Annals of Forest Science (2019) 76: 53
https://doi.org/10.1007/s13595-019-0836-9

RESEARCH PAPER

Performance of alternative spatial models in empirical Douglas-fir and simulated datasets

Eduardo Pablo Cappa[1,2] · Facundo Muñoz[3] · Leopoldo Sanchez[3]

Received: 7 September 2018 / Accepted: 15 April 2019 / Published online: 13 May 2019

Abstract

• ***Key message*** **Based on an empirical dataset originating from the French Douglas-fir breeding program, we showed that the bidimensional autoregressive and the two-dimensional P-spline regression spatial models clearly outperformed the classical block model, in terms of both goodness of fit and predicting ability. In contrast, the differences between both spatial models were relatively small. In general, results from simulated data were well in agreement with those from empirical data.**

• ***Context*** Environmental (and/or non-environmental) global and local spatial trends can lead to biases in the estimation of genetic parameters and the prediction of individual additive genetic effects.

• ***Aims*** The goal of the present research is to compare the performances of the classical a priori block design (block) and two different a posteriori spatial models: a bidimensional first-order autoregressive process (AR) and a bidimensional P-spline regression (splines).

• ***Methods*** Data from eight trials of the French Douglas-fir breeding program were analyzed using the block, AR, and splines models, and data from 8640 simulated datasets corresponding to 180 different scenarios were also analyzed using the two a posteriori spatial models. For each real and simulated dataset, we compared the fitted models using several performance metrics.

• ***Results*** There is a substantial gain in accuracy and precision in switching from classical a priori blocks design to any of the two alternative a posteriori spatial methodologies. However, the differences between AR and splines were relatively small. Simulations, covering a larger though oversimplified hypothetical setting, seemed to support previous empirical findings. Both spatial approaches yielded unbiased estimations of the variance components when they match with the respective simulation data.

• ***Conclusion*** In practice, both spatial models (i.e., AR and splines) suitably capture spatial variation. It is usually safe to use any of them. The final choice could be driven solely by operational reasons.

Keywords Global and local spatial trends · Forest genetics trials · Autoregressive residual · Two-dimensional P-splines

Handling Editor: Ricardo Alia

Contributions of the co-authors EPC, FM, and LS conceived and designed the research; FM, LS, and EPC planned and carried out the simulations; EPC and FM analyzed the data; EPC, FM, and LS wrote the original paper; LS supervised the work and coordinated the research project.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13595-019-0836-9) contains supplementary material, which is available to authorized users.

✉ Eduardo Pablo Cappa
cappa.eduardo@inta.gob.ar

[1] Bosques Cultivados, Centro de Investigación en Recursos Naturales, Instituto Nacional de Tecnología Agropecuaria (INTA), Instituto de Recursos Biológicos, De Los Reseros y Dr. Nicolás Repetto s/n, 1686, Hurlingham, Buenos Aires, Argentina

[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina

[3] UMR BioForA, INRA, 45160 Ardon, France

1 Introduction

Global and local spatial trends are well known empirically in forestry field trials as a result of environmental factors such as variations in soil characteristics and land topography. Tree breeders have attempted to account for this environmental variability using a priori designs (i.e., by design), including randomized complete and incomplete blocks and lattices. However, in most cases, even under the most efficient experimental layout, the spatial heterogeneity is conveniently revealed at the evaluation stage (Fu et al. 1999). Thus, it is often necessary to model such variability a posteriori within the model of evaluation (i.e., by analysis). Additionally, the spatial heterogeneity can also include other non-environmental factors such as competition, non-random arrangements of genotypes, or any other unexplained variation (Gilmour et al. 1997), which could further affect the performance of evaluation methods.

Although environmental heterogeneity has been often considered a nuisance in forest genetic evaluation, where the main goal is the prediction of breeding values, ignoring such a source of heterogeneity can lead to biases in the estimation of genetic parameters and the prediction of individual additive genetic effects otherwise known as breeding values (Magnussen 1993, 1994). Several a posteriori approaches called "spatial models" have been developed and widely applied to forest genetic trials in order to account accurately for site heterogeneity. The impact of small-scale spatial heterogeneity is accounted for by including a spatially correlated structure into the model residuals expressed as a Kronecker product of first-order autoregressive processes for rows and columns (Gilmour et al. 1997). Other alternatives to model the small-scale spatial variability use nearest neighbor techniques (Anekonda and Libby 1996; Joyce et al. 2002; Kroon et al. 2008; Gezan et al. 2010) or kriging (Hamann et al. 2002; Zas 2006). Large-scale continuous spatial variation has been modeled following a variety of approaches, like post-blocking (Ericsson 1997; Lopez et al. 2002; Gezan et al. 2006), the inclusion of spatial coordinates expressed as either classification variables such as polynomials (Thomson and El-Kassaby 1988; Federer 1998; Saenz-Romero et al. 2001), or smoothing splines (Gilmour et al. 1997; Verbyla et al. 1999). Of these methods, a Kronecker product of first-order autoregressive residual (co)variance structure has become commonly used in forest tree breeding. In one dimension (either in rows or in columns), the resulting first-order autoregressive structure is equivalent to a geostatistical model with an exponential covariance function. Many forest genetic studies using this spatial model approach displayed a consistent reduction in the error variance and increases in both heritabilities and accuracies of predicted breeding values with respect to the a priori model with block design (e.g., Costa e Silva et al. 2001, Dutkowski et al. 2002, 2006, Ye and Jayawickrama 2008). In a first-generation Douglas-fir progeny trial, Ye and Jayawickrama (2008) showed that the spatial autoregressive model removed on average 14 to 34% of residual variance due to spatial heterogeneity, which resulted in 20% increase in accuracy of breeding value prediction with respect to the classical non-spatial model.

An alternative approach to model complex patterns of environmental heterogeneity proposed by Cappa and Cantet (2007) is to use a mixed model representation of a two-dimensional P-spline regression (Eilers and Marx, 2003) with spatially structured coefficients. Over a series of studies with forest genetic trials involving large scale (Cappa and Cantet 2007), small scale (< 1 ha) (Cappa et al. 2011), and both large and small scales (Cappa et al. 2015a), the P-splines approach also displayed a consistent reduction in the residual variance with respect to the blocks' model, together with increases in heritability and accuracy of the predicted breeding values of parents and offspring.

Many forest genetic studies have already compared a priori design models with one of the several a posteriori spatial approaches available. Gezan et al. (2010) compared the results from the a priori incomplete block design with those obtained with a model with an autoregressive residual structure, by using simulated data of unrelated genotypes with different surface patterns. The comparison indicated that the incorporation of this autoregressive structure yielded the highest correlations between the predicted and true treatment effects. Few studies, however, have systematically compared several a posteriori methods over a range of trials and parametric scenarios. Saenz-Romero et al. (2001) combined quadratic polynomials and autoregressive approaches into a mixed model to fit simultaneously global and local trends in a nursery trial. The study showed that the combination of the two methods was one of the best analytical choices. Rodríguez-Álvarez et al. (2018) compared the autoregressive approach with a more sophisticated spatial model based on splines in an agricultural context.

