CONTROL DE LA AUTOCORRELACIÓN ESPACIAL MEDIANTE DISEÑOS EXPERIMENTALES Y MÉTODOS DE ANÁLISIS ESPACIAL EN ENSAYOS DE PROGENIE DE *PINUS PINASTER*

Raúl de la Mata Pombo [1], Rafael Zas Arregui [1] y Esther Merlo Sánchez [2]

[1] Centro de Investigacións Ambientais. CINAM-Lourizán. Apdo 127. 36080-PONTEVEDRA (España). Correo electrónico: rmata.cifal@siam-cma.org
[2] CIS-Madeira. Avda. de Galicia 5. Parque Tecnológico de Galicia. 32901-SAN CIBRAO DAS VIÑAS (Ourense, España)

Resumen

Es frecuente que los ensayos genéticos forestales evalúen variables espacialmente autocorrelacionadas, lo que supone la violación del requisito de independencia del análisis de varianza convencional. Para absorber parte de la variación espacial del ambiente se recurre a la instalación de dichos ensayos bajo diseños experimentales. Sin embargo, ante autocorrelaciones espaciales intensas y a pequeña escala, éstos pueden no ser suficientes. En el presente trabajo se comprueba la eficiencia estadística de distintos diseños experimentales en referencia a dos técnicas de análisis espacial. Se analiza la altura 2 años después de la plantación en tres ensayos de progenie de *Pinus pinaster* Ait. situados en tres localidades del noroeste de la península ibérica, instalados bajo un diseño de bloques incompletos resoluble. Los semivariogramas de los residuos revelan una estructura espacial no aleatoria en los tres sitios, comprobándose que el diseño de bloques incompletos mejora la eficiencia estadística del análisis en todos los casos, si bien ante patrones intensos resulta claramente insuficiente. En esta situación, el modelo de errores espacialmente correlacionados mejora notablemente el análisis y en todo caso el procedimiento de ajuste espacial ISA (*Iterative Spatial Analysis*) muestra ser más eficiente que las técnicas anteriores.

Palabras clave: *Patrón espacial, Eficiencia estadística, Geoestadística, Semivariograma, Pino marítimo*

INTRODUCCIÓN

Una de las técnicas estadísticas más utilizadas por los investigadores de las más variadas disciplinas es el análisis de la varianza (ANOVA). Este procedimiento estadístico goza de gran robustez, pero exige que los datos cumplan tres requisitos fundamentales: normalidad en la distribución, homogeneidad de varianzas (homocedasticidad) e independencia de las observaciones. Los dos primeros requisitos suelen ser comprobados de manera regular, mientras que el requisito de independencia raramente es testado, confiando en muchos casos ciegamente en que la aleatorización sea suficiente para su cumplimiento.

Sin embargo, y como han demostrado gran número de autores (ver ZAS, 2007 y referencias citadas allí), la independencia de las observaciones puede verse condicionada cuando los datos

ISSN: 1575-2410

presentan una estructura espacial no aleatoria a mayor o menor escala, es decir, cuando existe heterogeneidad espacial. Esto es algo muy común en cualquier ambiente o ecosistema, en el que la distribución de los condicionantes tanto físicos como biológicos no es aleatoria ni uniforme (LEGENDRE, 1993). Esta situación cobra especial relevancia cuando se trata de ensayos genéticos forestales, debido a que habitualmente se instalan en terrenos abruptos y con una fuerte heterogeneidad. La presencia de manchas y gradientes es, de hecho, la norma general en este tipo de ensayos (e.g. DUTKOWSKI et *al*., 2006; FU et *al*., 1999; HAMANN et *al*., 2002; ZAS, 2006, 2007; ZAS et *al*., 2007).

Es por ello que surge la necesidad de instalar los ensayos en base a diseños experimentales cuyo objetivo es absorber la heterogeneidad espacial presente. Los más comunes son el de bloques completos al azar (BC), con todos los niveles del factor en cada una de las réplicas del ensayo, y el diseño de bloques incompletos (BI) en el que no aparecen todos los niveles del factor en cada bloque.

