Inferencia Bayesiana en modelos mixtos con datos faltantes, efectos de competencia genética y tendencias espaciales para la evaluación genética forestal

Eduardo Pablo Cappa

Director: Rodolfo J. C. Cantet
Consejero: Martín O. Grondona

UBA FAUBA

Escuela para Graduados Alberto Soriano
Facultad de Agronomía
Universidad de Buenos Aires

EPG

Hoja de ruta

1. Introducción. **(14)**

- Naturaleza de los datos genéticos forestales.
- Naturaleza de los datos genéticos forestales y Modelos Mixtos.
- Modelo Mixto de árbol individual unicarácter.
- EMM, GLS, BLUP.
- Inferencia Bayesiana.

2. Motivaciones y Objetivos Generales. **(2)**

3. Capítulos. **(38)**

- Capítulo 3: Inferencia Bayesiana para un modelo mixto de árbol individual multicarácter con datos faltantes vía “Full Conjugate Gibbs”.
- Capítulo 4: Efectos aditivos directos y de competencia en mejoramiento forestal: estimación Bayesiana en un modelo mixto de árbol individual.
- Capítulo 5: Estimación Bayesiana de una superficie para modelar la tendencia espacial utilizando un modelo mixto semiparamétrico de árbol individual.

4. Conclusiones / Contribuciones. **(3)**

Naturaleza de los datos genéticos forestales

Característica	Ensayos Genéticos Agrícolas	Ensayos Genéticos Forestales
Unidad experimental	Parcela o muchas plantas.	Individuo (árbol).
Tamaño de ensayos	Menor.	Mayor, en el número de unidades experimentales y superficie.
Número de Replicas	Menor.	Mayor, para compensar bajo número de plantas por parcelas.
Uniformidad de Sitios	Terrenos más uniformes.	Terrenos menos uniformes: sitios marginales, con desnivel.
Distribución	En filas y columnas.	En grilla de filas y columnas, regularmente espaciadas.
Mediciones	Única medición anual.	Repetidas a lo largo del tiempo.
Material Genético	Líneas generalmente endocriadas.	Mayor diversidad genética.
Competencia	Menos importante ya que se da a nivel de parcela.	Más importante ya que se da a nivel de árbol individual.

Fuente: adaptado de Costa e Silva *et al.* (2001)

Naturaleza de los datos genéticos forestales

Como consecuencia de estas características, los mejoradores genéticos forestales se ven obligados a conducir el análisis de EGF considerando:

- relaciones aditivas entre individuos emparentados
- poblaciones compuestas
- correlaciones genéticas entre caracteres
- desbalance de las observaciones, tanto con caracteres múltiples como con medidas repetidas en el tiempo
- heterogeneidad ambiental
- competencia entre unidades vecinas

Naturaleza de los datos GF y Modelos mixtos

- Para analizar datos con estas características, los genetistas forestales utilizan la metodología de MM con predicciones BLUP de los valores de cría.
- Las predicciones BLUP de los valores de cría dependen del MM utilizado para el análisis, en particular de la estructura de covarianza asumida.
- La presente tesis aborda la estimación de componentes de (co)varianza en tres situaciones de evaluación:
  - el análisis multicarácter con datos faltantes;
  - la presencia de efectos genéticos aditivos de competencia;
  - la modelación de la (co)variabilidad ambiental empleando una función *spline* en dos dimensiones.

Modelo mixto aditivo de árbol individual unicarácter

Función matemática lineal con ***efectos "fijos"*** y ***variables aleatorias.***

Expresión matricial

$$\boldsymbol{y} = \underbrace{\boldsymbol{X}\boldsymbol{\beta}}_{\text{parte "fija"}} + \underbrace{\boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e}}_{\text{parte aleatoria}}$$

Momentos

$$\mathrm{E}(y) = \boldsymbol{X}\,\boldsymbol{\beta}$$

Matriz de relaciones aditivas — **Varianza aditiva**

$$\mathrm{Var}(y) = \mathrm{Var}\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & 0\\ 0 & \boldsymbol{R}\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}\sigma^2_A & 0\\ 0 & \boldsymbol{I}\sigma^2_e\end{bmatrix}$$

Varianza del error

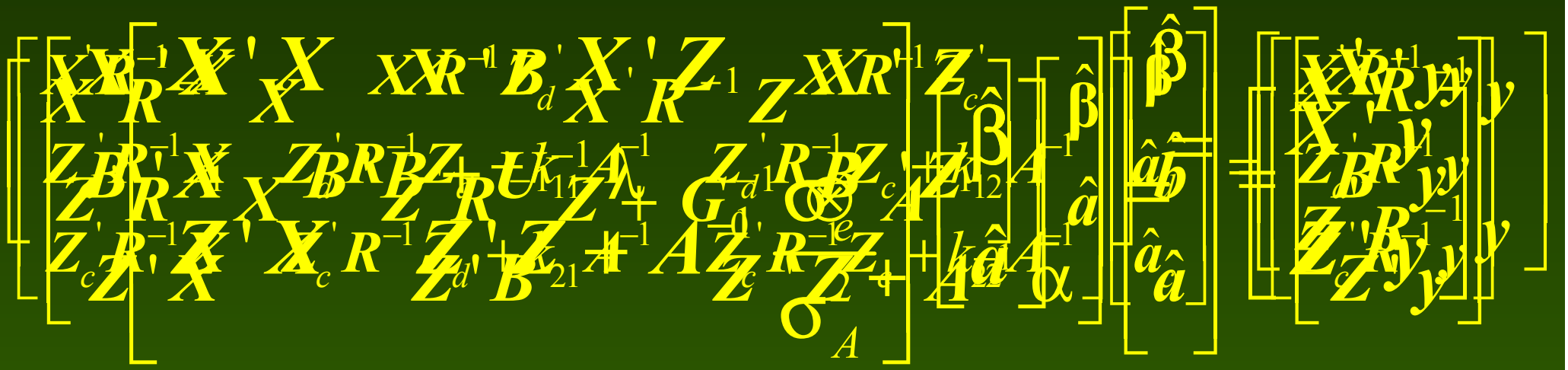
Supuestos

Matrices $\boldsymbol{G}$ y $\boldsymbol{R}$ de elementos conocidos y positivas definidas.

$$p(\boldsymbol{y}, \boldsymbol{a}) \sim \mathrm{NMV}$$

MME – GLS - BLUP

Ecuaciones de Modelo Mixto



Asumiendo los componentes de varianza conocidos, las soluciones a las MME:

‘mínimos cuadrados generalizados’ (GLS) de los efectos fijos

‘predicciones lineales insesgadas de mínima varianza’ (BLUP) de los efectos aleatorios (valores de cría).

EGF y estimación de componentes de (co)varianza

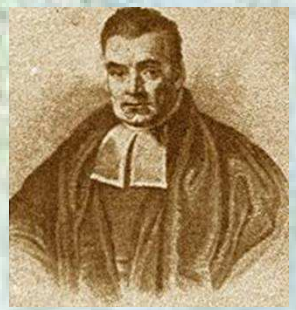
En todos los casos estos métodos consideran los parámetros de dispersión en las MME como "verdaderos".

- **No toma en cuenta los errores de estimación de los componentes de dispersión.**

- **No es posible obtener analíticamente la varianza de muestreo de las estimaciones de los parámetros de dispersión.**


Si nada parece
funcionar,

intente con el Reverendo Thomas Bayes



$$p(\theta|y) = \frac{p(y|\theta)\ p(\theta)}{p(y)} \propto p(y|\theta)\ p(\theta)$$

Inferencia Bayesiana

La teoría Bayesiana, dado los datos (y), cuantifica la incertidumbre de los parámetros (θ) a través de probabilidades.

Distribución "*a posterior*" = *actualización de los conocimientos acerca de* θ *luego de observar* ***y***

Distribución a priori o "*prior*" = *representa el conocimiento previo de los parámetros*

$$p(\theta \mid y) \propto p(y \mid \theta) \times p(\theta)$$

Verosimilitud = *refleja la información sobre* θ *aportada por los datos*

Inferencia Bayesiana

- Distribuciones marginales "*a posteriori*" son calculadas con el algoritmo de Cadenas de Markov Monte Carlo (MCMC): ***Muestreo de Gibbs***.

- Distribuciones condicionales "*a posteriori*" son obtenidas analíticamente.