The goal of the present research is to compare the performances of alternative approaches to account for various patterns of spatial autocorrelation on forest

genetic data. Competing methods are the classical a priori block design and two different spatial models: a bidimensional separable first-order autoregressive spatial process and a bidimensional P-spline regression (hereafter, "block," "AR," and "splines" models, respectively). These are the three most widely used models to accommodate the spatial heterogeneity in forest genetic trials. The block model is the classical approach to handle environmental variation and dates back to the development of the theory of experimental design almost a century ago. The AR model is a more modern and flexible approach that estimates a nearly continuous surface of environmental effects. Finally, the splines model has been gaining more recognition recently. Besides their popularity, they represent three fundamentally different approaches to environmental variation. The block model is based on the aggregation of individuals into groups that share similar environments and assume the effect of the environment being constant in each group. The AR model estimates the correlation between neighboring residual effects. Finally, the splines model belongs to the family of kernel smoothers, representing the spatial effect as a linear combination of a set of basis functions. In addition, the present study showcases different modalities for diagnosing competing models, as an additional but important step in the analysis of spatial data. One of the most common tools is the analysis of the residual semivariograms, which helps to quantify the unaccounted spatial autocorrelation from a given model. Another approach involves cross-validation, a measure of model's adequacy independent of any model assumptions (Grondona et al. 1996).

For generality, we based our study on simulated and real data. Simulated data offer the advantage of covering large parametric ranges that are difficult to find in real datasets, and allow for a more accurate assessment of model performance. Our simulated datasets involved contrasting spatial patterns generated by both AR and splines models under different scenarios. Each dataset was then fitted with AR and splines models and their respective performances assessed and compared under several criteria. Among the alternative scenarios, the cases of single- and multiple-tree plot designs, and half-full-sibs, and clonal genetic structures were studied. We also used data from the French Douglas-fir tree improvement program to complement simulated scenarios and confront the alternative methods to real datasets. These datasets involved a number of mature progeny trials that were established across a range of contrasting sites but yet appropriate for productivity in the context of the species in France. Precise genetic evaluation at this stage is of critical importance before selection. For this reason, previous analyses based on a priori design were re-analyzed with these alternative a posteriori spatial models (i.e., AR and splines) and the new results compared with the previous ones. We discuss the implications of our findings for the French Douglas-fir breeding program led by the French National Institute for Agricultural Research (INRA), and wider implications for diagnosis of spatial modeling and genetic evaluation in general.

2 Materials and methods

2.1 Genetic material, mating design, and trial description

Data from eight trials of the French Douglas-fir breeding program of an appropriated age for genetic evaluation were used in the current study (Bastien et al. 2013). These trials correspond to progeny tests involving either breeding stock for the renewal of the breeding population (2.707.1, 2.707.3, 2.708.1, and 2.708.3) or progeny-tested genotypes from first-generation seed orchards going into genetic thinning (3.704.2, 3.713.1, 3.713.1, and 3.713.1). Trials were roughly the same age, 18 to 21 years after planting, and were chosen to be representative of the French trial network for Douglas-fir breeding program, accounting for geography, climate, and basic trial disposition. The features of each of the trials are given in Table 1. The standard design in all eight trials corresponded to a randomized complete block setting following a single-tree plot design, with as many blocks as replicates per family. The number of replicates per trial can be found in Table 1.

Two traits, height and diameter at breast height, were measured at 6–8 and 16–20 years, respectively (Table 1). However, not all traits were available for every trial.

2.2 Statistical models of analysis and inference

We compared three individual-tree models of general form:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Bb} + \boldsymbol{Za} + \boldsymbol{e}$$

with alternative formulations of the spatial random effect $\boldsymbol{Bb}$: block, AR, and splines. These three models

Table 1 Design and phenotypic information across the eight Douglas-fir progeny trials

Site		2.707.1	2.707.3	2.708.1	2.708.3	3.704.2	2.713.1	2.713.2	2.713.3
No. of trees with records		5434	6248	6709	8977	2759	3976	3458	1459
Parents number		183	198	221	235	68	134	117	66
Family number	OP[a]	183	198	219	228	68	95	78	29
	FS[b]	–	–	6	8	–	38	30	31
No. of provenances		7	9	6	7	–	6	8	8
Replicate (blocks)		88	88	100	101	91	53	40	34
Row		196	120	234	164	55	63	90	79
Column		140	153	83	109	74	172	97	51
Spacing (m)		3 × 3	3.6 × 2.75	3 × 3	4 × 2.5	3.5 × 3	3 × 3	3.5 × 3	2.5 × 3.5
Age at height assessment		8	7	7	8	8	7	–	6
Age at diameter at breast height assessment		19	–	–	–	20	16	16	16
Height mean (mm) (and SD[c])		413.0 (123.4)	405.8 (89.8)	416.4 (105.9)	572.4 (97.1)	443.1 (99.4)	495.9 (109.3)	–	345.7 (96.5)
Diameter at breast height (cm) (and SD)		415.7 (155.8)	–	–	–	636.3 (161.9)	486.5 (130.0)	530.2 (106.0)	449.9 (153.9)
Altitude (m a.s.l.)		330	700	425	720	550	500	710	650
Slope (%)		5–10	10	10	10	8	0	5–30	2–20
Mean annual precipitation (mm)		1100	1440	900	1140	–	1200	1300	1100

[a] OP: open-pollinated families, [b] FS: full-sib families, [c] SD: standard deviation

were evaluated for each combination of site and trait in the Douglas-fir dataset. Although genetic evaluation is generally conducted with multiple-trait and multiple-site models, splitting data analyses in this way was justified by the need of collecting a broad spectrum of spatial patterns. The vector $\boldsymbol{y}$ contains the phenotypic data, and all three models included a fixed effect of provenance (β, genetic group) to account for different sub-population means. A set of genetic ($\boldsymbol{a}$) and residual ($\boldsymbol{e}$) random effects were also considered in all models. The former was a normally distributed random additive genetic effect (breeding values) with (co)variance matrix $\boldsymbol{A}\sigma_a^2$, where $\boldsymbol{A}$ is the additive relationship matrix among all trees (Henderson 1984), and σ_a^2 is the additive genetic variance. The residuals were independently distributed with mean 0 and residual variance σ_e^2. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the incidence matrices relating the observations ($\boldsymbol{y}$) to the model effects β and $\boldsymbol{a}$.

The spatial random effect was modeled differently in the three alternative models. For the block model, $\boldsymbol{B}$ was the incidence matrix for blocks and $\boldsymbol{b}$ was a normally distributed random effect with mean 0 and variance σ_s^2. The AR model considers $\boldsymbol{b}$ a random effect at individual level with a covariance structure σ_s^2 $[AR1(\rho_{row})\otimes AR1(\rho_{col})]$ given by the Kronecker product of first-order autoregressive processes $AR1(\rho_.)$ in the rows (*row*) and the columns (*col*) with spatial variance parameter σ_s^2, while $\boldsymbol{B}$ is a permutation matrix to sort observations by columns within rows (Gilmour et al. 1997). Finally, the splines model fits a two-dimensional surface built as a tensor product of cubic B-splines basis (Eilers and Marx 2003). The matrix $\boldsymbol{B}$ contains the two-dimensional B-splines basis evaluated in the corresponding row and column for each tree, while the vector of regression coefficients $\boldsymbol{b}$ is normally distributed with mean 0 and covariance matrix $\boldsymbol{U}\sigma_s^2$, where $\boldsymbol{U}$ is a fixed spatial structure and σ_s^2 is the spatial variance parameter. A more detailed explanation of the two-dimensional surface and the covariance structure used in this work can be found in Cappa and Cantet (2007). Note that the three formulations (block, AR, and splines) are mutually exclusive. In particular, the AR and splines models do not include the effect of the blocks.

In order to make the alternative parameters of spatial variance comparable, we scaled the covariance matrices so that σ_s^2 represented exactly the characteristic marginal variance of the spatial effect (Sørbye and Rue 2014).

Residual maximum likelihood (REML, Patterson and Thompson 1971) was used to estimate the variance components for the random effects in the three spatial models described previously, by using the function remlf90 in the R-package breedR (Muñoz and Sanchez, 2015), which is based on the programs REMLF90 and AIREMLF90 of the BLUPF90 library (Misztal 1999). While REML90 uses the "expectation maximization" algorithm for the variance component estimation, AIREMLF90 is based on the "average information" approach. The former is slower but more robust to initial values and was used for the more complex splines models, while the latter was used for the AR models.