El requisito fundamental que deben cumplir los diseños en bloques, es que éstos deben ser homogéneos, lo que en muchas ocasiones no se cumple, resultando ineficientes e incapaces de absorber la variación espacial presente. La heterogeneidad interna de los bloques se debe en primer lugar a su gran tamaño en ensayos genéticos forestales, a que el replanteo del diseño experimental en campo se realiza de manera previa al conocimiento del patrón espacial verdadero, y a que los bloques suponen fronteras artificiales bruscas, mientras que la variación ambiental tiene carácter continuo y suave (LEGENDRE, 1993; DUTILLEUL, 1993).

Se hace entonces necesario prestar especial atención a la posible autocorrelación espacial de los datos cuando se analizan variables de experimentos en campo, ya que su presencia puede llegar a ser dramática e invalidar los resultados del análisis (DUTILLEUL, 1993). Se ha demostrado que aplicar el análisis de varianza cuando existe heterogeneidad espacial sin realizar ningún tipo de ajuste deriva en un cálculo erróneo del nivel de significación de los efectos del modelo, de la proporción de varianza explicada por cada factor, y/o de la estimación de los efectos del modelo (ZAS, 2007).

En las últimas décadas han aparecido diversas técnicas basadas en estadística espacial que tratan de estudiar la variabilidad ambiental y su patrón espacial. A causa de la falta de eficacia de los diseños experimentales, varios autores han procurado aplicar estas técnicas en la evaluación de ensayos genéticos en campo, tanto de tipo agronómico como forestal (e.g. QIAO et *al*., 2000; HONG et *al*., 2005; DUTKOWSKI et *al*., 2006; ZAS, 2006).

El objetivo de este trabajo es comprobar la eficiencia estadística de tres diseños experimentales convencionales, comparándolos con dos de estos métodos alternativos basados en estadística espacial.

MATERIAL Y MÉTODOS

Los datos utilizados corresponden a tres ensayos de progenie de *Pinus pinaster* Ait., instalados en la primavera de 2005 en Galicia. El material vegetal instalado en dichos ensayos se trata de familias de medios hermanos procedentes de 116 árboles superiores seleccionados en masas naturales de la zona costera de Galicia y clonados en el huerto semillero de primera generación de Sergude (A Coruña). Se incluyen además en el ensayo a modo de testigos o controles, tres lotes de semilla comercial no mejorada para repoblación, uno empleado en la zona costera de Galicia, otro en la zona de interior, y el tercero en Francia.

Los ensayos siguen un diseño de bloques incompletos (BI) resoluble, donde los BI pueden ser agrupados en bloques completos (BC) de tal manera que cada tratamiento (familia) aparece representado una vez en cada réplica. Dichos ensayos cuentan con 96 bloques (BI) con forma rectangular de 10 familias cada uno y 3 plantas contiguas de la misma familia por unidad experimental. Cada réplica o BC está formado por 12 BI. El número de plantas en cada ensayo asciende a 2856 que, a marco de plantación de 3x2 m, ocupan una superficie cercana a las 2 ha. El tamaño promedio del BC es de 2160 m^2, con formas que van desde las cuadradas a las rectangulares alargadas, los BI tienen forma cuadrada,

ocupan una superficie de 180 m2, y la longitud aproximada de su lado es de 12 metros.

La variable objeto de estudio fue la altura total dos periodos vegetativos después de la plantación. La posición relativa de cada planta fue determinada mediante levantamiento topográfico empleando una estación total.

