APLICACIÓN DE LOS MODELOS MIXTOS A UN CASO PRÁCTICO DE MODELIZACIÓN DEL CRECIMIENTO Y PRODUCCIÓN DE LAS MASAS FORESTALES

R. Calama Sainz y G. Montero González

Grupo Selvicultura Mediterránea. CIFOR-INIA. Ctra. Coruña km 7,5. 28040–MADRID (España). Correo electrónico: rcalama@inia.es

Resumen

Los modelos mixtos conforman una herramienta de interés para la modelización del crecimiento y la producción de variables de interés forestal, al permitir el ajuste de funciones lineales sobre observaciones que no verifiquen el supuesto de independencia. En el presente trabajo se presenta una aproximación teórica a los modelos mixtos y se muestra un caso práctico de aplicación para la modelización de la relación diámetro–altura.

Palabras clave: *Modelización, Estocástico, Componentes aleatorios*

INTRODUCCIÓN

Gran parte de los modelos que se aplican en la actualidad en el ámbito de la Ciencia Forestal se formulan como un modelo general lineal (SEARLE, 1971):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad (1)$$

Donde y representa el vector n x 1 que contiene las observaciones correspondientes a la variable aleatoria de interés del modelo; **X** es una matriz n x p, que incluye el valor de las covariables explicativas; $\boldsymbol{\beta}$ es un vector p x 1 que incluye los parámetros asociados a las covariables del modelo y **e** es un vector n x 1 que incluye los términos residuales del error. La formulación anterior se puede corresponder con una regresión lineal múltiple o con un modelo de análisis de la varianza. El primer y segundo momento (esperanza y varianza) para la distribución de **y** son:

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta};\ \text{Var}(\mathbf{y}) = \text{var}(\mathbf{e}) = \mathbf{V}$$

Donde **V** es la matriz n x n de varianza covarianza para las observaciones. El objetivo principal de la resolución de un modelo general lineal es obtener un estimador para el vector $\boldsymbol{\beta}$. El estimador lineal, insesgado y de mínima varianza es el estimador de mínimos cuadrados generalizados $\hat{\boldsymbol{\beta}}_{MCG}$, definido como:

$$\hat{\boldsymbol{\beta}}_{MCG} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^- \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \qquad (2)$$

Sin embargo, en el ajuste de la mayor parte de los modelos se utiliza el estimador de mínimos cuadrados ordinarios (MCO), $\hat{\boldsymbol{\beta}}_{MCO}$, que queda definido por:

$$\hat{\boldsymbol{\beta}}_{MCO} = (\mathbf{X}^T \mathbf{X})^- \mathbf{X}^T \mathbf{y} \qquad (3)$$

$\hat{\boldsymbol{\beta}}_{MCO}$ es un estimador insesgado de $\boldsymbol{\beta}$, pero no de mínima varianza. Únicamente en el caso de que la estructura de **V** sea del tipo $\sigma^2\mathbf{I}_n$, donde $\mathbf{I}_n$ representa la matriz identidad de orden n, $\hat{\boldsymbol{\beta}}_{MCO}$ coincide con $\hat{\boldsymbol{\beta}}_{MCG}$. El cumplimiento de esta estructura requiere dos supuestos básicos: homocedasticidad en la varianza de los residuos e independencia entre los residuos. El aplicar téc-

ISSN: 1575-2410

nicas de MCO sin que se verifique el cumplimiento de ambos supuestos implica una estimación sesgada del intervalo de confianza para el estimador del parámetro, con el peligro de considerar como significativas variables que realmente no lo son. La heterocedasticidad se corrige aplicando transformaciones de la variable dependiente o incluyendo matrices de ponderación en el estimador MCO. La falta de independencia entre las observaciones requiere la aplicación de técnicas de mínimos cuadrados generalizados y, por tanto, la definición de una estructura para la matriz **V** y la estimación de sus componentes.

FALTA DE INDEPENDENCIA EN LOS DATOS UTILIZADOS EN LA MODELIZACIÓN FORESTAL

En la toma de datos para la construcción de modelos en el ámbito forestal es muy común tomar distintas mediciones de la variable de interés en una misma unidad de muestreo. La unidad de muestreo puede ser un árbol, en el que se mide el diámetro de sección (árbol tipo) o el número de anillos (análisis de tronco) en distintos puntos del fuste; o puede ser una parcela, donde de distintos árboles se mide un atributo determinado (altura, volumen individual, producción de piña…). A su vez, las parcelas pueden agruparse en rodales, montes o regiones….