Verosimilitud y distribuciones "a priori"

Verosimilitud

$$p\left(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{R}\right)\sim N\left[\boldsymbol{X\beta}+\boldsymbol{Za},\boldsymbol{R}\right]$$

Distribuciones a priori

- *Efectos fijos:* $\boldsymbol{\beta}\sim N(\boldsymbol{0},\boldsymbol{K})$, matriz de covarianzas diagonal $\boldsymbol{K}$ ($K_i > 10^8$) ("proper prior": Hobert and Caselllla, 1996).
- *Valores de cría* $[\boldsymbol{a}]=[\boldsymbol{a}]\,|\,\boldsymbol{G}_0,\boldsymbol{A}\sim N([\boldsymbol{0}],\boldsymbol{G}_0\otimes\boldsymbol{A})$
- *Coeficientes del producto tensorial de B-spline:* $\boldsymbol{b}\sim N_b(\boldsymbol{0},\boldsymbol{U}\,\sigma^2{}_b)$
- *Varianzas de* $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{e}$: $\sigma_i^2\sim$ Inv Chi Square $(\boldsymbol{\delta}_i^2,\upsilon_i)$
- *Matriz de (co)varianzas de los valores de cría:* $\boldsymbol{G_0}\sim$ IW $(\boldsymbol{G^*_0},\upsilon_A)$
- *Varianza del error:* $\boldsymbol{R_0}\sim$ IW $(\boldsymbol{R^*_0},\upsilon_k)$

$$p\left(\boldsymbol{R}_0\,|\boldsymbol{R}_0^*,\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K,\nu_k\right)\propto\prod_{k=1}^{K}\left|\boldsymbol{M}_k\boldsymbol{R}_0\boldsymbol{M}_k'\right|^{-\frac{(\nu_k+2r_k+1)}{2}}\times\quad\exp\left\{-\tfrac{1}{2}\operatorname{tr}\left[\boldsymbol{M}_k\boldsymbol{R}_0^*\boldsymbol{M}_k'\left(\boldsymbol{M}_k\boldsymbol{R}_0\boldsymbol{M}_k'\right)^{-1}\right]\right\}$$

Distribuciones condicionales "a posteriori"

Condicional β, ***a***

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{a} \end{bmatrix} \mid \boldsymbol{G}_0, \boldsymbol{R}_0, \boldsymbol{y} \sim N\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{a}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}_0^{-1} \otimes \boldsymbol{A}^{-1} \end{bmatrix}^{-1} \right)$$

Condicional de varianzas ***a***, ***b***, ***e***: $\sigma_i^2 \sim \text{Inv Chi Square}\left(\tilde{\upsilon}_i, \tilde{\delta}_i^2\right)$

Condicional $\boldsymbol{G_0}$ $\boldsymbol{G}_0 \sim \text{IW}\left(\left(\boldsymbol{G}_0^* + \boldsymbol{S}\right)^{-1}, \upsilon_A + \#\text{arboles ped.} + 3\right)$

Condicional $\boldsymbol{R_0}$ (Cantet *et. al.* 2004)

$$p\left(\boldsymbol{R}_0 \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{G}_0\right) \sim \text{IW}\left(\left(\sum_{k=1}^{K} \boldsymbol{R}_k^* \right), \sum_{k=1}^{K} \upsilon_k + n + (K-1)(r+1) \right)$$

Diagnósticos de convergencia

- Luego de un periodo de ‘calentamiento’ las muestras generadas crean una cadena de Markov cuya distribución de equilibrio es la densidad posterior.

- Análisis de convergencia empleando diagnósticos en BOA (***R***).

  - Métodos gráficos: Promedio acumulado por iteración.

  - Cálculos de estadísticos: Geweke (1992), Raftery y Lewis (1992).

  - Autocorrelación en función de la separación entre muestras (*lag*) y función de autocorrelación.

Motivaciones

- Posibilidad de adaptar los Modelos Lineales Mixtos a características que son comunes en datos de EGF.

- Escasa a nula aplicación de técnicas Bayesianas de estimación en la EGF.

Objetivos generales

a) Describir nuevos Modelos Lineales Mixtos para analizar datos de ensayos genéticos forestales unicarácter o multicarácter, que contemplen observaciones faltantes, competencia genética y heterogeneidad espacial.

b) Proponer el enfoque metodológico Bayesiano para estimar componentes de (co)varianza en Modelos Lineales Mixtos, aplicados a las observaciones provenientes de ensayos genéticos forestales.

Capítulo 3

Inferencia Bayesiana para un modelo mixto de árbol individual multicarácter con datos faltantes vía "Full Conjugate Gibbs".

Eduardo P. Cappa, and Rodolfo J. C. Cantet. (2006). Bayesian inference for normal multiple trait individual tree models with missing records via Full Conjugate Gibbs. Can. J. For. Res. **36:** 1276-1285.

"State of the art"

- **Van Tassell C.P., and Van Vleck L.D.** 1996. J. Anim. Sci. 74: 2586-2597.

- **Cantet, R.J.C., Birchmeier, A.N., and Steibel, J.P.** 2004. Genet. Sel. Evol. 36: 49–64.

- **Spiegelhalter, D.J., Best, N. G., Carlin, B.P., and Van der Linde, A.** 2002. Journal of the Royal Statistical Society. Series B, 64: 583-639.

Introducción

Observaciones perdidas son comunes en datos provenientes EGF.

La estimación de heradabilidades y correlaciones genéticas con datos faltantes es un problema estadístico complejo.

Van Tassel y Van Bleck (1996) emplearon el algoritmo MCMC de "*Data Augmentation*"; sin embargo, este presenta una baja convergencia.

Cantet *et. al.* (2004) propusieron un método MCMC: El algoritmo "***Full Conjugate Gibbs***" (FCG) para modelos con caracteres normales múltiples.

Introducción

La estimación Bayesiana en Modelos Lineales Mixtos de árbol individual es condicional a un modelo particular.

Distintos Modelos Lineales Mixtos pueden competir.

Spiegelhalter *et*. *al*. (2002) propusieron un estadístico Bayesiano: "*Deviance Information Criterion*" (DIC).

Objetivos

- Aplicar el algoritmo de FCG propuesto por Cantet *et*. *al*. (2004) para estimar componentes de (co)varianza, o funciones de los mismos, en modelos de árbol individual multicarácter con datos faltantes.
- Obtener una expresión del DIC que nos permita seleccionar modelos de árbol individual con caracteres normales múltiples.
- Ilustrar la metodología presentada con datos de pinos híbridos.

Modelo mixto aditivo de árbol individual multicarácter

Expresión matricial

$$\begin{bmatrix} y_1 \\ . \\ y_r \end{bmatrix} = \begin{bmatrix} X_1 & . & 0 \\ . & . & . \\ 0 & . & X_r \end{bmatrix} \begin{bmatrix} \beta_1 \\ . \\ \beta_r \end{bmatrix} + \begin{bmatrix} Z_r & . & 0 \\ . & . & . \\ 0 & . & Z_r \end{bmatrix} \begin{bmatrix} a_1 \\ . \\ a_r \end{bmatrix} + \begin{bmatrix} e_1 \\ . \\ e_r \end{bmatrix}$$

$$y = \sum_{i=1}^{r} \oplus X_i \beta + \sum_{i=1}^{r} \oplus Z_i a + e$$

Momentos

$$\mathrm{Var}\begin{bmatrix} a_1 \\ a_2 \\ . \\ a_r \end{bmatrix} = G = \begin{bmatrix} g_{1,1}A & g_{1,2}A & . & g_{1,r}A \\ g_{2,1}A & g_{2,2}A & . & g_{2,r}A \\ . & . & . & . \\ g_{r,1}A & g_{r,2}A & . & g_{rr}A \end{bmatrix} = \begin{bmatrix} g_{1,1} & g_{1,2} & . & g_{1,r} \\ g_{2,1} & g_{2,2} & . & g_{2,r} \\ . & . & . & . \\ g_{r,1} & g_{r,2} & . & g_{rr} \end{bmatrix} \otimes A = G_0 \otimes A$$

Asumiendo que todos los individuos tienen registro:

$$\mathrm{Var}\begin{bmatrix} e_1 \\ e_2 \\ . \\ e_r \end{bmatrix} = R = \begin{bmatrix} r_{1,1}I & r_{1,2}I & . & r_{1,r}I \\ r_{2,1}I & r_{2,2}I & . & r_{2r}I \\ . & . & . & . \\ r_{r,1}I & r_{r,2}I & . & r_{r,r}I \end{bmatrix} = R_0 \otimes I$$

Matriz de (co)varianzas del error con datos faltantes

$$\mathrm{Var}(\boldsymbol{e}) = \boldsymbol{R} = \begin{bmatrix} \boldsymbol{I}_{n_1} \otimes \boldsymbol{M}_1 \boldsymbol{R}_0 \boldsymbol{M}_1' & \boldsymbol{0} & . & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_2} \otimes \boldsymbol{M}_2 \boldsymbol{R}_0 \boldsymbol{M}_2' & . & \boldsymbol{0} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \boldsymbol{0} & \boldsymbol{0} & . \; . \; . & \boldsymbol{I}_{n_k} \otimes \boldsymbol{M}_K \boldsymbol{R}_0 \boldsymbol{M}_K' \end{bmatrix}$$

ENFOQUE BAYESIANO

Esquema de muestreo del FCG

1. Construir y resolver las MME;
2. Muestrear $\boldsymbol{\beta}$ y $\boldsymbol{a}$;
3. Calcular los residuales: $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{a}$;
4. Muestrear, para cada patrón, los elementos de las matrices de hipercovarianzas para los caracteres faltantes;
5. Muestrear $\boldsymbol{R}_0$;
6. Calcular $\boldsymbol{S}$;
7. Muestrear $\boldsymbol{G}_0$;

Comparación de modelos

$$\mathrm{DIC} = \bar{D}(\theta_M) + p_D$$

Media posterior de la "*Desviación*"

N° efectivo de parámetros

$$\mathrm{DIC} = 2\,\bar{D}(\theta_M) - D(\bar{\theta}_M)$$

$$p_D = \bar{D}(\theta_M) - D(\bar{\theta}_M)$$

$$D(\theta) = -2\log p(y|\theta, R)$$

$$= \mathrm{N}\log(2\pi) + \log|R| + (y - X\beta - Za)'\,R^{-1}\,(y - X\beta - Za)$$

Suma de cuadrados ponderada

$$= \mathrm{N}\log(2\pi) + \sum_{k=1}^{K} n_k \log|M_k R_0 M_k'| + \sum_{k=1}^{K} \mathrm{tr}\left[(M_k R_0 M_k')^{-1} E_{(\theta)k}\right]$$