La variable se analizó mediante tres modelos convencionales, que respectivamente consideran las parcelas de ensayo bajo tres diseños experimentales diferentes. Para el diseño completamente aleatorizado (CA), en el que no se considera los efectos de los bloques, y el de bloques completos al azar (BCA), sus expresiones matemáticas son:

$Y_{ijk} = \mu + G_i + \varepsilon_{ijk}$ $Y_{ijk} = \mu + G_i + BC_j + \varepsilon_{ijk}$

donde Y_{ijk} es el promedio de la variable (media de las tres plantas) para la unidad experimental "k" de la familia "i" en la réplica "j", μ es la media global, G_i es el efecto fijo de la familia "i" (i = 1-119), BC_j es el efecto aleatorio del bloque completo "j" (j = 1-8) y ε_{ijk} es el error experimental.

Para el diseño de bloques incompletos (BI), el modelo es:

$$Y_{ijkl} = \mu + G_i + BC_j + BI_k\,(BC_j) + \varepsilon_{ijk}$$

donde, además de los términos de la primera ecuación se incluye el efecto fijo del bloque completo "j" (BC_j, j = 1-8), y el efecto aleatorio del bloque incompleto "k" (k = 1-96) perteneciente a la réplica "j" ($BI_k(BC_i)$).

Los tres modelos se han analizado mediante el procedimiento MIXED en SAS (SAS INSTITUTE, 1999), explorándose la independencia de los residuos mediante la construcción de semivariogramas con el procedimiento VARIOGRAM y ajustando el semivariograma teórico mediante regresión no lineal con el procedimiento NLIN en SAS (SAS INSTITUTE, 1999).

Frente a los modelos convencionales, se han empleado dos metodologías basadas en técnicas de estadística espacial para el análisis de las plotmeans. Por un lado se ha empelado un modelo de errores espacialmente correlacionados (EC), que incorpora la estructura espacial no aleatoria del error mediante el procedimiento MIXED de SAS (SAS INSTITUTE, 1999). La estructura espacial del error se define mediante el comando REPEATED que nos permite especificar distintas estructuras de autocorrelación (LITTELL et *al*., 1996; HONG et *al*., 2005). El modelo EC se basa en un diseño completamente aleatorizado (sin considerar la estructura de bloques) estimándose los parámetros del modelo espacial de la distribución del error mediante el propio proceso REML del procedimiento MIXED. El otro procedimiento de análisis espacial se trata de un proceso de ajuste denominado ISA (*Iterative Spatial Analysis*) (ZAS, 2006), basado en corregir la variable original mediante semivariogramas y kriging para eliminar la correlación espacial de una manera iterativa. Una vez eliminada la parte de la variación residual que depende de la heterogeneidad espacial, y obtenida una nueva variable, ésta es analizada de manera similar al diseño CA con el procedimiento MIXED en SAS.

Los distintos modelos se han comparado en términos del error estándar promedio de la estimación de las medias ajustadas (LSMEANS) para el efecto fijo de la familia, indicativo de la precisión de las mismas, mediante el estadístico de máxima verosimilitud *-2L*, que cuantifica la bondad del ajuste del modelo, siendo ésta mayor cuanto menor sea el valor de *-2L* (LITTELL et *al*., 1996), y mediante el F ratio de los efectos fijos.

RESULTADOS

Los semivariogramas de los residuos de la variable, eliminado el efecto familiar, revelan la presencia de una estructura espacial no aleatoria en los 3 sitios de ensayo, reflejando que los valores de los vecinos más próximos son más parecidos que los de aquellos más alejados (Figura 1). En cada uno de los sitios se muestran las diferentes tipologías de patrones existentes: en gradiente, en manchas de tamaño grande y en manchas de pequeño tamaño. Para las distribuciones en manchas, el patrón muestra además una muy alta intensidad con alrededor del 60% de la variación residual explicada por la estructura espacial. El tamaño de mancha (rango) fue de 75 y 33 metros en cada caso (Tabla 1).