Las observaciones procedentes de una misma unidad de muestreo presentan alta correlación, originada por una serie de factores que influyen comunes a todas las mediciones procedentes de una misma unidad de muestreo. Estos factores pueden estar relacionados con el tamaño del árbol, su genotipo, atributos del rodal, variables de tipo edáfico, climático… En el caso de que no todos estos factores puedan incluirse en el modelo los residuos procedentes de una misma unidad tenderán a ser parecidos entre sí, la matriz V no tendrá estructura del tipo $\sigma^2\mathbf{I_n}$, y el estimador $\hat{\boldsymbol{\beta}}_{\mathbf{MCO}}$ no será un estimador eficiente de $\boldsymbol{\beta}$.

MODELOS MIXTOS LINEALES

Una aproximación al problema del ajuste de modelos lineales sobre datos no independientes viene dada por los modelos mixtos lineales (SEARLE et *al.,* 1992; LITTELL et *al.,* 1996). Los modelos mixtos se fundamentan en la partición del error no explicado en una componente común a las observaciones procedentes de una misma unidad de muestreo y un término residual del error, propio de cada observación, y, en principio, independiente de los términos residuales del resto de las observaciones. El modelo mixto lineal incluye en su formulación parámetros fijos, comunes a toda la población, y parámetros aleatorios, específicos de cada unidad de muestreo. Los parámetros aleatorios se consideran realizaciones aleatorias de un proceso de media cero, y cuya varianza define la componente del error asociada a la unidad de muestreo.

La formulación de un modelo mixto lineal para las ni observaciones incluidas en la unidad de muestreo "*i*" es:

$$\mathbf{y_i} = \mathbf{X_i}\,\boldsymbol{\beta} + \mathbf{Z_i}\,\mathbf{b_i} + \mathbf{e_i} \qquad (4)$$

Donde $\mathbf{y_i}$ es el vector n_i x 1 $[y_{i1},\ldots,y_{ini}]^T$ que contiene las observaciones medidas en la unidad X_i; es la matriz n_i x p donde su fila j, $\mathbf{x_{ij}} = [x_{1ij},\ldots,x_{pij}]$, representa el valor de las covariables explicativas para la observación j de la unidad i; $\boldsymbol{\beta}$ es un vector p x 1 de parámetros fijos; $\mathbf{Z_i}$ es la matriz de n_i x q en la que su fila j, $\mathbf{z_{ij}} = [z_{1ij},\ldots,z_{qij}]$, representa el valor de las covariables explicativas asociadas a los parámetros aleatorios; $\mathbf{b_i}$ es un vector q x 1 de parámetros aleatorios específicos para la unidad i $[u_{1i},\ldots,u_{qi}]^T$, tal que $\mathbf{b_i} \sim \mathbf{N\,(0\ ,\ D_i)}$; $\mathbf{e_i}$ es un vector n_i x 1 de términos del error con distribución $\mathbf{e_i} \sim \mathbf{N\,(0\ ,\ R_i)}$. $\mathbf{R_i}$ es una matriz n_i x n_i, normalmente del tipo $\sigma^2{}_e\mathbf{I_{ni}}$; $\mathbf{D_i}$ es la matriz q x q de varianza de los parámetros aleatorios, y en el caso más general tiene la forma:

$$D_i = \begin{pmatrix} \sigma^2_{u1} & \sigma_{u1u2} & \ldots & \sigma_{u1uq} \\ \sigma_{u1u2} & \sigma^2_{u2} & \ldots & \sigma_{u2uq} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{u1uq} & \sigma_{u2uq} & \ldots & \sigma^2_{uq} \end{pmatrix}$$

La formulación del modelo mixto lineal para las r unidades de muestreo que componen la muestra (n observaciones en total) es:

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{Z}\,\mathbf{b} + \mathbf{e} \qquad (5)$$

Donde **y** es el vector n x 1 $[\mathbf{y_1}^T,\ldots,\mathbf{y_r}^T]$T; **X** es la matriz n x p, cuyas filas son los vectores $\mathbf{x_{ij}}$; $\boldsymbol{\beta}$ es el vector de parámetros fijos antes definido; **Z** es la matriz en bloque diagonal cuyos

bloques son las r matrices $\mathbf{Z_i}$, **b** es el vector $[\mathbf{b_1}^T,\ldots,\mathbf{b_r}^T]^T$ que incluye los q parámetros aleatorios para cada una de las r parcelas, tal que $\mathbf{b \sim N (0 , D)}$; e es un vector **n x 1** que contiene los términos residuales del error, tal que $\mathbf{e \sim N (0 , R)}$. En el caso más común **D** es la matriz en bloque diagonal compuesta por las r matrices $\mathbf{D_i}$ $\mathbf{(D_1=\ldots=D_i=\ldots=D_r)}$, y **R** es una matriz diagonal n x n, cuya única componente es la varianza residual del modelo σ^2_e. Los dos primeros momentos asociados a la distribución de $\mathbf{y_i}$ y de **y** son:

$$E(\mathbf{y_i}) = E(\mathbf{y}|\mathbf{b_i}) = \mathbf{X_i}\,\boldsymbol{\beta} + \mathbf{Z_i}\,\mathbf{b_i};\ \text{Var}(\mathbf{y_i}) = \mathbf{V_i} = \text{var}(\mathbf{Z_i}\,\mathbf{b_i} + \mathbf{e_i}) = \mathbf{Z_i}\,\mathbf{D_i}\,\mathbf{Z_i}^T + \mathbf{R_i}$$

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta};\ \text{Var}(\mathbf{y}) = \mathbf{V} = \text{var}(\mathbf{Zb} + \mathbf{e}) = \mathbf{Z\,D\,Z}^T + \mathbf{R}$$

MOTIVACIONES PARA EL USO DE UN MODELO MIXTO LINEAL

- Los modelos mixtos permiten la estimación eficiente del vector de parámetros fijos $\hat{\boldsymbol{\beta}}$ que definen el patrón común de la población, aplicando técnicas de MCG:

$$\hat{\boldsymbol{\beta}} = (X^T \hat{V}^{-1} X)^- X^T \hat{V}^{-1} y \qquad (6)$$

- Predicción eficiente del vector de parámetros aleatorios $\hat{\mathbf{b}}_\mathbf{i}$ específico para la unidad de muestreo i, que define el patrón común de desviación de los componentes de la unidad con respecto de la media. En la predicción se usa la teoría de los mejores predictores lineales insesgados (SEARLE et *al.,* 1992):

$$\hat{\mathbf{b}} = \mathbf{DZ}^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\,\hat{\boldsymbol{\beta}}) \qquad (7)$$

- Descomposición de la varianza total en la varianza entre las unidades de muestreo y la varianza dentro de cada unidad. Estimación de los componentes de la varianza que definen las matrices de varianza estimadas $\hat{\mathbf{D}}$ y $\hat{\mathbf{R}}$. La matriz de varianza estimada para las observaciones queda definida como $\hat{\mathbf{V}} = \mathbf{Z}\hat{\mathbf{D}}\,\mathbf{Z}^T + \hat{\mathbf{R}}$. En la estimación de los componentes de la varianza se aplican métodos de máxima verosimilitud (HARVILLE, 1977) o métodos ANOVA (HENDERSON, 1953).
- Calibración del modelo para una unidad k no muestreada, al poder predecir el vector de parámetros aleatorios $\hat{\mathbf{b}}_\mathbf{k}$ si se dispone de al menos una medición de la variable de interés en la nueva unidad.

APLICACIÓN DE UN MODELO MIXTO A UN CASO PRÁCTICO DE MODELIZACIÓN DE LA ALTURA EN FUNCIÓN DEL DIÁMETRO

Para mostrar la aplicación de un modelo mixto lineal se utilizan los datos correspondientes a la altura individual y el diámetro normal de los árboles contenidos en 125 parcelas temporales de producción instaladas por el INIA en masas regulares de *Pinus pinea* en la provincia de Valladolid. Del total, 115 parcelas (con 2300 árboles) se usan como muestra de ajuste y 10 parcelas (200 árboles) se reservan como muestra de validación. Para modelizar la relación entre el diámetro y la altura se utiliza la expresión propuesta por CURTIS (1967):

$$\log(h) = a_1 + a_2 \log(d) + e$$

Donde h es la altura del árbol (m) y d es el diámetro normal (cm). En primer lugar suponiendo homoscedasticidad e independencia se procede al ajuste de la relación aplicando regresión lineal por mínimos cuadrados ordinarios. El modelo resultante es (el término entre paréntesis se refiere al error estándar de estimación de los parámetros):

$$\log(h_j) = \underset{(0.0283)}{-0.9881} + \underset{(0.0085)}{0.9404}\log(d_j) + e_j; \qquad (8)$$

$$R^2=0.841;\ \sigma^2_e = 0.0307$$

El análisis gráfico de los residuos (fig. 1-1) no muestra desviaciones patentes respecto al patrón de homoscedasticidad en la varianza σ^2_e de los residuos. Sin embargo, en la fig 1-2 queda señalado (sobre una muestra de parcelas) como los residuos procedentes de una misma parcela son más parecidos entre sí que la media (es decir tienden a ser todos positivos o negativos), lo que indica falta de independencia entre los residuos.