Ejemplo de trabajo: Análisis de P. híbridos

- ***Pinus elliottii.*** **var. *elliotti.*** **×** ***Pinus caribaea*** **var. *hondurensis*. E, F1: E × H y H × E, Backcross: E × F1, H × F1, F2.**

- **DAP (1598) y Altura (1570) a los 7 años de edad.**

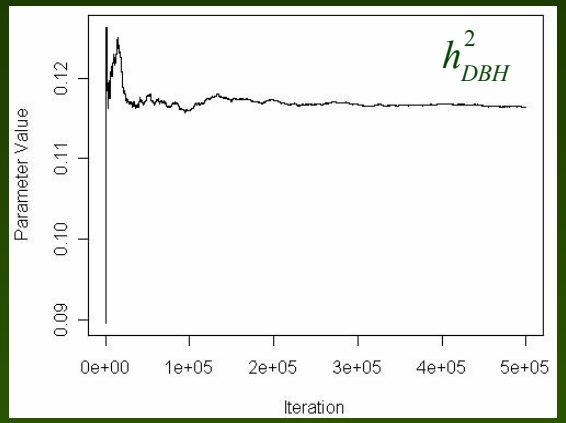
- **DBCA con 4 repeticiones a un espaciamiento de 3.5 m × 3.5 m y 4 m × 3 m.**

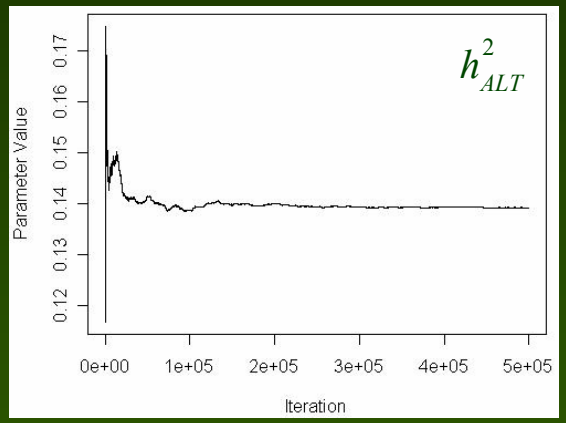
Inferencia Bayesiana posterior

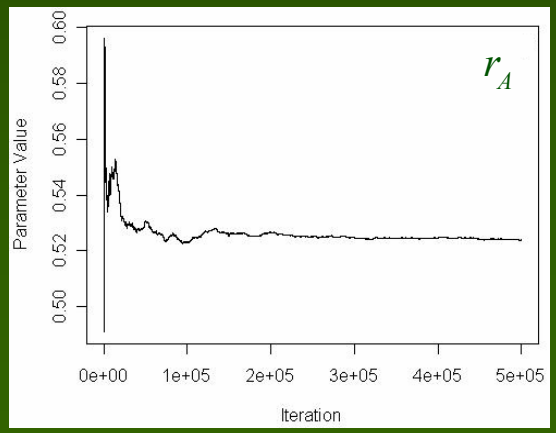
- Se generaron 1.010.000 muestras en una cadena simple, las primeras 10.000 iteraciones fueron descartadas debido al período de calentamiento.
- Los valores de hipervarianzas para g_{11}, g_{22}, r_{11} y r_{22} fueron calculadas en una corrida previa utilizando GS para cada carácter y para la misma base de datos.
- Las covarianzas *a priori* para g_{12} y r_{12} fueron seleccionadas con valores pequeños (pero no 0) y positivos.
- Se generaron corridas preliminares con cadenas simples para distintos valores de hipervarianzas.

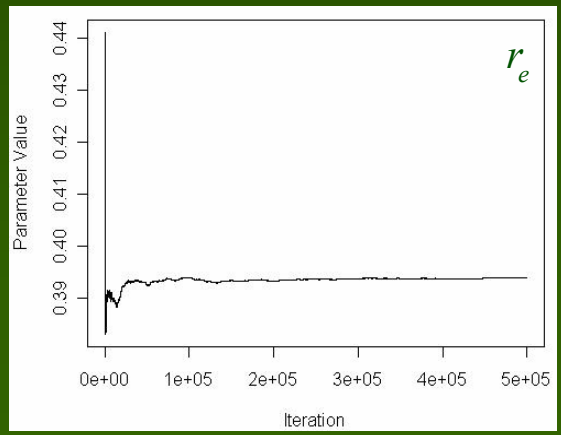
Diagnósticos de convergencia

- Promedio acumulado por iteración:









Ejemplo de trabajo: Análisis de P. híbridos

Cuadro 3.3. Esperanza de la desviación Bayesiana $\bar{D}(\boldsymbol{\theta})$, desviación Bayesiana evaluada en la media posterior de los parámetros del modelo $D(\bar{\boldsymbol{\theta}})$, número efectivo de parámetros (p_D), y "*Deviance Information Criterion*" (DIC) para los ocho modelos analizados.

Modelo[a]	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$\bar{D}(\boldsymbol{\theta})$	3943.355	3944.455	3945.583	3946.168	3944.593	3944.913	3944.908	3946.408
$D(\bar{\boldsymbol{\theta}})$	3940.733	3941.806	3942.928	3943.502	3941.956	3942.256	3942.262	3943.767
p_D	2.622	2.649	2.655	2.666	2.637	2.657	2.646	2.641
DIC	**3945.977**	**3947.104**	**3948.238**	**3948.834**	**3947.230**	**3947.570**	**3947.554**	**3949.049**

Nota:[a] **Modelo 1**: A. **Modelo 2**: A + D. **Modelo 3**: A + (A × A). **Modelo 4**: A + (D × D).
Modelo 5: A + D + (A × A). **Modelo 6**: A + D + (D × D).
Modelo 7: A + D + (A × D). **Modelo 8**: A + (A × A) + (A × D).

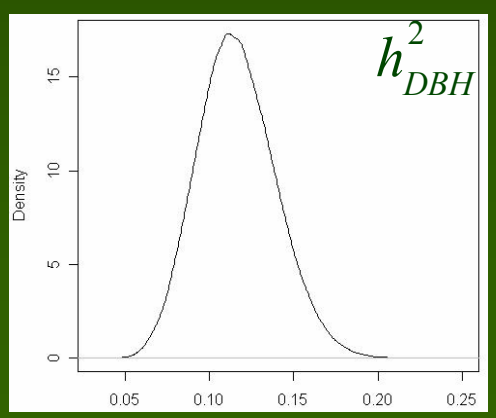
r_A

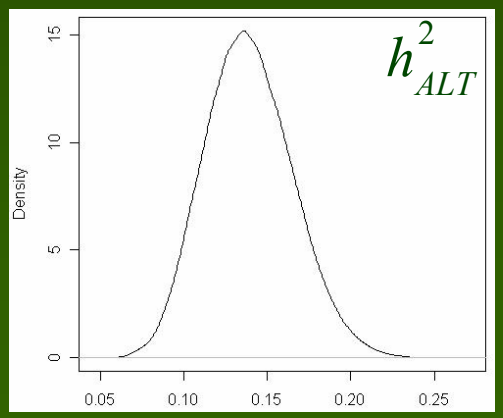
Ejemplo de trabajo: Análisis de P. híbridos

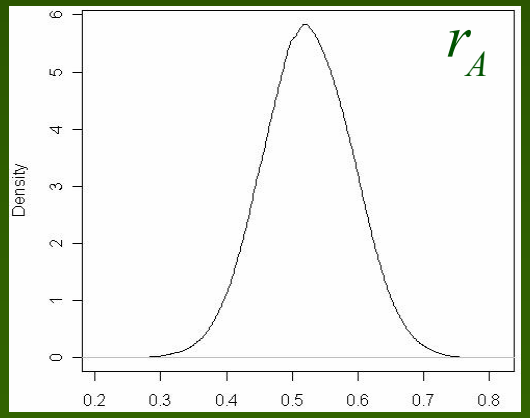
Cuadro 3.4. Estadísticos posteriores para la heredabilidad del diámetro (h^2_{DAP}), heredabilidad de la altura (h^2_{Altura}), correlación aditiva (r_A) y ambiental (r_E).

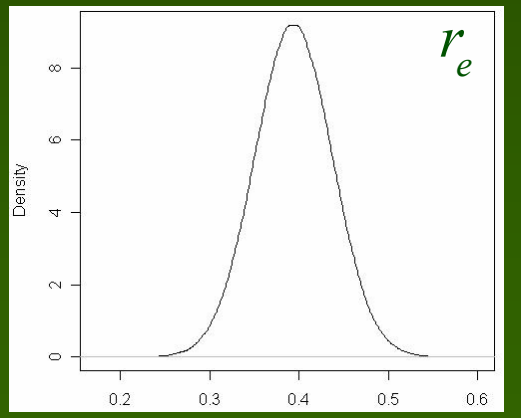
	Media	Mediana	Modo	SE Efectivo	95% HPD	ESS
h^2_{DAP}	0.116	0.115	0.114	0.019	0.080 – 0.156	5207
h^2_{Altura}	0.139	0.138	0.121	0.022	0.098 – 0.184	5631
r_A	0.524	0.524	0.515	0.057	0.411 – 0.636	5306
r_E	0.394	0.394	0.387	0.042	0.323 – 0.465	56744

SE Efectivo= Error Efectivo Estándar; **95% HPD=** Intervalo de alta Densidad Posterior**; ESS**= Tamaño Efectivo de Muestra.