Si se comparan los semivariogramas de la variable original, y los de los residuos de los tres modelos convencionales en cualquiera de los tres ensayos (Figura 1), podremos entender mejor el efecto de los distintos diseños en el análisis estadístico. Por un lado, el semivario-

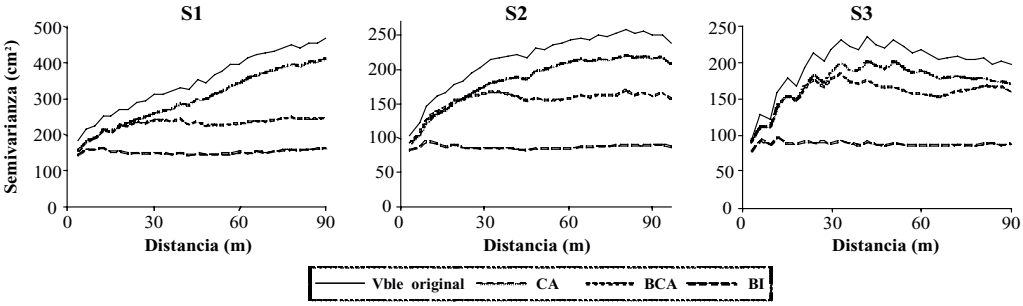

Figura 1. *Semivariogramas empíricos para la variable original, los residuos del diseño completamente aleatorizado (CA) y los residuos de los diseños de bloques completos (BCA) e incompletos (BI)*

SITIO	Tipo de patrón	Completamente aleatorizado			Bloques completos al azar			Bloques incompletos		
		IP[1]	R^2	rango (m)	IP[1]	R^2	rango (m)	IP[1]	R^2	rango(m)
S1	gradiente	-	0,99***	-	49%	0,89***	-	0%	0,00ns	-
S2	mancha grande	67%	0,99***	75	53%	0,94***	23	25%	0,28*	7
S3	mancha pequeña	59%	0,89***	33	55%	0,84***	25	60%	0,61***	5

Tabla 1. *Eficiencia de los tres diseños convencionales en los tres sitios de ensayo, definida por la intensidad de la variación espacial residual de los modelos (IP) y por la bondad de ajuste del semivariograma teórico (R^2). Se incluye el rango o tamaño de la mancha del patrón.* [1] *Intensidad del patrón estimada mediante el ratio C/(C+Co), siendo C la varianza estructural y Co el nugget*

grama de los residuos del modelo CA presenta valores menores que los de la variable original, pero ambas curvas son totalmente paralelas. El diseño de BCA genera un semivariograma residual casi coincidente con el del diseño CA a distancias cortas, pero que se aplana a largas distancias. Por último el semivariograma de los residuos del diseño en BI se muestra casi totalmente plano en los tres sitios de ensayo, salvo a muy cortas distancias lo que indica la persistencia de un ligero patrón a pequeña escala.

Los valores del estadístico *-2L* (Tabla 2) van disminuyendo desde el diseño convencional más simple, el CA, hasta los modelos de análisis espacial. De la misma manera, tanto el F ratio de los efectos fijos, como la precisión de la estimación de los efectos del modelo mejoran en el mismo orden, aumentando desde el modelo CA hasta alcanzar el máximo con el método de ajuste ISA, indicando una mayor robustez de este último.

En la tabla 1 se compara la eficiencia de los diseños convencionales para absorber la variación espacial en los tres sitios de ensayo. Para patrones en gradiente, los residuos del diseño de BCA todavía mantienen una autocorrelación espacial con intensidad cercana al 50%, mientras que en el caso del diseño de BI los residuos son independientes (semivariograma plano) resultando por lo tanto eficientes. En el caso de distribuciones espaciales en manchas, ni siquiera los diseños de BI son capaces de absorber la variación espacial, mostrando residuos con estructuras no aleatorias, si bien esto sólo cobra especial gravedad cuando la escala del patrón es menor, es decir, en manchas pequeñas.