The autocorrelation parameters in the AR models and the number of row and column knots in the splines models, respectively, were estimated by a grid-search approach that fits the model using several combinations of these values and selects the one with highest log-likelihood. Specifically, for the empirical Douglas-fir data, we used a two-step procedure for AR, initially fitting the models on a coarse grid with all combinations of the values (−0.8, −0.2, 0.2, 0.8) for the autoregressive parameters in each direction. A refined grid of the same size around the previous best combination was used in the second step, considering a variation of roughly 20% at each side in each dimension in the logit scale. For the spline models, the initial number of knots in each dimension is a power function of the size of the field given as a default by breedR. We tested all the combinations of five values around this initial value in each dimension, similarly to AR. As a result, search process for the empirical Douglas-fir data assumed anisotropy for the AR and splines models; i.e., different autoregressive parameters and number of knots were fitted in row and column directions.

Standard errors for estimated variances and functions of variances (i.e., heritabilities, see equation below) were calculated via Monte Carlo (Manly 1991). This implied sampling random realizations of datasets from the corresponding analytical models and estimated variance components.

2.3 Model comparison and diagnostics for the empirical Douglas-fir datasets

We compared the predictive ability of the fitted models with the Akaike information criterion (AIC, Akaike 1974), which is an approximation of the average out-of-sample deviance based on the marginal likelihood. A smaller AIC value indicates a better trade-off between goodness of fit and parsimony.

In addition, a 12-fold cross-validation was conducted. Each dataset was subdivided into 12 equally sized independent samples. Each of the samples was used for validation after fitting the model. The predicting ability was obtained as the Pearson correlation between observed phenotypes and fitted values. A complementary measure of the prediction quality was obtained as the root mean square error (RMSE) of the fitted values.

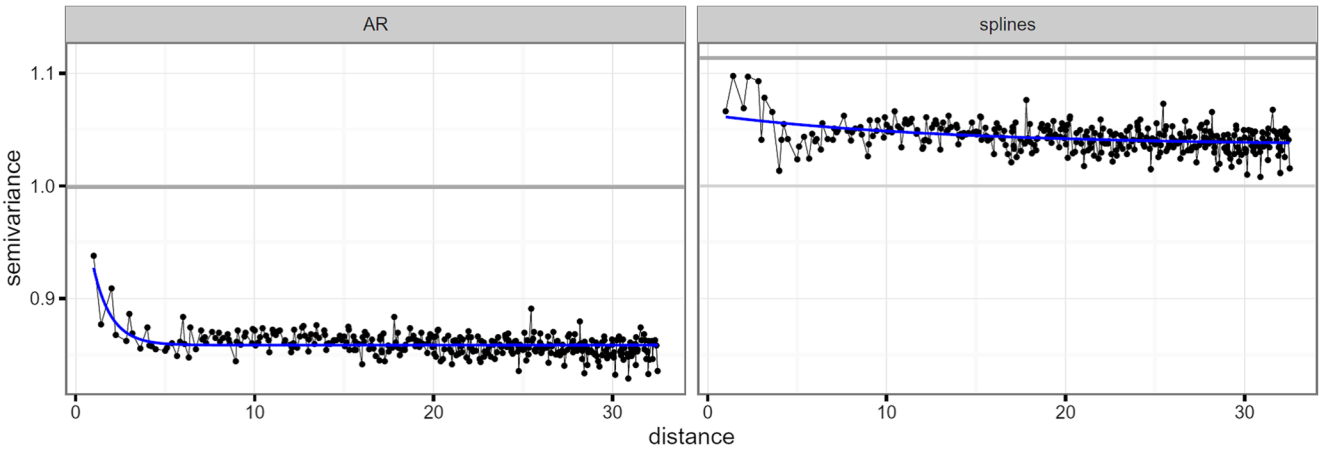


Fig. 1 Empirical semivariograms of residuals from the bidimensional separable first-order autoregressive (AR) and bidimensional spline regression (splines) models fitted to a simulated dataset with a high-variance short-ranged AR spatial effect under a single-tree plot structure. The horizontal (darker) lines are the estimated values of the residual variances for a true value of 1

Single-site narrow-sense individual heritability $\left(\hat{h}^2\right)$ was estimated as $\hat{h}^2 = \hat{\sigma}_a^2/\left(\hat{\sigma}_a^2 + \hat{\sigma}_e^2\right)$, where $\hat{\sigma}_a^2$ is the estimated additive genetic variance, and $\hat{\sigma}_e^2$ is the estimated residual variance. We explicitly excluded the spatial variance from the denominator in order to get comparable estimates of heritability across sites. This simple definition of heritability does not adhere to the standard quantitative genetics assumptions (Cullis et al. 2006), and so this measure should only be interpreted as a descriptive measure of precision (or ability) to detect additive genetic differences among the three models studied.

Further model comparison was provided by the predictive accuracy of breeding values, computed as:

$$r = \sqrt{\frac{1-SE^2}{\hat{\sigma}_a^2}}$$

where *SE* stands for the standard error of predicted breeding values using the "best linear unbiased predictors" (BLUPs) of parents and offspring. Finally, Spearman correlations between the best 10% candidate breeding values obtained by any two alternative models were also calculated, in order to check whether the ranking of breeding values differed between alternatives and therefore selection decisions were impacted by the choice of the model.

We also computed the empirical semivariogram of residuals as a standard model diagnostics tool. The isotropic semivariogram represents the half-average variation in pairwise residual differences as a function of the distance (Cressie 1993). Since the synthetic residuals were simulated independently and identically distributed, a logical expectation is that the empirical semivariogram of the residuals from a model properly accounting for all relevant effects will be approximately constant at the estimated residual variance. Thus, it might be disturbing to obtain empirical residual semivariograms such as those shown in Fig. 1 after fitting an AR or splines model to an ideal, simulated dataset.

We have seen this pattern emerging often from models fitted to real datasets as well (e.g., progeny trial 3.704.2; see Fig. 2b), most noticeably with AR spatial models. Specifically, the two challenging aspects are:

1. a discrepancy between the variance of the empirical residuals and the estimated residual variance, and
2. a remarkable non-flat slope in the first few lags.