Capítulo 4

Efectos aditivos directos y de competencia en mejoramiento forestal: estimación Bayesiana en un modelo mixto de árbol individual

---

Cappa Eduardo P., and Rodolfo J. C. Cantet. (2007). Direct and competition additive effects in tree breeding: Bayesian estimation from an individual tree mixed model. Silvae Genetica (*Aceptado*).

"State of the art"

- **Muir, William M.** 2005. Genetics **170**: 1247-1259.
- **Arango, J. et al.** 2005. J. Anim. Sci. **83**: 1241-1246.
- **Van Vleck L. D. and Cassady J. P.** 2005. J. Anim. Sci. **83**: 68-74.

Introducción

FENOTIPO = EFECTOS DIRECTOS + EFECTOS INDIRECTOS

DE SUS PROPIOS GENES

DE CONTRIBUCIONES DE OTROS GENOTIPOS

Los efectos genéticos directos ocurren cuando los genes del individuo influyen directamente sobre su fenotipo, mientras que los efectos genéticos indirectos se expresan en el fenotipo de otro individuo.

En la EGF un efecto genético indirecto es la **COMPETENCIA**.

Muir y Schinkel (2002) introdujeron la idea de predecir ambos efectos para animales compitiendo en corral, utilizando las ecuaciones de modelos mixtos.

Muir (2005) extendió el modelo en plantas, ignorando las consecuencias del número variable de competidores sobre la varianza aditiva de competencia.

Van Vleck *et. al.* (2005) simularon datos de animales criados en corrales, utilizando un número fijo de competidores e ignorando las relaciones aditivas y las consecuencias sobre el sesgo en la estimación de los componentes de varianza de este modelo.

Van Vleck *et. al.* (2005) y **Arango *et. al*. (2005)** encontraron dificultades para estimar los parámetros de dispersión bajo la metodología **REML** en animales.

Objetivos

- Presentar un modelo genético aditivo de árbol individual para la EGF que incluya efectos directos y de competencia, teniendo en cuenta el número y la posición de los árboles.

- Estimar los parámetros de dispersión del modelo propuesto utilizando un enfoque Bayesiano por medio del algoritmo GS.

- Ilustrar la metodología presentada con datos de diámetro a la altura del pecho de *Pinus taeda* L. a los 13 años de edad.

Modelo estadístico – Ef. de competencia

MM de árbol individual con efectos de competencia:

$$y = \underbrace{X\beta}_{\text{parte fija}} + \underbrace{Z_d a_d + Z_c a_c + e}_{\text{parte aleatoria}}$$

Genético directo **Genético de competencia**

La competencia total que recibe el árbol *i* por parte de los valores de cría de sus vecinos es igual a:

$$Z_c a_c = \sum_{j=1}^{m} f_{ij}\, \boldsymbol{a}_{c_j} = f_{i1}\, \boldsymbol{a}_{c_1} + f_{i2}\, \boldsymbol{a}_{c_2} + \ldots + f_{im}\, \boldsymbol{a}_{c_m}$$

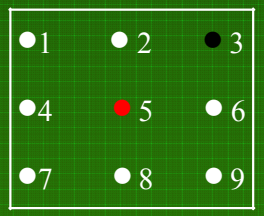
Es razonable suponer que f_{ij} es una función de la distancia entre el árbol *i* y su competidor *j* (*j*= 1,....., *m*).

- en la misma fila o columna $f_{ijR\text{-}C} = 1/d.$
- en la diagonal $f_{ijD} = 1/\,2^{\,½}\, d.$

Efectos genético aditivo de competencia

Tomando en cuenta el número variable de competidores y las relaciones aditivas:

$$f_{ijR-C} = \left[\sum_{j=1}^{m} 2 f_{ijK}^2 = n_{R-C} \right]^{1/2} f_{ijR-C} \quad f_{ijD} + n_D f_{ijD}^2 = 1 \Big/ \left[2n_{R-C} + n_D \right]^{1/2}$$



$$\left[1/\sqrt{11} \quad \sqrt{2}/\sqrt{11} \quad \sqrt{2}/\sqrt{11} \quad 0 \quad \sqrt{2}/\sqrt{11} \quad 1/\sqrt{11} \quad \sqrt{2}/\sqrt{11} \quad 1/\sqrt{11} \right]$$

$$f_{ijD} = 1/\,[2*4+3]^{1/2} = 1/(11)^{1/2}$$

$$f_{ij\text{R-C}} = [2/(2*4+3)]^{1/2} = (2/11)^{1/2}$$

La varianza genética total es:

$$\text{var}(a_{d_i}) + \text{var}\left(\sum_{j=1}^{m} f_{ijk}\, a_{c_j} \right) + 2\,\text{cov}\left(a_{d_i}, \sum_{j=1}^{m} f_{ijk}\, a_{c_j} \right) =$$

$$= (1+F_i)\,\sigma^2_{\text{A}d} + \left[\left(n_{\text{R-C}}\, f^2_{i\,1\text{R-C}} + n_{\text{D}}\, f^2_{i\,1\text{D}} \right)(1+F_j) + 2 \sum_{j \neq j'}^{m_i} f_{ij'k}\, f_{imk}\, A_{jj'} \right] \sigma^2_{\text{A}c} + 2 \sum_{j=1}^{m_i} f_{ijk}\, A_{ij}\, \sigma_{\text{A}d\text{A}c}$$

Momentos

$$\mathrm{E}(y) = \boldsymbol{X}\,\beta$$

Varianza aditiva del VC directo

$$V\begin{bmatrix} \boldsymbol{a}_d \\ \boldsymbol{a}_c \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}_0 \otimes \boldsymbol{A} & 0 \\ 0 & \boldsymbol{R} \end{bmatrix} = \begin{bmatrix} A\sigma^2_{Ad} & A\sigma_{AdAc} & 0 \\ A\sigma_{AdAc} & A\sigma^2_{Ac} & 0 \\ 0 & 0 & I\sigma^2_e \end{bmatrix}$$

Covarianza entre VC directo y VC de competencia

Varianza aditiva del VC de competencia

ENFOQUE BAYESIANO

Ejemplo de trabajo: Análisis de P. taeda L.

- 20 familias de progenies de polinización abierta de *Pinus taeda* L., de origen Marion (Florida, USA).
- DAP (932) a los 13 años de edad.
- DBCA con 8 repeticiones de 5 árboles en línea y un espaciamiento de 3.5 m × 3.5 m.

Cuadro 4.1. Media y número de observaciones de la base de datos de *Pinus taeda* L.

Árboles con observaciones (*n*)	932
Familias	20
Individuos en el pedigree (*q*)	957
Media diámetro (DAP, cm) (SD)	27.21 (4.56)

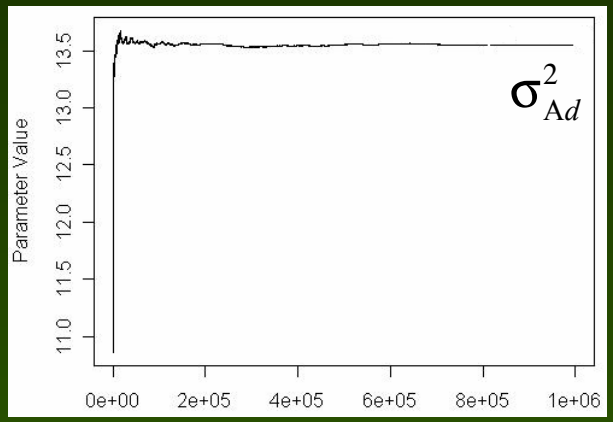
SD= Desvío Estándar

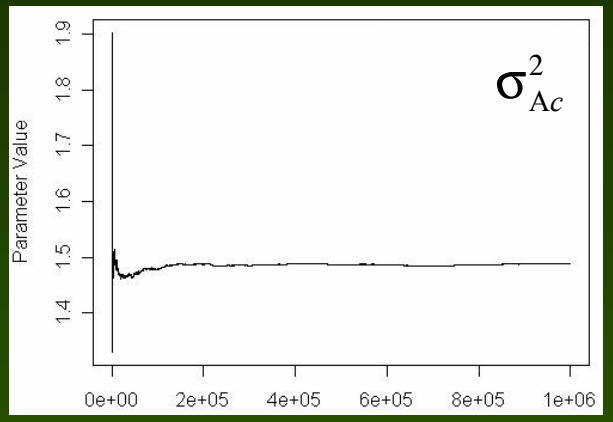
Inferencia Bayesiana posterior

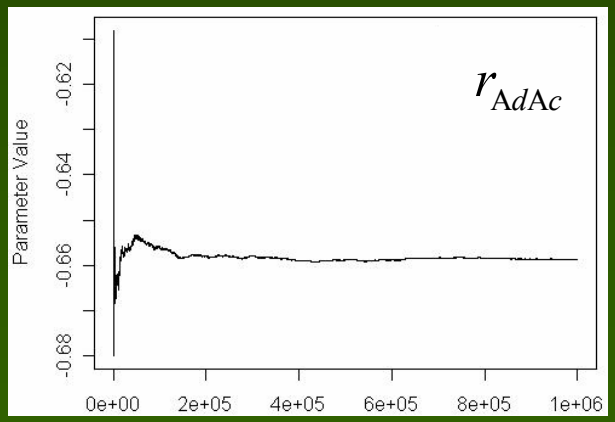
- Se generaron 1.010.000 muestras en una cadena simple, las primeras 10.000 iteraciones fueron descartadas debido al período de calentamiento.
- Los valores para las hipervarianzas de σ^2_{Ad} y σ^2_{e} se calcularon a partir de los mismos datos, utilizando un enfoque Bayesiano empírico *vía* GS.
- Se corrieron distintas cadenas de Markov con diferentes valores *a priori* de σ_{AdAc} (positivos, cero y negativos) y de σ^2_{Ac} (un valor alto y uno bajo, relativos a σ^2_{Ad}). Los resultados para todas las corridas fueron similares; se utilizó la cadena con la mejor convergencia para estimar los componentes de (co)varianza.