DISCUSIÓN

Se comprueba una vez más, que en lo referente a ensayos genéticos forestales la heterogeneidad espacial es lo común y no la excepción (e.g. DUTKOWSKI et *al*., 2006; FU et *al*., 1999; HAMANN et *al*., 2002; ZAS, 2006), pudiendo ser modelizada dicha heterogeneidad con gran precisión mediante semivariogramas teóricos con coeficientes de regresión no lineal (R^2) muy altos en todos los casos y con intensidades del

MODELO	SITIO 1			SITIO 2			SITIO3		
	-2L	F	SE	-2L	F	SE	-2L	F	SE
CA	7600	0,9	7,73	6683	1,0	5,62	6911	1,0	5,26
BCA	7227	1,5	7,71	6512	1,3	5,56	6813	1,1	5,25
BI	7160	1,8	5,62	6331	1,6	4,26	6599	1,3	4,15
EC	6994	1,9	12,74	6209	1,6	4,93	6421	1,4	5,24
ISA	6677	2,5	4,40	5301	4,3	2,31	5590	3,1	2,34

Tabla 2*. Robustez del modelo de análisis definida por el estadístico de ajuste (Residual log likelihood, -2L), F ratio de los efectos fijos y promedio de los errores estándar (SE) de la estimación de los efectos familiares (LSMEANS) según el modelo empleado para el análisis en los tres sitios de ensayo.*
CA: completamente aleatorizado, BCA: bloques completos al azar, BI: bloques incompletos, EC: Errores correlacionados, ISA: Iterative spatial analysis

patrón elevadas, a pesar de tratarse de observaciones a edades tempranas. Esta variación espacial puede distribuirse según diversos tipos de patrón (gradiente o manchas de mayor o menor tamaño) dependiendo de las condiciones propias del ambiente de ensayo, condicionando de forma diferente la eficiencia de los diseños experimentales convencionales.

La semivarianza residual del diseño completamente aleatorizado (CA) decrece con respecto a la variable original, debido a la eliminación del efecto familiar. Sin embargo, las fluctuaciones en la curva del semivariograma se mantienen, lo que indica que son debidas a la propia heterogeneidad espacial del ensayo y no al efecto de la familia.

La varianza máxima alcanzada por los semivariogramas residuales de los diseños de BC es menor que en los diseños CA, debido a que los bloques son capaces de absorber entorno a un 50% de la variación espacial presente en el ensayo independientemente del tipo de patrón. Sin embargo estos diseños sólo absorben aquella parte de la variación que se produce a gran escala, pero no la que se produce a cortas distancias donde ambas curvas se solapan. Esto se debe a la presencia de heterogeneidad a nivel de micrositio que se plasma en una heterogeneidad dentro de los bloques completos, lo que se comprueba comparando los 2160 m^2 de superficie de éstos, frente a los 75 y 33 metros de rango en el caso de los patrones en manchas.

Los bloques incompletos son más eficientes a la hora de absorber la variación espacial existente, mostrando semivariogramas de los residuos prácticamente planos lo que indica una baja autocorrelación residual. El menor tamaño de los bloques (180 m^2 – 12 m de lado) favorece que éstos encajen mejor sobre el patrón de variabilidad espacial. Sin embargo, su eficiencia depende del tipo de patrón espacial presente. Mientras que para situaciones en gradiente son capaces de absorber la práctica totalidad de la heterogeneidad, cuando los parches son de pequeño tamaño, los BI sólo absorben un escaso 40% de la variación espacial, resultando ineficientes.

La distinta capacidad de los diseños para absorber la heterogeneidad espacial determina la robustez de sus estimaciones. Los dos modelos de análisis basados en técnicas de estadística espacial, muestran estadísticos de ajuste (*-2L*) claramente más bajos que cualquiera de los modelos convencionales y F ratios notablemente superiores, ratificando que son modelos más adecuados. Además, el promedio de los errores estándar (SE) asociados a las estimaciones de los LSMEANS de los efectos familiares son más bajos en estos casos, indicando una mayor precisión en tales estimaciones. Entre los modelos de análisis espacial evaluados, el ajuste espacial iterativo ISA ofrece mejores resultados que el modelo de errores espacialmente correlacionados (EC).