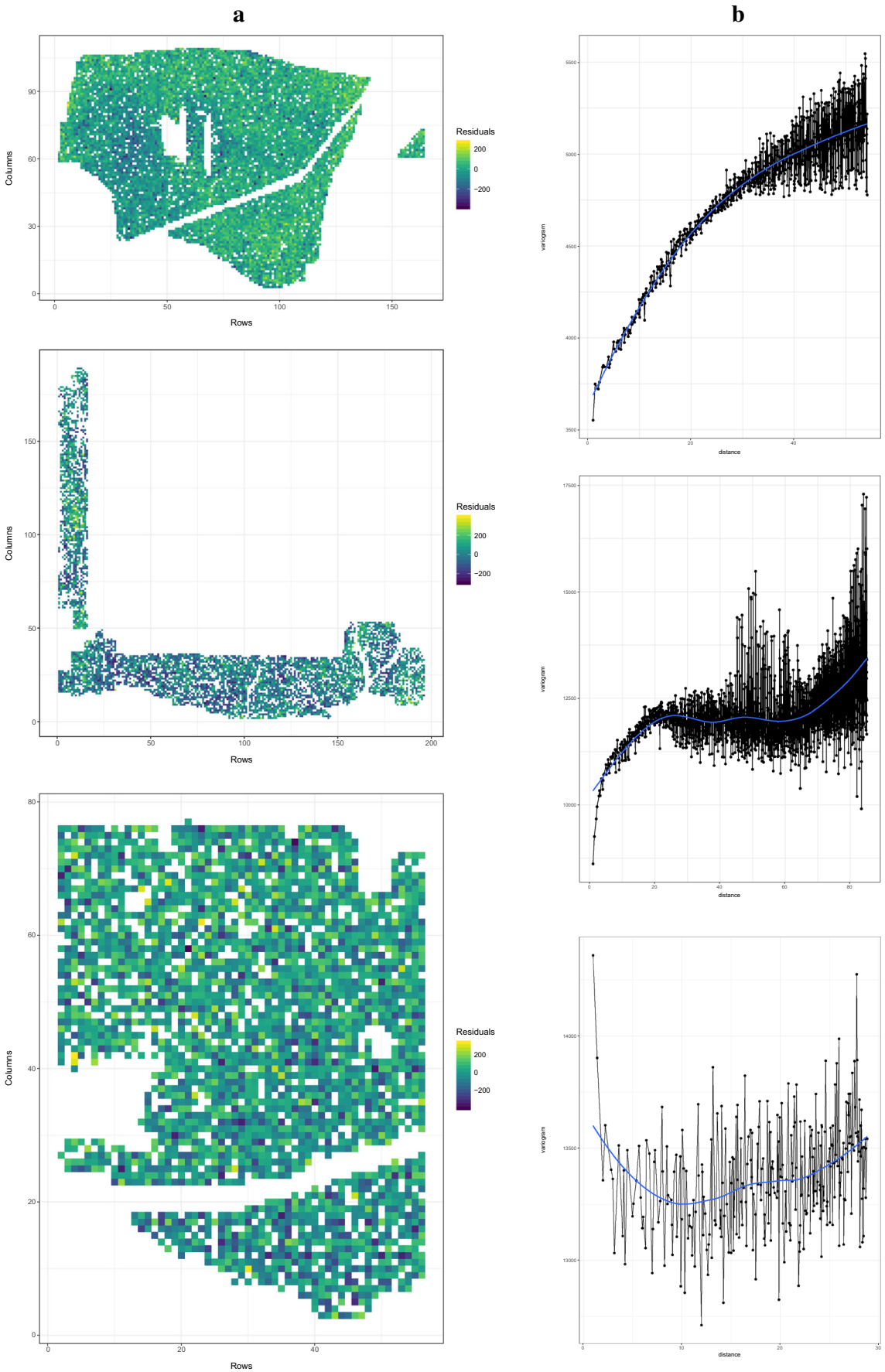
Such behavior could lead to thinking that there remains some unaccounted structure in the residuals, or that the spatial effect is overfitting the observations. To assess this, we devised two metrics for quantifying the extent of these effects in the simulation experiment (see Section 2.5). For each fitted model, we measured the discrepancy between the empirical and estimated residual variances (i.e., empirical vs. estimated differences; RV_disc) and the slope of the empirical semivariogram of residuals at distance 0 (vgslope).

2.4 Selection of representative trials according to spatial patterns

Twelve site-trait combinations from height difference progeny trials were analyzed in order to assess the patterns of environmental heterogeneity. We examined the spatial distribution of residuals (heat map) and their isotropic empirical semivariograms from a model with fixed genetic groups whenever present and individual-tree level additive genetic effects (breeding values).

Although all cases were used for comparing the performance of the different spatial models, only three of them were used as representative cases of the different spatial pattern scenarios (Fig. 2) for further illustration in Section 3. The first selected trial (2.708.3, height) is a typical case exhibiting large-scale variation in two dimensions, with a heat map displaying large patches of similar residuals and a semivariogram with a clear trend of increasing variation with the distance between trees. A trial with mostly short-scale variation is shown in the next panel (2.707.1, height), where the semivariogram depicts a rapid change in variation at close distances with a leveling out at medium distances. The heat map displayed smaller aggregates of similar tones than in the previous case. Finally, the last selected trial (3.704.2, diameter) displays a very short-scale spatial pattern, with the heat map showing antagonistic residuals in neighboring cells, and a semivariogram with no general trend but at the very short range.

Fig. 2 Heat maps (**a**) and isotropic empirical semivariograms (**b**) from the residuals of a basic genetic model not accounting for spatial autocorrelation, for three contrasting cases of the eight Douglas-fir progeny trials ▶


a
b
Columns
Rows
Residuals
200
0
−200
variogram
distance

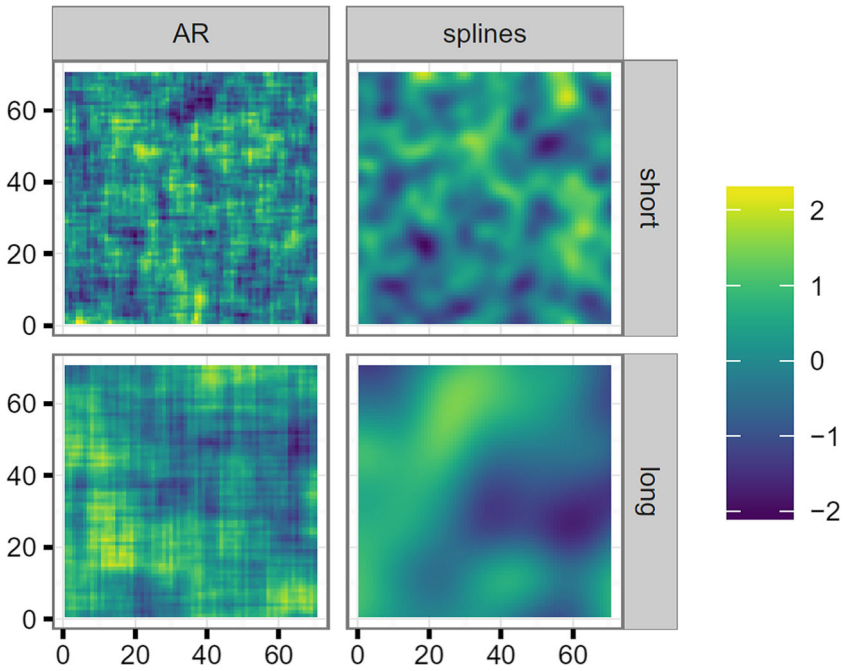


Fig. 3 Random realizations of examples of simulated spatial patterns with the bidimensional separable first-order autoregressive (AR) and bidimensional spline regression (splines) models at short and long spatial ranges

2.5 Performance assessment of AR and splines approaches under simulated scenarios

We designed an extensive simulation experiment where a number of datasets were sampled under several alternative scenarios. We fitted AR and splines models to each dataset and compared their relative performance according to a few selected measures.