Diagnósticos de convergencia

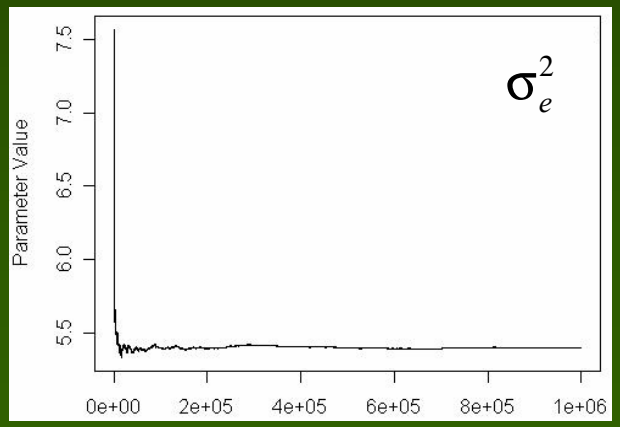
- Promedio acumulado por iteración:







r_A



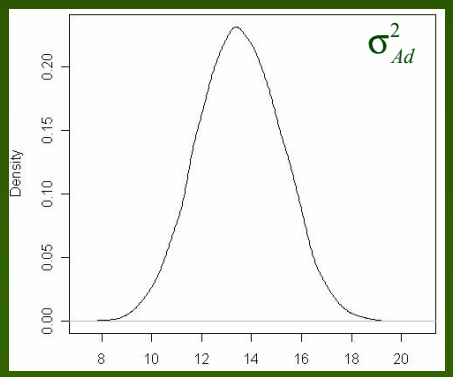
- Puntuaciones *z* (“z-score”) de la prueba de Geweke: 0.667 para σ^2_{Ad}, -2.04 para σ^2_{Ac}, 1.638 para r_{AdAc} y -0.1778 para σ^2_e.

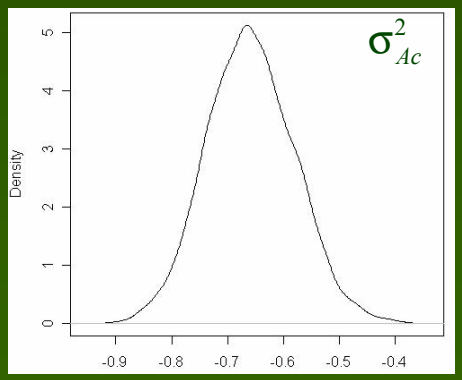
Ejemplo de trabajo: Análisis de P. taeda L.

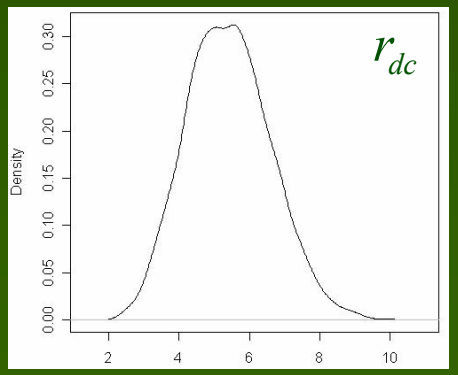
Cuadro 4.2. Estadísticos posteriores para la varianza aditiva directa σ^2_{Ad}, varianza aditiva de competencia σ^2_{Ac}, correlación aditiva directa y de competencia r_{AdAc} y varianza del error σ^2_e.

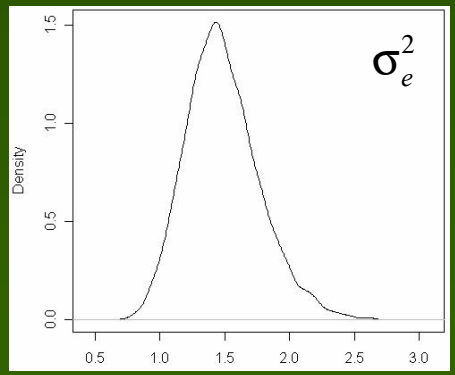
	Media	Mediana	Modo	SD	95% HPD	ESS
σ^2_{Ad}	13.527	13.505	10.908	1.675	10.788 – 16.244	16708
σ^2_{Ac}	1.488	1.459	1.100	0.289	1.060 – 1.998	16332
r_{AdAc}	-0.659	-0.661	-0.747	0.079	-0.785 – -0.523	22372
σ^2_e	5.417	5.369	4.091	1.207	3.511 – 7.485	14150

SD= Desvío Standard; **95% HPD=** Intervalo de alta Densidad Posterior**; ESS**= Tamaño Efectivo de Muestra.



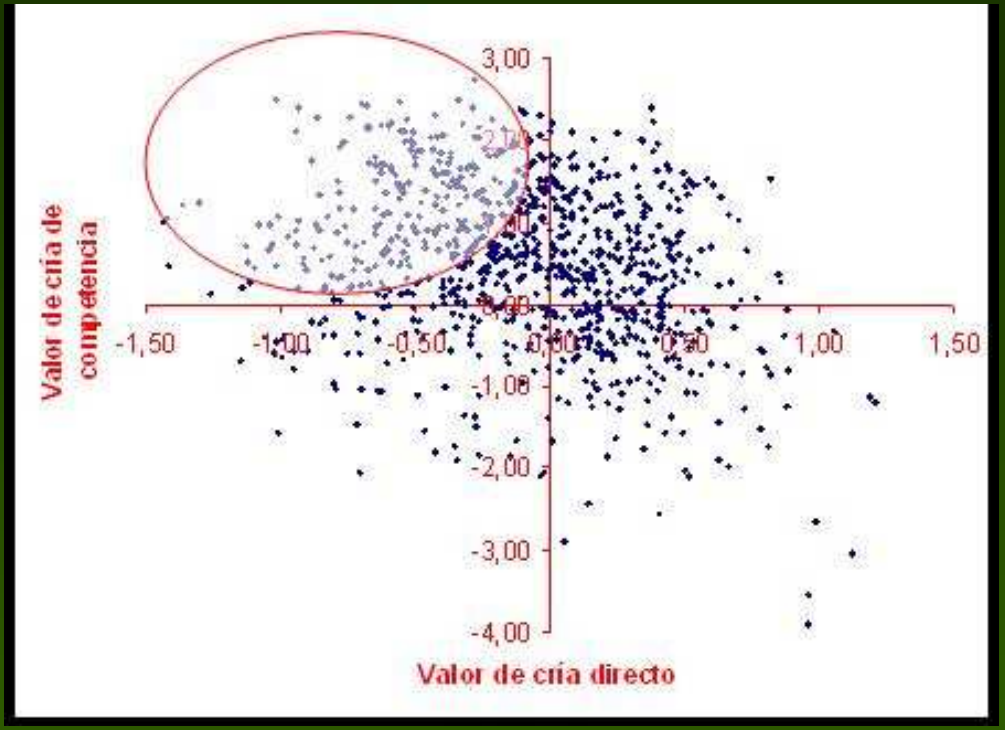






Ejemplo de trabajo: Análisis de P. taeda L.

Objetivo de selección: Genotipos no competitivos para incrementar el rendimiento por unidad de área.



26.46 % (1.21 cm)
29.62 % (1.23 cm)

Exactitud de los valores de cría de progenitores		Exactitud de los valores de cría de progenies		Correlación de Spearman de los valores de cría	
Modelo ED	Modelo EDC	Modelo ED	Modelo EDC	Progenitores	Progenies
0.904	0.941	0.758	0.879	0.985	0.987

+ 4 %

+ 16 %

Capítulo 5

Estimación Bayesiana de una superficie para modelar la tendencia espacial utilizando un modelo mixto semiparamétrico de árbol individual.

Cappa Eduardo P., and Rodolfo J. C. Cantet. (2006). Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. Can. J. For. Res. (*En prensa*).

"State of the art"

- **Costa e Silva, J., Dutkowski, G.W., Gilmour, A.R.** 2001. Can. J. For. Res. 31: 1887-1893.

- **Dutkowski, G.W., Costa e Silva, J., Gilmour, A.R., and Lopez, G.A.** 2002. Can. J. For. Res. 32: 2201-2214.

- **Wand, M.P.** 2003. Comput. Stat. 18: 223–249.

- **Eilers, P.H.C., and Marx, B.D.** 2003. Chemometr. Intell. Lab. Syst. 66: 159-174.

Introducción

En ensayos genéticos la heterogeneidad ambiental sesga la estimación de los parámetros genéticos y la predicción de los valores de cría.

Tradicionalmente la heterogeneidad ambiental se tiene en cuenta a través de diseños *"a priori"*: BCA, BI (Látice).