Ante esta falta de independencia se propone la formulación de la relación altura-diámetro como un modelo de coeficientes aleatorios, caso particular del modelo mixto lineal en el que **Z** contiene únicamente unos y ceros, o si contiene covariables explicativas, éstas también forman parte de **X**. Bajo este supuesto, la expresión del modelo de coeficientes aleatorios para la observación j medida en la unidad i sería:

$$\log(h_{ij}) = a_1 + u_i + (a_2 + v_i)\log(d_{ij}) + e_{ij}$$

$$\text{Donde}\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D_i} = \begin{pmatrix} \sigma^2_u & \sigma_{uv} \\ \sigma_{uv} & \sigma^2_v \end{pmatrix}\right];\ e_{ij} \sim N(0,\sigma^2_e)$$

Al resolver el modelo anterior utilizando el procedimiento MIXED del paquete estadístico SAS obtenemos estimadores para a_1 y a_2 (componentes de $\hat{\boldsymbol{\beta}}$), los componentes de las matrices $\mathbf{D_i}$ y $\mathbf{R_i}$, y el valor predicho de u_i y v_i para cada unidad de muestreo:

$$\log(h_{ij}) = \underset{(0.0801)}{0.3367} + u_i + \underset{(0.0206)}{(0.5459} + v_i)\log(d_{ij}) + e_{ij}; \quad (9)$$

$\sigma^2_u = 0.5580$; $\sigma^2_v = 0.0335$; $\sigma_{uv} = -0.1323$; $\sigma^2_e = 0.0064$; EF= 0.969

Al comparar estos resultados con los obtenidos se observa que el estimador MCO subestima el error de estimación de los parámetros. En el modelo mixto la varianza del error se descompone en una componente asociada a la unidad de muestreo, definida por σ^2_u, σ^2_v y σ_{uv}, y un término residual σ^2_e, indicando que la variabilidad entre unidades de muestreo explica gran parte del error total. EF indica la eficiencia (estadístico análogo al R^2 de la regresión lineal) del modelo mixto completo, que para una variable y se define como:

$$EF = 1 - \frac{SSE}{SST} = 1 - \frac{y_{pred} - y_{obs}}{y_{pred} - y_{medio}} \qquad (10)$$

Si se desea hacer estimaciones en nuevas unidades es posible aplicar la parte del modelo mixto que únicamente incluye los parámetros fijos (modelo de efectos fijos). La eficiencia predictiva del modelo se puede incrementar si se predice o explica el valor de los parámetros u_i y v_i para las nuevas unidades. Estos parámetros definen la desviación de las observaciones de la unidad de muestreo con respecto al valor medio para la población, desviación debida a una serie de atributos de la unidad. La identificación y entrada de estos atributos en el modelo reducirá la variabilidad no explicada por la parte fija del modelo.

Una tercera alternativa es la posibilidad de predecir el valor de los parámetros aleatorios u_i y v_i para una nueva unidad de muestreo a partir de una serie de mediciones complementarias de la variable dependiente medidas en la nueva unidad. En esta calibración del modelo se utiliza la expresión (7). Supongamos que en una nueva unidad k se mide en dos árboles el diámetro, d_{k1} y d_{k2} y la altura, h_{k1} y h_{k2}. El valor de los parámetros predichos u_k y v_k sería:

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix} = \underbrace{\begin{pmatrix} 0.5580 & -0.1323 \\ -0.1323 & 0.0335 \end{pmatrix}}_{\hat{\mathbf{D}}_K} \underbrace{\begin{pmatrix} 1 & 1 \\ \log(d_{k1}) & \log(d_{k2}) \end{pmatrix}}_{\mathbf{Z_k}^T}$$

$$\underbrace{\hat{\mathbf{V}}_k^{-1}}\left[\underbrace{\begin{pmatrix} h_{k1} \\ h_{k2} \end{pmatrix}}_{\mathbf{y_k}} - \underbrace{\begin{pmatrix} 1 & \log(d_{k1}) \\ 1 & \log(d_{k2}) \end{pmatrix}}_{\mathbf{X_k}} \underbrace{\begin{pmatrix} 0.3367 \\ 0.5459 \end{pmatrix}}_{\beta}\right]$$

Donde $\hat{\mathbf{V}}_k = \mathbf{Z_k}\,\hat{\mathbf{D}}_k\,\mathbf{Z_k}^T + \hat{\mathbf{R}}_k$, con $\hat{\mathbf{R}}_k = 0.0064\,\mathbf{I_2}$