Specifically, we simulated 180 datasets of dimension 70 × 70 individuals in each of the 12 scenarios which arise from the combination of the following:

- generating model: AR and splines
- range: short and long
- variance ratio of spatial and residual variances: low, medium, and high

Given that the environmental spatial variation may interact with other non-environmental factors like additive genetic resemblance between neighbors or plot structure and this interaction can affect differently the performance of alternative spatial modeling, we simulated four additional scenarios involving

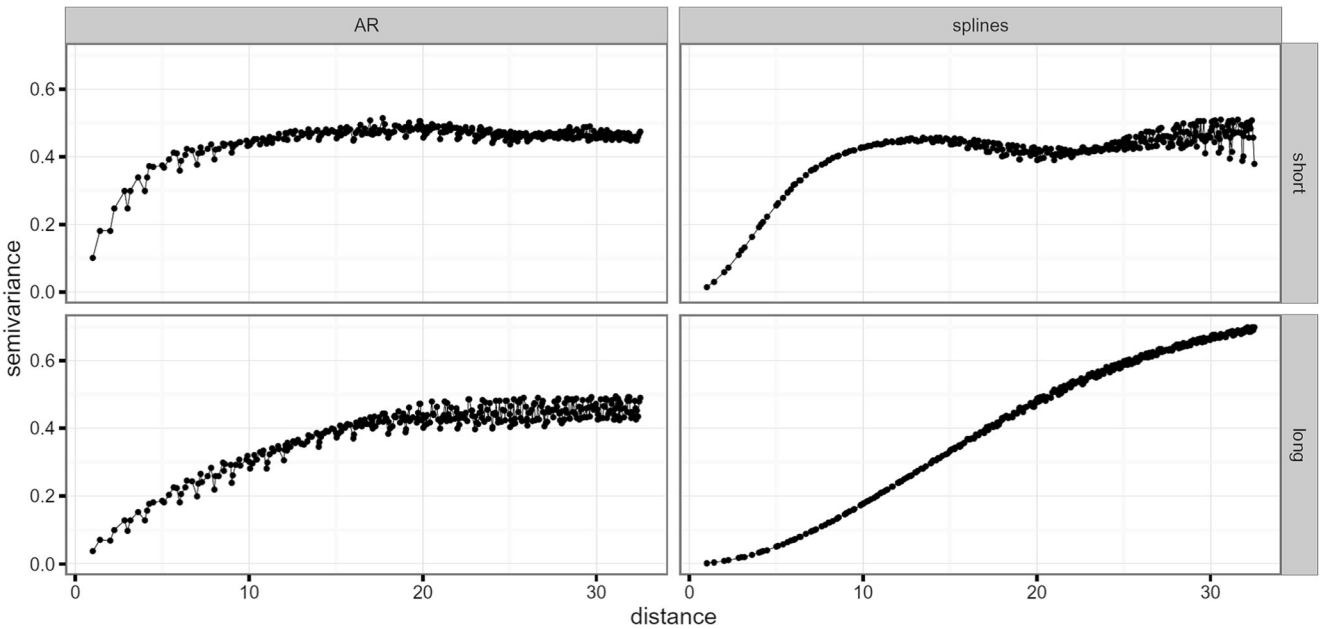


Fig. 4 Empirical semivariograms of realized spatial effects by model (bidimensional separable first-order autoregressive-AR-and bidimensional spline regression -splines-) and spatial range (short and long)

combinations of plot design and within-plot genetic relatedness. The basic scenario was a single-tree plot design (STP) with a Gaussian residual variance of 1. No genetic effects were simulated in this scenario, since their effect on the estimation of the spatial component would be indistinguishable from the residuals due to the stochastic independence. Three other alternative scenarios with multiple-tree plot designs (MTP) of seven trees in line plots differed in their within-plot genetic composition, with clonal, full-sib, and half-sib family structures with independent progenitors. The STP designs were completely randomized, while the MTP designs were arranged into 10 bands of randomized line plots with seven individuals each (see Supplementary Fig. S1). For the MTP designs, we assumed an additive genetic variance of half the residual variance (1/3 and 2/3 respectively), keeping the total at 1 for comparability, and simulated breeding values according to the specific family structure. We rescaled the genetic and residual effects to keep the total variance comparable to that of the STP design.

A short (long) range was simulated with an autocorrelation parameter of 0.80 (0.93) for the AR model and 20 (7) internal knots for the splines model. In all cases, the same parameter was used for both the row and column directions. The specific values were selected to produce spatial patterns with qualitatively different ranges of approximately equivalent extent across models. Figure 3 shows four random realizations of spatial patterns by model and range (more examples in Supplementary Fig. S2).

The differences in perceived smoothness between the spatial patterns from AR and splines are intrinsic to the nature of these models. Specifically, this feature is caused by a fundamental difference in the curvature of the semivariogram at 0 (e.g., see Cressie 1993, p. 60). Figure 4 displays the corresponding empirical semivariograms for the realized spatial patterns in Fig. 3 (and Supplementary Fig. S3, the corresponding empirical semivariograms for the realized spatial patterns in Supplementary Fig. S2).

The categories of the ratio of the spatial effect variance (σ_s^2) with respect to the residual variance (σ_e^2) (i.e., σ_s^2/σ_e^2) were 0.0625, 0.25, and 0.5 for low, medium, and high, respectively. The R code used for simulating the entire datasets used in this study is given as an Electronic Supplementary Material (Supplementary R_code S1). Additionally, an R code to generate a reduced simulation study mimicking the entire datasets is given also as an Electronic Supplementary Material (Supplementary R_code S2).

Each of the resulting 180 × 12 × 4 = 8640 datasets was fitted with both AR and splines models using a grid search for their respective parameters (i.e., autocorrelation parameter and

Table 2 Akaike information criterion (AIC), estimates of variance components (Genetics, Spatial, Residual), heritabilities (Heritability), and average of the correlation coefficients and root mean square errors (RMSE) from the cross-validation analyses for each of the three spatial mixed models studied (see text for models' abbreviations). Standard errors are in brackets. The corresponding traits analyzed are total height for sites 2.708.3 and 2.707.1, and diameter for site 3.704.2. The lowest AIC and RMSE values, and highest correlations values are highlighted in italics

Trial		Model[a]		
		block	AR	splines
2.708.3	AIC	104,886	*104,454*	104,523
	Genetic	2507 (172)	2484 (182)	2514 (173)
	Spatial	1889 (320)	1786 (172)	1649 (341)
	Residual	4625 (107)	4089 (114)	4359 (111)
	Heritability	0.35 (0.02)	0.38 (0.002)	0.37 (0.02)
	Correlation coefficients	0.54 (0.04)	*0.57* (*0.03*)	0.56 (0.03)
	RMSE	23.65 (1.09)	*23.1* (*1.06*)	23.22 (1.03)
2.707.1	AIC	64,179	*63,722*	63,905
	Genetic	1723 (180)	1882 (224)	1770 (217)
	Spatial	2050 (403)	4804 (378)	4693 (890)
	Residual	11,353 (262)	7998 (163)	10,430 (292)
	Heritability	0.13 (0.01)	0.19 (0.02)	0.15 (0.02)
	Correlation coefficients	0.38 (0.03)	*0.49* (*0.03*)	0.44 (0.04)
	RMSE	33.02 (1.21)	*31.07* (*1.15*)	32.03 (1.24)
3.704.2	AIC	33,830	33,799	*33,798*
	Genetic	7833 (971)	7719 (978)	7731 (1170)
	Spatial	286 (137)	907 (354)	578 (268)
	Residual	18,155 (799)	18,060 (839)	18,070 (824)
	Heritability	0.30 (0.03)	0.30 (0.03)	0.30 (0.04)
	Correlation coefficients	0.24 (0.07)	0.26 (0.07)	*0.27* (*0.07*)
	RMSE	45.38 (1.91)	45.07 (2.04)	*45.05* (*2.08*)

[a] block: block design model, AR: bidimensional separable first-order autoregressive model, splines: bidimensional spline regression model

number of knots, respectively). The search process assumed isotropy (i.e., same parameter value in row and column directions) and fitted seven different values around the true simulated parameter value. The specific values for ρ were defined in a logit scale as $\text{logit}(\rho_0) + (k-4)/2$, $k = 1, \ldots, 7$, where ρ_0 is the value of the autocorrelation parameter used for simulation. The specific values for the number of knots were defined as $\exp(\log(k_0 - 2) + (k-4)/6 + 2)$, $k = 1, \ldots, 7$, where k_0 is the number of knots used for simulation. The resulting values were rounded to two and zero decimal places, respectively. These somewhat arbitrary scales were designed to produce approximately uniformly distributed values in qualitative terms, since, for example, a 0.05 increase in correlation from 0.5 has a very different relevance than from 0.9. The model with the highest likelihood was selected as the outcome of the grid-search processes for the autoregressive and knots parameters.