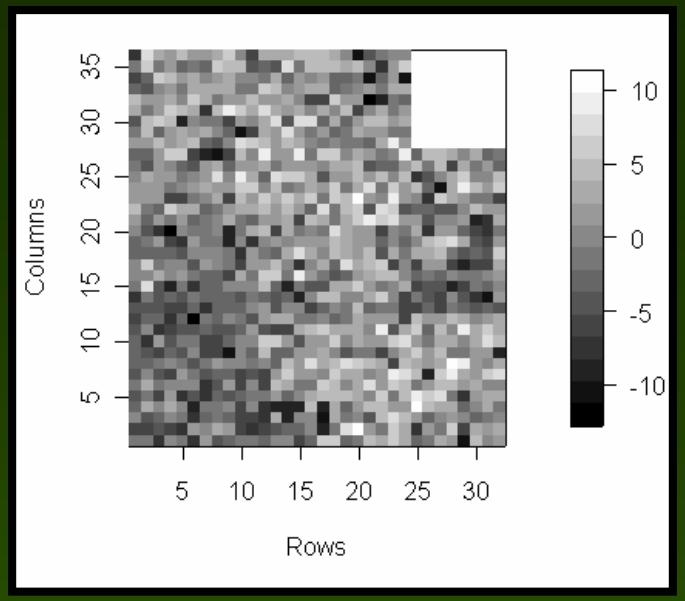
Los diseños "*a priori*" suelen ser ineficientes para remover la tendencia espacial.

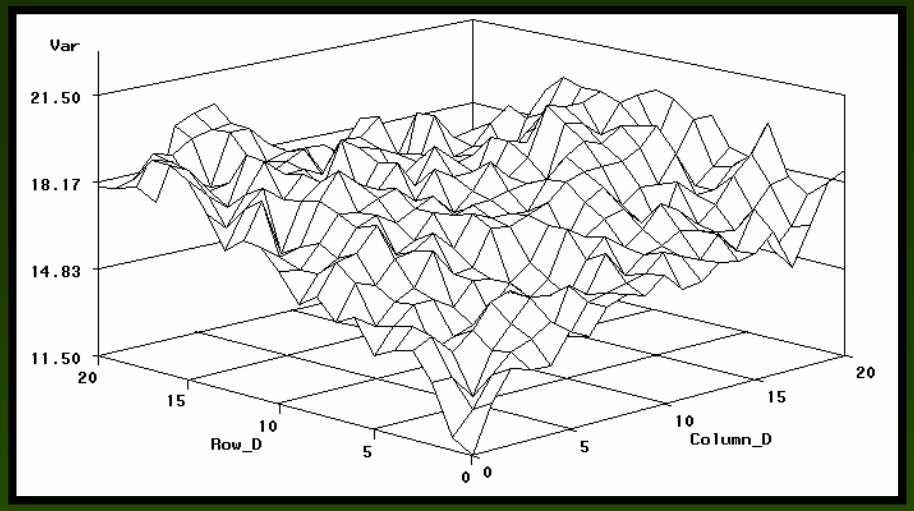
Humedad

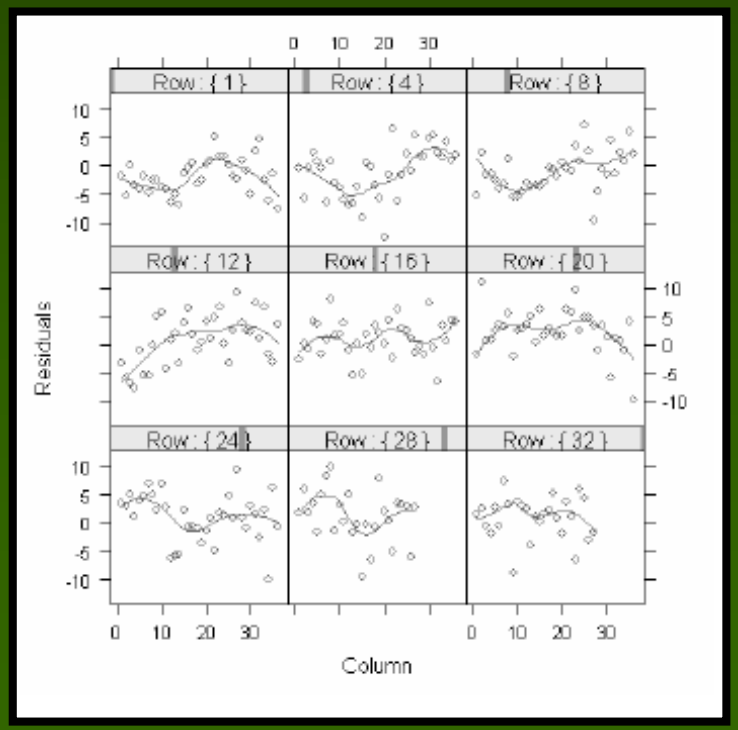
Profundidad del suelo

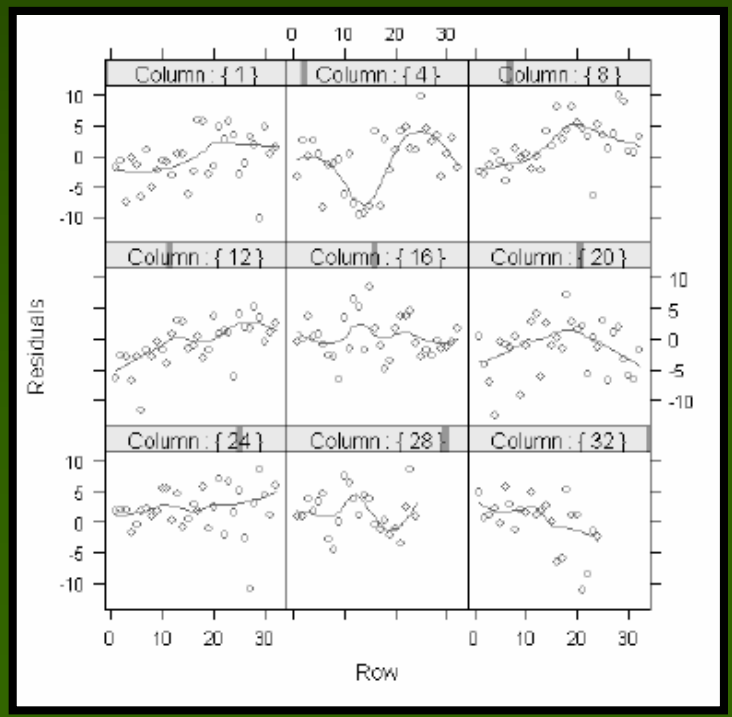
Pendiente

Fertilidad


Columns
Rows
5
10
15
20
25
30
35
10
5
0
-5
-10


Var
21.50
18.17
14.83
11.50
Row_D
Column_D
20
15
10
5
0


Row : { 1 }
Row : { 4 }
Row : { 8 }
Row : { 12 }
Row : { 16 }
Row : { 20 }
Row : { 24 }
Row : { 28 }
Row : { 32 }
Residuals
Column
0
10
20
30
10
5
0
-5
-10


Column : { 1 }
Column : { 4 }
Column : { 8 }
Column : { 12 }
Column : { 16 }
Column : { 20 }
Column : { 24 }
Column : { 28 }
Column : { 32 }
Residuals
Row
0
10
20
30
10
5
0
-5
-10

Para considerar la variación espacial continua, tradicionalmente se utilizan variables clasificatorias (Costa e Silva et al. 2001) o covariables continuas (Dutkowsky et al. 2002) mediante funciones polinómicas de las coordenadas de filas y/o columnas.

El ajuste por polinomios posee algunos inconvenientes dado que produce un ajuste global.

Una solución alternativa es el uso de funciones polinómicas 'a trozos', bajo ciertas condiciones de 'suavidad', llamadas splines.

Ruppert *et al.* (2003) mostraron como expresar smoothing splines con modelos mixtos en una dimensión.

Eilers y Marx (2003) extendieron las splines penalizadas a datos en dos dimensiones utilizando producto tensorial de bases B-Splines cúbicas.

Objetivos

- Describir el ajuste del producto tensorial de bases B-splines en un modelo mixto, con una estructura de (co)varianza entre coeficientes de una grilla de nodos en dos dimensiones.
- Utilizar la expresión obtenida para modelar la variación espacial continua en EGF a través de un modelo mixto de árbol individual.
- Ilustrar las metodologías desarrolladas con datos de diámetros a la altura del pecho en progenies de *Eucalyptus globulus.*

"Producto Tensorial" B-spline en dos dimensiones

Curva con B-spline cúbica en dos dimensiones

Para un cierto conjunto de nodos, la superficie $\alpha(row,col)$ puede ser aproximada utilizando la siguiente expresión matricial:

$$\boldsymbol{vec}\left(\alpha(row,col)\right) = \boldsymbol{Bb}$$

$$\boldsymbol{B} = \left(\boldsymbol{B}_r \otimes \mathbf{1}'_{n_r}\right) \odot \left(\mathbf{1}'_{n_c} \otimes \boldsymbol{B}_c\right)$$

$$\boldsymbol{b} = vec\left(\gamma_{rc}\right)$$

Eilers y Marx (2003) obtuvieron los estimadores penalizados ($\boldsymbol{b}$) $\left(B'B + \lambda_r P_r + \lambda_c P_c\right)\hat{b} = B'y$

$$\boldsymbol{P}_r = \boldsymbol{I}_{n_r} \otimes \boldsymbol{D}'\boldsymbol{D} \quad , \quad \boldsymbol{P}_c = \boldsymbol{D}'\boldsymbol{D} \otimes \boldsymbol{I}_{n_c}$$

Modelo estadístico

MM de árbol individual con B-splines en dos dimensiones:

$$\boldsymbol{y} = \underbrace{\boldsymbol{X}\beta}_{\textit{parte fija}} + \underbrace{\boldsymbol{Ba} + \boldsymbol{Za} + \boldsymbol{e}}_{\textit{parte aleatoria}}$$

