	-14,664	-14,645	-15,423	9,4
Murchison River	-3,829	-3,722	-3,823	-4,573	9,4
Mingenew	-8,900	-8,751	-8,727	-9,568	9,6
Lake Indoon	-7,313	-7,127	-7,061	-7,396	9,5
Moora	-0,032	-0,005	-0,080	-0,888	10,4
Muntadgin Rock	-6,992	-6,843	-6,937	-7,900	10,3
Lake Muir	-21,618	-21,558	-21,565	-23,044	10,1
Boyatup Hill	-18,188	-18,044	-18,036	-19,041	10,1
Lancelin	34,369	34,399	34,289	35,654	9,7
Ravensthorpe 2	46,215	46,264	46,014	44,901	10,3

[1]DE: Desviación estándar

CONCLUSIONES

Inferencia Bayesiana, vía la variante de los métodos MCMC, cadenas independientes, puede ser utilizada eficientemente en el análisis genético de especies forestales. A la luz de los resultados, se recomienda ampliamente la utilización de inferencia Bayesiana, vía este algoritmo, del mismo modo que el análisis clásico de modelos lineales mixtos vía REML/BLUP.

REFERENCIAS

Astorga, M. y F. Mora. 2005. Componentes de varianza e interacción variedad-sitio del vigor, producción y productividad de *Olea europaea*, en Chile. Cerne 11: 25-33.

Gianola, D. and R.L. Fernando. 1986. Bayesian methods in animal breeding theory. Journal of Animal Science 63: 217-244.

Mora, F. y E. Arnhlod. 2006. Inferencia Bayesiana y metodología de modelos lineales mixtos aplicados al mejoramiento del maíz. Ciencia e Investigación Agraria, 35.

Nogueira, D.A., T. Sáfadi, E. Bearzoti, and J.S.S. Bueno-Filho. 2003. Análises clássica e Bayesiana de um modelo misto aplicado ao melhoramento animal: uma ilustração. Ciênc. agrotec. 27:1614-1624.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). Annals of Statistics, 23: 1701-1762.

Wolfinger, R.D. and R.E. Kass. 2000. Non-conjugate Bayesian analysis of variance component models. Biometrics, 56: 768-774.