In summary, for each approach (i.e., AR or splines), we fitted seven models to each of the 8640 datasets, which resulted in more than 120K model fits.

For each dataset and approach, we calculated the following performance metrics:

- AIC: Akaike Information Criterion, see above
- RMSE: root mean square error of prediction of the expected phenotype
- RVRD: residual variance relative deviation, i.e., $\frac{\hat{\sigma}_e^2 - \sigma_e^2}{\sigma_e^2}$
- SCor: correlation between predicted and true spatial effects based on the simulations
- SVRD: spatial variance relative deviation, i.e., $\frac{\hat{\sigma}_s^2 - \sigma_s^2}{\sigma_s^2}$

Note that AIC and RMSE are always positive, and lower values are better. For RVRD and SVRD, values closer to 0 are preferable, while SCor ranges between 0 and 1, and higher values are superior.

Finally, in order to compare the relative performances of the AR and splines approaches, we computed the differences between the absolute values of their metrics ($|m_{Splines}| - |m_{AR}|$) for each simulated dataset. This resulted in a Monte Carlo approximation (with 8640 samples coming from the 180 different scenarios) of the sampling distribution of these relative-performance metrics where positive (negative) values favor the AR (splines) approach, except for SCor where this relationship is reversed.

Data availability The datasets generated and analyzed during the current study are available in the Zenodo repository (Cappa et al. 2019) at https://doi.org/10.5281/zenodo.2629151.

3 Results

3.1 Results from Douglas-fir case studies

Only one of the 12 traits by site combinations resulted in the block model being the best performing analytical method. We present the measures of model comparison for the three representative case studies in Table 2. The equivalent measurements for the rest of the trial by trait combinations are shown in Supplementary Table S1. According to AIC, the block model resulted in the worst fit (i.e., highest AIC), while AR yielded the lowest AICs for the cases with large- and short-scale environmental heterogeneity (2.708.3 and 2.707.1, respectively). For the case with very-short-scale environmental heterogeneity (3.704.2), splines gave the best fit although differences were very small compared to AR (33,798 versus 33,799, respectively). Both AR and splines yielded smaller residual variance estimates than the block model, which resulted ultimately in higher heritabilities. Measures of predictive ability from cross-validation revealed a similar picture of advantages for AR and splines over the block model (correlation coefficient and RMSE in Table 2). In general, differences between the two best fitting models (AR and splines) were more important when environmental variation was at short scale than when compared at larger scales, with AR outperforming splines in terms of AIC and predicting ability.

Apart from fitting quality and heritability increase, accuracy of breeding values and eventual change in candidate ranking between models are also of concern for the breeder. Table 3 shows

Table 3 Accuracy of prediction of breeding values from the block design (block), bidimensional separable first-order autoregressive (AR) and bidimensional spline regression (splines) models, and Spearman correlations coefficients between predicted breeding values of the 10% best candidates for all pairs of models. The corresponding traits analyzed are total height for sites 2.708.3 and 2.707.1, and diameter for site 3.704.2

Trail	Accuracy of breeding values						Spearman correlation of breeding values					
	block		AR		splines		block/AR		block/splines		splines/AR	
	Parents	Offspring	Parents	Offspring	Parents	Offspring	Parents	Offspring	Parents	Offspring	Parents	Offspring
2.708.3	0.87	0.66	0.88	0.67	0.88	0.67	0.95	0.83	0.95	0.86	0.96	0.93
2.707.1	0.70	0.47	0.73	0.51	0.71	0.49	0.81	0.74	0.93	0.88	0.90	0.83
3.704.2	0.86	0.63	0.86	0.63	0.86	0.63	0.96	0.95	0.96	0.95	1.00	1.00

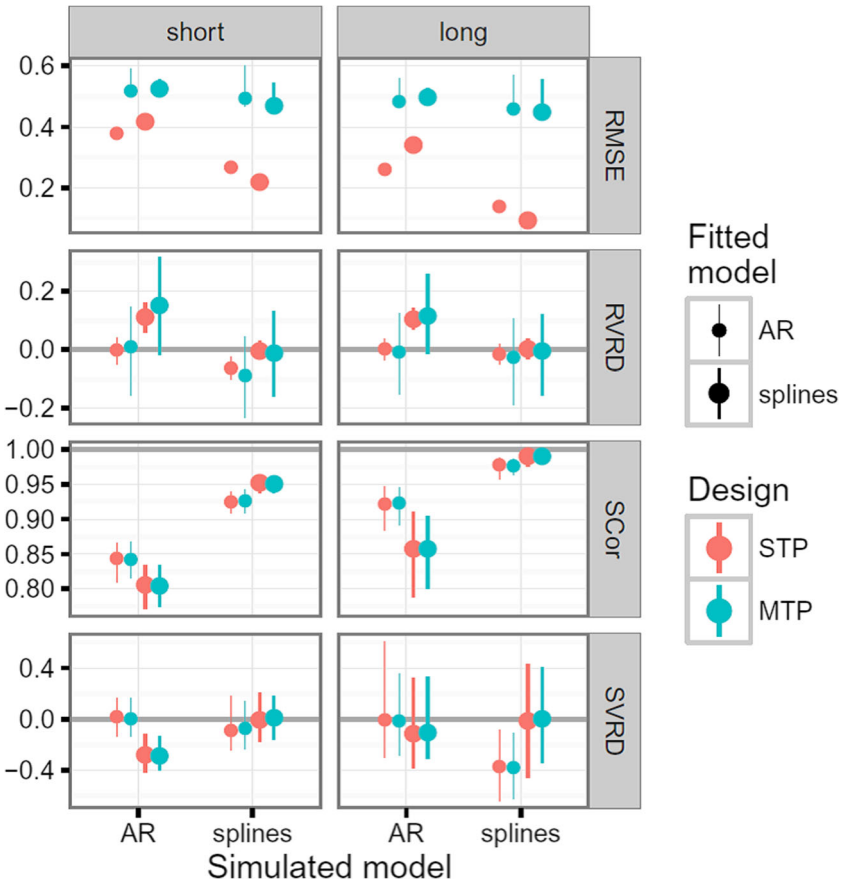


Fig. 5 Median and 90% central quantile of sampling distribution of metrics (root mean squared error of prediction of the expected phenotype -RMSE-; residual variance relative deviation -RVRD- ; correlation between predicted and true simulated spatial effects -Scor-; and spatial variance relative deviation -SVRD-) by fitted model (bidimensional separable first-order autoregressive -AR- and bidimensional spline regression -splines-), spatial range (short and long), and plot structure (design, single-tree plot -STP- and multiple-tree plot -MTP-)

some statistics concerning resulting breeding values from the three models. As expected, parents reached higher accuracies than offspring, and accuracies increased with heritabilities. There were no clear differences between AR and splines models for accuracies, except for the short-scale environmental variation case study, for which AR model presented slightly higher accuracies. Block model accuracies were similar or lower than those of AR and splines models. However, we observed some changes in ranking of best 10% candidates, which were particularly important between block and AR models, presumably because the former takes mostly account for large-scale environmental variation only.

3.2 Results from simulated scenarios

Figure 5 shows a summary of the sampling distributions of the considered performance metrics by fitted model, spatial range, and plot structure. For ease of interpretation, we removed some scenarios that did not add any further insight. Specifically, we kept only the scenarios with highest ratio of spatial to residual variances, which displayed the same patterns than lower ratios but at a higher scale. Furthermore, we show only STP and half-sib MTP designs, since they behave very similar to clone-MTP and full-sib MTP designs, respectively. The full version of the Fig. 5, for all scenarios and mating designs with two additional metrics considered later, is available in Supplementary Fig. S4.