Superficie suavizada

MME

$$\begin{bmatrix} \boldsymbol{X`X} & \boldsymbol{X`B} & \boldsymbol{X`Z} \\ \boldsymbol{B`X} & \boldsymbol{B`B} + \boldsymbol{U}^{-1}\lambda & \boldsymbol{B`Z} \\ \boldsymbol{Z`X} & \boldsymbol{Z`B} & \boldsymbol{Z`Z} + \boldsymbol{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X`y} \\ \boldsymbol{B`y} \\ \boldsymbol{Z`y} \end{bmatrix}$$

$$\boldsymbol{U} = (\Sigma_r \otimes \Sigma_c) \qquad \lambda = \sigma_e^2 / \sigma_b^2, \quad \alpha = \sigma_e^2 / \sigma_A^2$$

ENFOQUE BAYESIANO

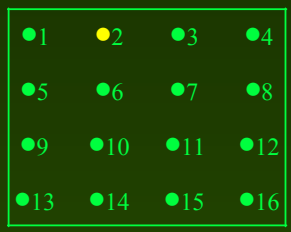
Modelo estadístico

To exemplify, suppose $nx_r = nx_c = 4$, then

One-dimensional covariance structure originally proposed by Green and Silverman (1994, page 13) and then used by Durban et al. (2001) to fit a fertility trend. In this matrix, correlations are non-zero for neighbor knots and are 0 otherwise.

$$\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}_c = \frac{1}{6}\begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix}$$

Grilla de nodos



$$\boldsymbol{U} = \boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_c = \frac{1}{6}\begin{bmatrix}
16 & 4 & 0 & 0 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 4 & 16 & 0 & 0 & 1 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 1 & 0 & 0 & 16 & 4 & 0 & 0 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 4 & 0 & 0 & 4 & 16 & 0 & 0 & 1 & 4 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 4 & 1 & 0 & 0 & 16 & 4 & 0 & 0 & 4 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 0 & 0 & 4 & 16 & 0 & 0 & 1 & 4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 1 & 0 & 0 & 16 & 4 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 0 & 0 & 4 & 16
\end{bmatrix}$$

Ejemplo de trabajo: Análisis de E. globulus

Datos

- **36 Flías. de polinización abierta de *Eucalyptus globulus* spp. *globulus.***
- **Suelo Paleudol petrocálcico fino, con un horizonte de tosca sub-superficial a profundidad variable.**
- **DAP (1080) a los 6 años de edad.**
- **El ensayo fue plantado en una grilla regular de 32 filas por 36 columnas.**

Modelos de análisis

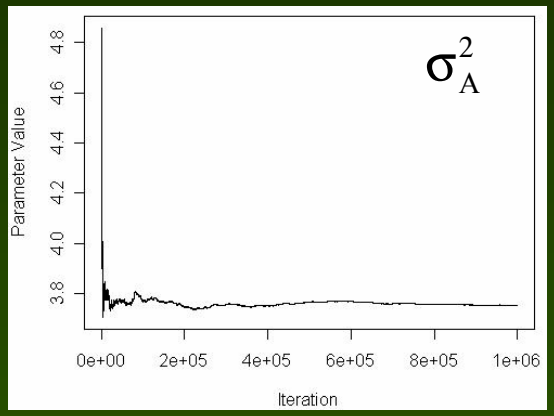
- **Se evaluaron cuatro modelos aditivos de árbol individual.**
- **Para comparar el ajuste se utilizó el "*Deviance Information Criterion*" (DIC, Spiegelhalter et al. 2002).**
- **Se obtuvo la exactitud de predicción y la correlación de Spearman de los valores de cría y la ganancia genética.**

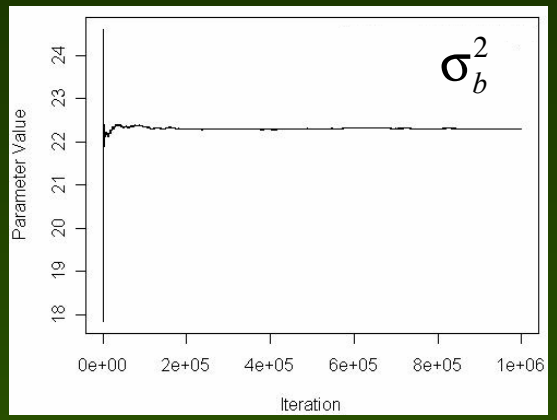
Inferencia Bayesiana posterior

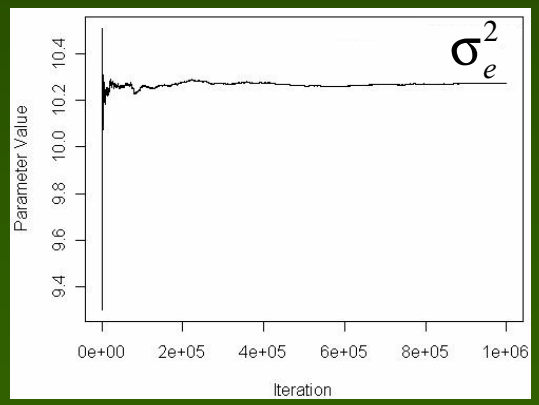
- Se generaron 1.010.000 muestras en una cadena simple, las primeras 10.000 iteraciones fueron descartadas debido al período de calentamiento.
- Los valores de las hipervarianzas de δ^2_A y δ^2_e se calcularon a partir de los mismos datos, utilizando un enfoque Bayesiano empírico *vía* GS.
- Dado que no se posee información *a priori* de la hipervarianza de δ^2_b, se estudiaron diferentes valores en el intervalo [0, δ^2_e) y se encontró que el algoritmo converge siempre a la misma media posterior de σ^2_b.

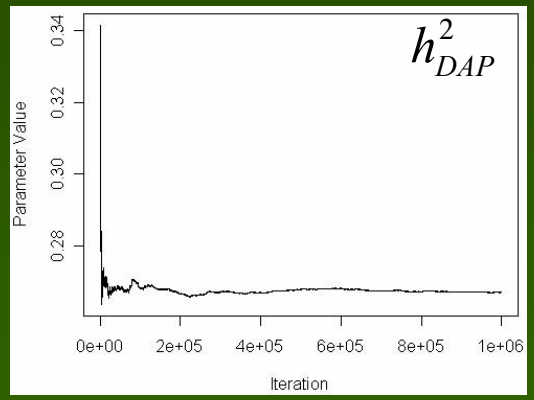
Diagnósticos de convergencia

- Promedio acumulado por iteración:





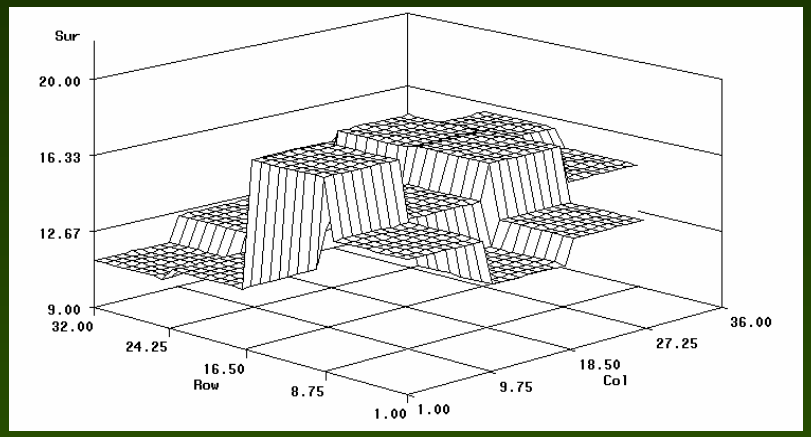




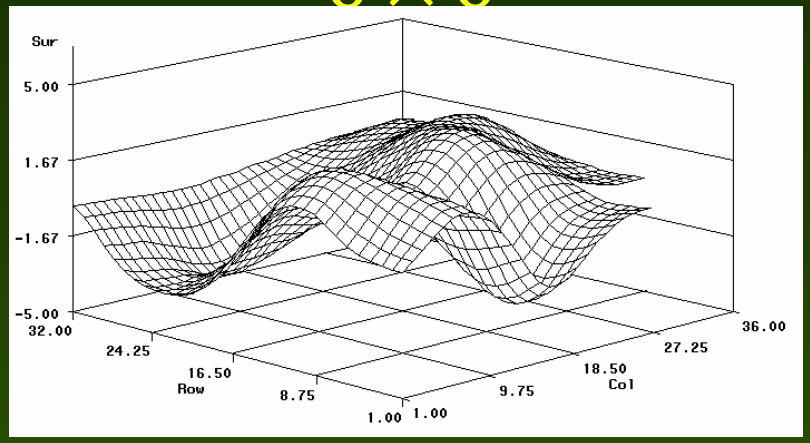
- Puntuaciones *z* ("*z*-score") de la prueba de Geweke: 1.217 para σ^2_A, 0.611 para σ^2_b, -1.132 para σ^2_e y 1.186 para h^2_{DAP}