En la figura 2 se compara el resultado de la calibración del modelo a partir de dos mediciones por parcela en dos de las parcelas incluidas en la muestra de validación con el obtenido al aplicar el modelo de efectos fijos (aplicando en todos los casos la transformación antilogarítmica). Al considerar la calibración para las 10 parcelas de la muestra de validación, se com-

Figura 1-1

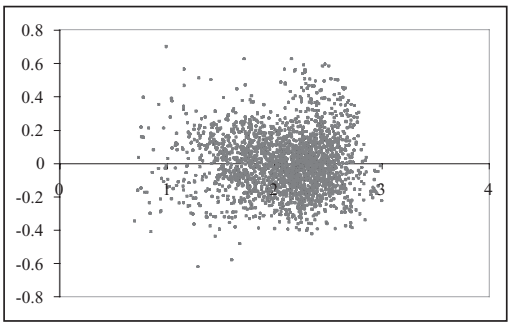

Figura 1-2

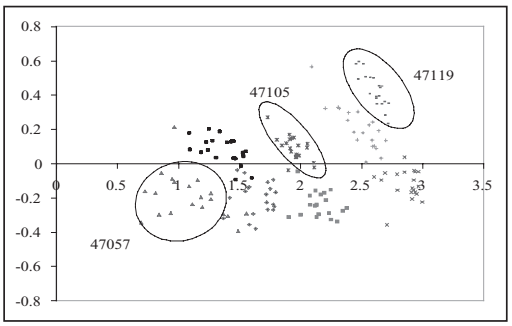

Figura 1-1. *Residuos correspondientes al ajuste por regresión lineal MCO del modelo (8)*
Figura 1-2. *Residuos del modelo (8) correspondientes a un grupo de parcelas de la muestra*

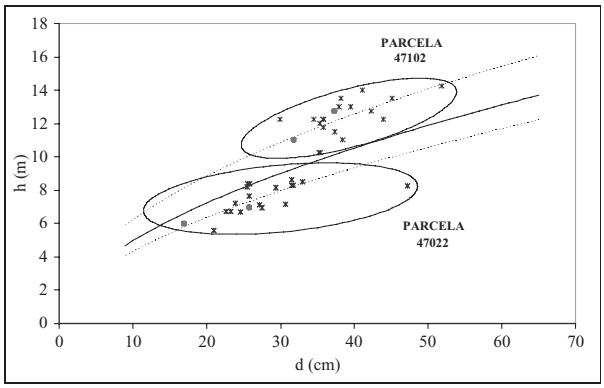

Figura 2. *Modelo de efectos fijos (continua) y modelo calibrado (discontinua) a partir de dos árboles (puntos) en dos parcelas de la muestra de validación*

prueba que la eficiencia predictiva del modelo que incluye los parámetros aleatorios predichos a partir de un único árbol por parcela (EF=0.928) mejora sustancialmente el resultado obtenido al aplicar el modelo de efectos fijos (EF=0.618).

CONCLUSIONES

Los modelos mixtos constituyen una alternativa al ajuste por regresión mediante mínimos cuadrados ordinarios cuando las observaciones no son independientes entre sí. El ajuste de modelo mixto permite obtener estimadores eficientes de los parámetros fijos del modelo, y predecir parámetros aleatorios específicos de cada unidad de muestreo, que reflejan el patrón de desviación con respecto de la media. La posibilidad de predecir estos parámetros y calibrar el modelo para nuevas localizaciones les confiere gran utilidad en el campo de la aplicación práctica de los modelos a la gestión forestal, al obtener estimaciones fiables sin necesidad de medir covariables adicionales.

BIBLIOGRAFÍA

CURTIS, R.O.;1967. Height-diameter and height-diameter-age equations for second growth Douglas Fir. *For. Sci.* 13: 365-375.

HARVILLE, D.A.; 1977. Maximum-likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Assoc.* 80: 132-138.

HENDERSON, C.R.; 1953. Estimation of variance and covariance components. *Biometrics* 9: 226-252.

LITTELL, R.C.; MILLIKE, G.A., STROUP, W.W., WOLFINGER, R.D. 1996. SAS ® *Ssytem for mixed models.* SAS Institute Inc. Cary NC.

SEARLE, S.R.; 1971. *Linear models.* John Wiley and Sons. New York.

SEARLE, S.R.; CASELLA, G. & MCCULLOCH, C.E; 1992. *Variante components.* John Wiley and Sons. New Cork.