From Fig. 5, we can confirm some expected results concerning the differences among scenarios. For instance, that STP designs generally yielded more accurate predictions (most clearly in terms of RMSE) than MTP, and that MTP designs had larger sampling variability (e.g., standard error), notably in the estimation of the residual variance. In general, long-ranged spatial effects were easier to fit, and spline-generated spatial effects were also more easily recovered than AR counterparts. Furthermore, both approaches yielded unbiased estimations of the variance components when they matched the generating spatial model (i.e., simulated model = fitted model; outermost couple of points in each panel). Conversely, both approaches are biased in the non-matching context. It is important to note that the relevant comparison here is that between fitted models, rather than that between generating models. These latter were simply two alternative ways of generating a "reality," the spatial heterogeneity, for which the experimenter has no clue of its nature.

The relative performances of the splines and AR approaches under these settings are more accurately assessed from Fig. 6, which displays the differences in absolute value between the splines and AR metrics by spatial range, simulated model, and plot structure. The distributions over the "AR" column tended to be concentrated further away from 0 (meaning larger differences between splines and AR metrics) than their corresponding counterparts over the "splines" column, and this happened for most

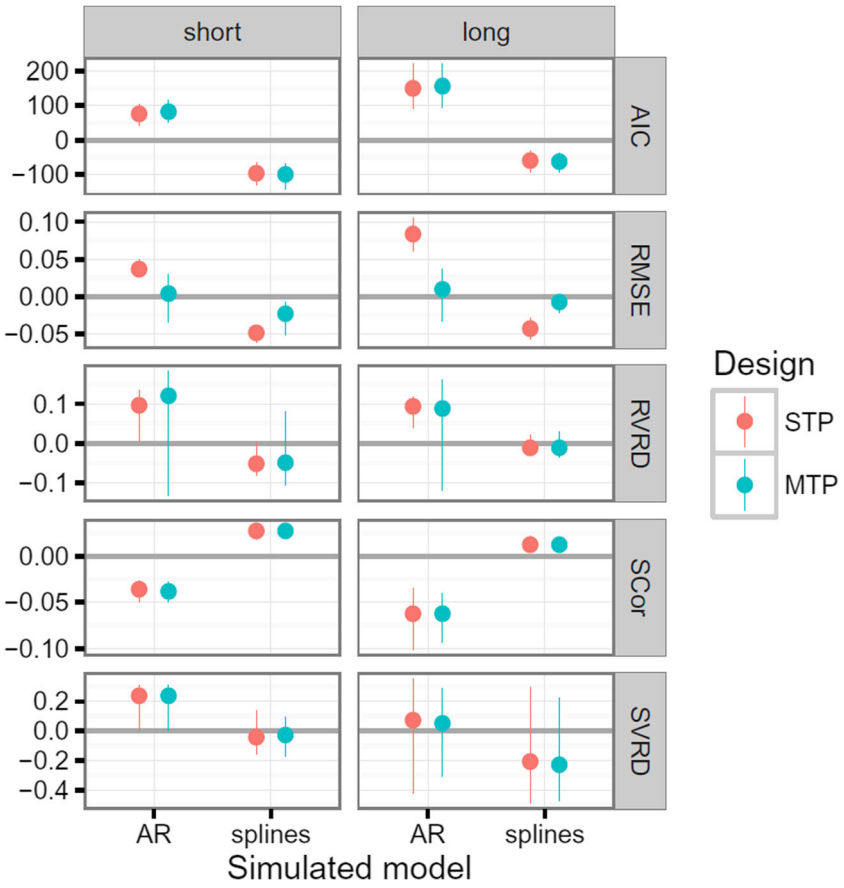


Fig. 6 Median and 90% central confidence interval of differences between the bidimensional spline regression (splines) and bidimensional separable first-order autoregressive (AR) absolute metrics (i.e., $|m_{Splines}| - |m_{AR}|$ for m equal to root mean square error of prediction of the expected phenotype -RMSE-; residual variance relative deviation -RVRD-; correlation between predicted and true simulated spatial effects -Scor-; and spatial variance relative deviation -SVRD-) by spatial range (short and long) and plot structure (design, single-tree plot -STP- and multiple-tree plot -MTP-). Positive (negative) values indicate a better performance by AR (splines), except for SCor where the relationship is the reverse

metrics except SVRD at long range which showed the contrary. For example, with spline-generated data, the AR approach is almost as good as the splines when it comes to estimating the residual variance, or in terms of correlation between simulated and predicted spatial effect. There is a more remarkable difference in these metrics in favor of AR when the spatial effect is generated by an AR model. We can conclude that even if the differences in performance are relatively small with respect to the target values, it is easier for the AR approach to fit spline-generated data than the converse. This is particularly true under a scenario of long-ranged spatial effect and STP structure (e.g., in terms of RMSE). Under MTP structure, however, the RMSE increases dramatically for both approaches and the advantage of the AR over splines vanishes almost completely.

Figure 7 displays the marginal distributions of the semivariogram metrics residual variance discrepancy (RV_disc) and semivariogram slope (vgslope) by (simplified) scenario, design, and approach. Note how the discrepancy between the empirical and estimated residual variance is most important under MTP designs, where there is the additional confounding of the spatialized genetic arrangement. For STP designs on the other hand, the discrepancy is more important for short- rather than for long-ranged spatial effects, and for the AR than for the splines approach, due to the weaker partial pooling. The slope of the empirical semivariogram is systematically negative, except for the case where the data were simulated with a long-ranged AR spatial process and fitted with a splines approach. This is certainly due to unaccounted autocorrelation at a shorter scale, and reflects the mismatch in the covariance model. Otherwise, for the AR approach under STP structure at short-ranged spatial scale, the negative slope is remarkably steeper than for the splines approach, both for AR- or splines-simulated data. This is likely a side effect of the increased correlation between spatial and residual BLUPs. It is not the semivariogram peak which is taller but the empirical variance which is lower. This can be seen from the negative correlations between these two metrics for the non-matching cases in Supplementary Fig. S5. From the same Fig. S5, we can see that neither the empirical discrepancy with the estimated residual variance (RV_disc) nor the semivariogram slope (vgslope) is associated with any other metric. This suggests that the discrepancy and slope are not signs of over- or under-fitting but a generally expected outcome. Moreover, the fact that the AR approach usually displays a steeper slope than splines does not hamper its predictive ability.

4 Discussion

In genetic evaluation and quantitative genetic analyses in forest trees, environmental heterogeneity is of relevance given