Ejemplo de trabajo: Análisis de E. globulus

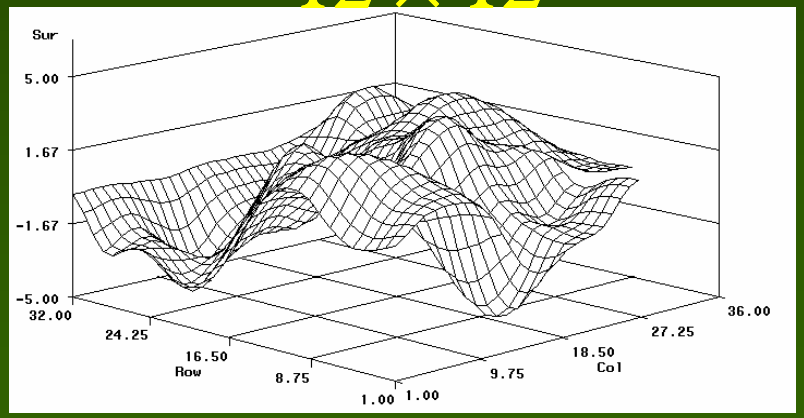
Blocks



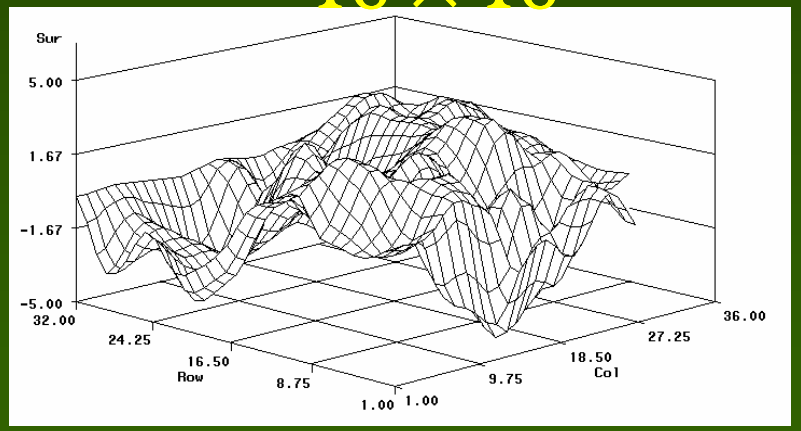
8×8



12×12



18×18



Cuadro 5.1. Criterio de Información de la Desviación (DIC), para los cuatro modelos analizados.

Modelo[a]	Bloques	8 × 8	12 × 12	18 × 18
DIC	3152.660	2868.6380	2833.463	2835.1230

Ejemplo de trabajo: Análisis de E. globulus

Cuadro 5.2: Estadísticos posteriores de la varianza genética aditivas (σ^2_A), varianza de los coeficientes del producto tensorial de B-splines (σ^2_b), varianza del error (σ^2_e) y heredabilidad del DAP (h^2_{DAP}).

Model[a]	Parm[b]	Media	Mediana	Modo	SD[c]	95% HPD[d]	ESS[e]
1	σ^2_A	1.835	1.801	1.609	0.37149	1.291 – 2.503	24119
	σ^2_e	23.043	20.144	14.070	8.69251	15.182 – 40.520	87274
	h^2_{DAP}	0.080	0.079	0.084	0.02520	0.040 – 0.123	43572
3	σ^2_A	3.754	3.643	2.933	1.00390	2.310 – 5.573	16474
	σ^2_b	22.317	21.649	23.716	5.47972	14.682 – 32.132	109973
	σ^2_e	10.275	10.301	9.900	1.01309	8.558 - 11.871	23568
	h^2_{DAP}	0.267	0.261	0.244	0.06872	0.167 - 0.389	16519

Tabla 3. Exactitud y correlación de los valores de cría para progenitores y progenies.

Exactitud de los valores de cría para Progenitores		Exactitud de los valores de cría para Progenies		Correlación de los valores de cría	
Bloques	**12 × 12**	**Bloques**	**12 × 12**	**Progenitores**	**Progenies**
0.405	0.612	0.323	0.542	0.972	0.939

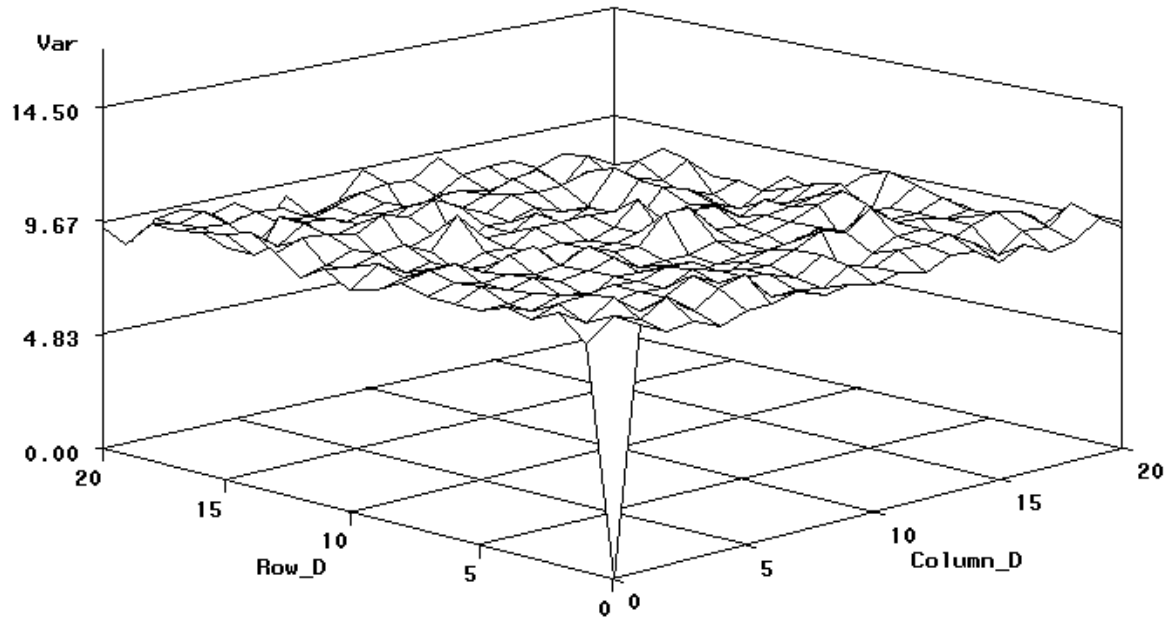
+ 66 %

+ 60 %

(0.23 cm) – (0.77 cm)

Ejemplo de trabajo: Análisis de E. globulus

Figura 6.2. Variograma muestral indicando la eliminación de la variación espacial continua producida por el modelo con 12 nodos para filas y 12 para columnas.



Conclusiones / Contribuciones

Capítulo 3:

- Se implementó el análisis Bayesiano vía el algoritmo de FCG en modelos de árbol individual multicarácter con datos faltantes;
- se obtuvo una expresión del DIC que permite seleccionar modelos de árbol individual con caracteres normales múltiples;
- inéditamente en la EG forestal se implementó, dentro de un modelo mixto gausiano, la parametrización de Hill (1982) para conocer qué efectos genéticos (A, D, E) afectan los caracteres estudiados, con los distintos genotipos parentales y sus cruzas;
- en datos de un ensayo de progenies de pinos híbridos, el modelo aditivo fue el de mejor ajuste y menor grado de complejidad. Medias marginales posteriores de h^2_{DAP} y h^2_{Altura}: 0.116 y 0.139. Media marginal posterior de r_{A} y r_{E}: 0.524 y 0.394.

Capítulo 4:

- Se logró una metodología que permite separar los efectos genéticos aditivos directos de los aditivos de competencia;
- se logró estimar con éxito los componentes de varianza y covarianza entre los efectos aditivos directos y de competencia utilizando un enfoque Bayesiano, a través del muestro de *Gibbs*;
- el modelo genético propuesto considera correctamente el efecto de una reducción en el número de competidores y las relaciones aditivas de cualquier árbol con sus competidores y entre los competidores;
- en datos de 20 progenies de *P. taeda* L. para el carácter DAP a la edad de 13 años, la σ^2_{Ad} fue 10 veces más grande que la σ^2_{Ac}. La correlación genética entre ambos efectos fue considerable (-0.659). Las exactitudes de los VC y las ganancias genéticas relativas fueron 4 y 16 % y 27 y 30 % superiores para progenitores y progenies, a favor del modelo con efectos de competencia.

Capítulo 5:

- Se logró incorporar de modo original, flexible y sencillo, el producto tensorial de bases B-splines cúbicas en MM de árbol individual;
- se obtuvieron P-splines en dos dimensiones con un enfoque Bayesiano;
- se extendió el producto tensorial de bases B-splines a un MM de árbol individual de forma tal que considere la variabilidad espacial continua;
- se mostró mediante la comparación de modelos por el estadístico Bayesiano DIC, que difícilmente la variación espacial continua en gran escala pueda removerse con la técnica "*a priori*" de bloqueo;
- el ajuste de una superficie en datos de progenies de *E. globulus*, redujo la media posteriores de σ^2_e aumentó la σ^2_A y h^2_{DAP}, incrementó las exactitudes de los VC y las ganancias genéticas relativas en un 66 y 60 % y 0.23 y 0.77 cm de DAP para progenitores y progenies.

Agradecimientos

- **A mi director: Dr. Rodolfo Cantet.**
- **A la Agencia Nacional de Promoción Científica y Tecnológica, FONCyT.**
- **A los docentes de la Cátedra de Mejoramiento Genético Animal. Facultad de Agronomía-UBA.**
- **Al Centro de Investigaciones y Experiencias Forestales, CIEF.**
- **A Bosques Cultivados, INTA Castelar.**
- **A la Ing. Ftal. Maria Elena Gauchat, INTA Montecarlo.**


Muchas Gracias