CONCLUSIONES

El diseño de bloques completos al azar fue en estos 3 casos insuficiente para absorber la estructura espacial de la variable.

El diseño de bloques incompletos se mostró como alternativa mejorando sustancialmente la eficiencia del análisis, pero ante patrones en manchas de pequeño tamaño (sitio 3), resultó

menos eficiente que las metodologías basadas en técnicas de estadística espacial.

Vista la intensidad de los patrones espaciales presentes en los ensayos genéticos forestales estudiados a tan tempranas edades, se recomienda analizar de manera rutinaria la dependencia espacial de los residuos del modelo estadístico utilizado mediante la representación del correspondiente semivariograma residual. Cuando el patrón encontrado tiene carácter no aleatorio será necesario optar por un método que tenga en cuenta dicho patrón o que ajuste espacialmente los datos, ya que de otra forma las estimaciones obtenidas con el modelo de ANOVA original arrastrarán errores considerables.

BIBLIOGRAFÍA

DUTILLEUL, P.; 1993. Spatial heterogeneity and the design of ecological field experiments. *Ecology* 74: 1646-1658.

DUTKOWSKI, G.W.; COSTA-E-SILVA, J.; GILMOUR, A.R.; WALLENDORF, H. & AGUIAR, A.; 2006. Spatial analysis enhances modelling of a wide variety of traits in forest genetic trials. *Can. J. For. Res.* 36: 1851-1870.

FU, Y.; YANCHUK, A.D. & NAMKOONG, G.; 1999. Spatial patterns of tree height variations in a series of Douglas-fir progeny trials: implications for genetic testing. *Can. J. For. Res.* 29: 714-723.

HAMANN, A.; NAMKOONG, G. & KOSHY, M.P.; 2002. Improving precision of breeding values by removing spatially autocorrelated variation in forestry field experiments. *Silvae Genet.* 51: 210-215.

HONG, N.; WHITE, J.G.; GUMPERTZ, M.L. & WEISZ, R.; 2005. Spatial analysis of precision agriculture treatments in randomized complete blocks: guidelines for covariance model selection. *Agron. J.* 97: 1082-1096.

LEGENDRE, P.; 1993. Spatial autocorrelation: a trouble or new paradigm? *Ecology* 74: 1659-1673.

LITTELL, R.C.; MILLIKEN, G.A.; STROUP, W.W. & WOLFINGER, R.D.; 1996. *SAS System for mixed models.* SAS Institute. Cary, North Carolina.

QIAO, C.G.; BASFORD, K.E.; DELACY, I.H. & COOPER, M.; 2000. Evaluation of experimental designs and spatial analyses in wheat breeding trials. *Theor. App. Gen.* 100: 9-16.

SAS INSTITUTE (1999). *SAS/STAT User´s guide, Version 8*. SAS Institute. Cary, North Carolina.

ZAS, R.; 2006. Iterative kriging for removing spatial autocorrelation in analysis of forest genetic trials. *Tree Genet. Genomics* 2: 177-186.

ZAS, R.; 2007 Autocorrelación espacial y el diseño y análisis de experimentos. *En*: F. Maestre, A. Escudero y A. Bonet (eds.), *Introducción al análisis espacial de datos en ecología y ciencias ambientales, Métodos y Aplicaciones*: en prensa. Universidad Rey Juan Carlos, Asociación Española de Ecología Terrestre y Caja de Ahorros del Mediterráneo. Madrid.

ZAS, R.; MARTINS, P.; DE LA MATA, R.; 2008. Autocorrelación espacial: un problema comúnmente olvidado. *Cuad. Soc. Esp. Cienc. For.* 24: 139-145.