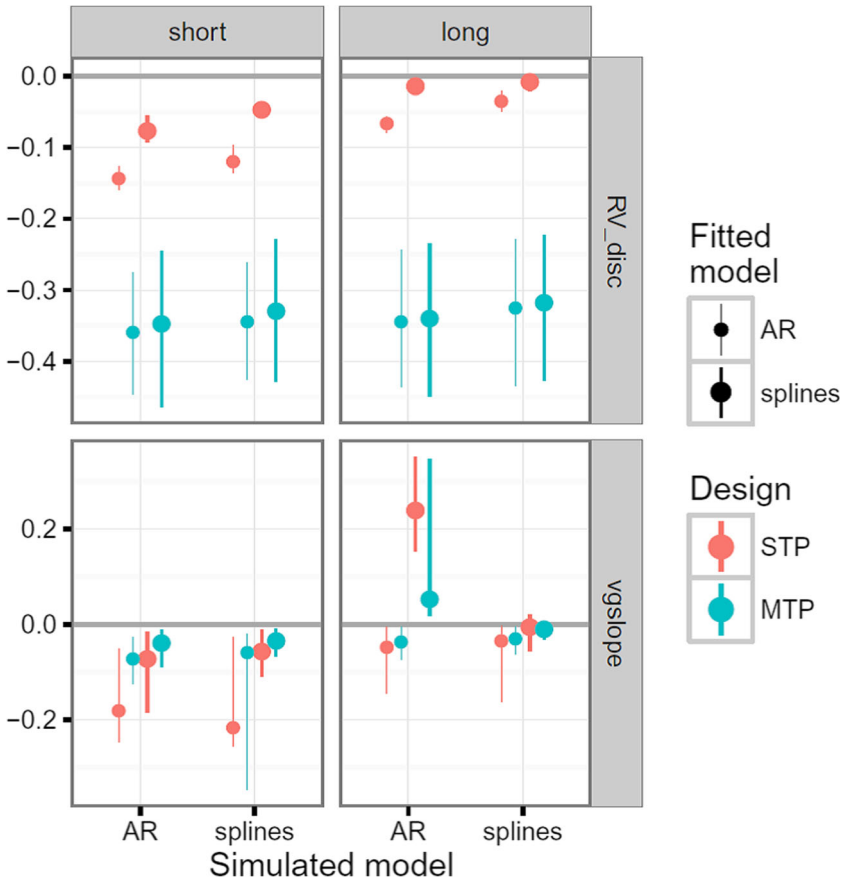


Fig. 7 Median and 90% central quantile of sampling distribution of variogram metrics (empirical and estimated residual variances differences -RV_disc- and slope of the empirical semivariogram of residuals at distance 0 -vgslope-) by fitted model (bidimensional separable first-order autoregressive -AR- and bidimensional spline regression -splines-), spatial range (short and long), and plot structure (design, single-tree plot -STP- and multiple-tree plot -MTP-)

that trials usually comprise large areas often established in marginal heterogeneous lands, which in turn can exacerbate the magnitude of environmental effects and the need to account precisely for them in the evaluation process. In the era of genome-driven accuracy, we should not forget that environmental terms also need to be assessed with precision in the populations used for calibration, in order to be able to split precisely genetic and environmental effects. In this study, we compared one of the most common and classical methods to account for environmental heterogeneity in experimental trials, the block model, which is an a priori method based on design, with two alternatives that are based on a posteriori analyses using an empirical Douglas fir data and the a posteriori AR and splines models under different simulated scenarios. Of these latter methods, the AR model is one of the most popular methods in the spatial literature in forest and crop evaluation (e.g., Qiao et al. 2000; Smith et al. 2001; Costa e Silva et al. 2001; Dutkowski et al. 2006). P-splines represents a methodologically distinct approach which is considered suitable to fitting trends over large scales than those usually associated to AR models.

We found that a posteriori analyses by AR and splines models clearly outperformed the block model. This was the case across a series of 11 out of the 12 traits by site combinations comprising eight sites from the Douglas fir genetic evaluation program that were used in the present study. Similar outcomes have been found in previous works involving forest genetic trials, when comparing a block model to AR (Costa e Silva et al. 2001; Dutkowski et al. 2006; Ye and Jayawickrama 2008) and splines (Cappa et al. 2011 and Cappa et al., 2015a) models. Other a posteriori approaches for spatial modeling showed also improvement of estimates of genetic parameters and breeding values over standard designs, like kriging (Hamann et al. 2002, Zas 2006), and nearest neighbor techniques (Anekonda and Libby 1996, Joyce et al. 2002). In summary, there is a considerable amount of work, of which we referenced just a few studies, all showing the benefits of a posteriori spatial analyses over classical a priori design-based approaches. However, there is a lack of comparisons between some of the best a posteriori methods.

This study presents one of the first comparisons between two methodologically distinct a posteriori methods by using both empirical and simulated data. Our empirical results revealed that differences between AR and splines models in terms of fitting and predicting ability, although in absolute terms favorable to the former, were relatively small or difficult to discern when accounting for replicate variation in the cross-validation analysis. These results are consistent with those obtained from Velazco et al. (2017) where the performance of the splines model was equivalent to the AR model when considering the improvement in the precision and the predictions of genetic values, even though they considered anisotropic models and a different covariance

structure in the splines model. Our results showed, however, that the differences between splines and AR were of somehow greater magnitude for cases where spatial heterogeneity happened at relatively short scale, which is well in agreement with what is considered a favorable scenario for the AR model (Gilmour et al. 1997). Therefore, the AR model could well be considered a good general option for most field testing situations, to the level of generality that can be drawn from the experimental coverage that was used in the present study. This study also comported a systematic comparison between the two a posteriori methods involving a comprehensive collection of simulated data covering a wide range of parametric situations. In general, results from simulated data were well in agreement with those from empirical data. Both methods performed similarly well, although AR model seemed to handle more easily all kinds of spatial scenarios. Notably, splines-based data resulted in AR fits being almost as good as those obtained with splines counterparts, while AR-based data were in general more challenging for the fit with splines. These advantages were, however, of very small magnitude in relative terms. Whenever MTP were fitted, any differences between the two models vanished as a consequence of a decrease in predictive ability for both models.

The result in the comparison between AR and splines models brings an important point here: the fact that the best model is a case-based choice and that for this there are now excellent alternatives to classical designs. This was already presented by Gilmour et al. (1997), and other authors who reached similar conclusions when facing analytical choices for spatial data in agriculture. Although the gain by perfecting the choice for each dataset between different a posteriori methods might appear as negligible in most cases, it is clear that the exercise can give the breeder an excellent insight into the way spatial heterogeneity affects genetic evaluation.

By means of the simulation experiment, we showed that some discrepancy between the variance of the empirical residuals and the estimated residual variance along with some non-flat slope in the first few lags of the empirical residual semivariogram (Fig. 7) are to be expected and are not an indication of model misspecification of any sort. Indeed, Stein (1999) shows that it is not generally possible to recover the covariance function of a Gaussian random field from observations in a bounded region (see Section 6.3 in Stein 1999). The reason is that the empirical residuals are predicted conditional to the observations, and thus they are neither independent nor identically distributed in general. In particular, they are prone to be positively correlated with other random effects in the model, which are also conditional on the data. This is enhanced by high-dimensional (e.g., individual level) random effects with little shrinkage (also known as partial pooling, or information sharing) such as short-ranged AR models. The positive correlation among BLUPs explains that the sum of the empirical variances of the effects is lower than the empirical variance of the sum (i.e., the phenotypic variance), and in turn, why the empirical semivariogram of residuals is shifted downwards from its theoretical level. On the other hand, the negative slope at distance 0 is a consequence of the partial pooling performed by the spatial effect, where the spatial BLUPs are shrunk towards the local average value. The remains of this shrinkage are left over as increased residual variance in the differences among neighboring residuals, which can be detected in the first few semivariogram lags.

Our simulation results also demonstrated that the AR and splines approaches yielded unbiased estimations of the variance components when they match the generating spatial model (Fig. 5). This contrasts with the results from Rodríguez-Álvarez et al. (2018), who found a slight but systematic bias in the estimates of the spatial and residual variances for the AR model (see Table 4 in Rodríguez-Álvarez et al. 2018). However, this bias is most apparent for extremely low values of the autocorrelation parameter ($\rho = 0.1$) which causes a lack of identifiability with the residual component. In our work, we did not consider these low levels of autocorrelation since they do not produce spatial effects of any practical interest as argued in Section 2. The remaining discrepancy can be explained by differences in the implementation. Specifically, our implementation was restricted to an isotropic grid search among 7 candidate values of the autocorrelation parameter, which is estimated within the REML algorithm in Rodríguez-Álvarez et al. (2018).

Our work has focused on positive correlations between two adjoining trees caused by small- and large-scale environmental variation. Moreover, it showed how these environmental variations can interact with other non-environmental factors occurring at the same spatial scale (i.e., mating design and plot structure), and how these interactions can affect the performance of the AR and splines models. Interplant competition may be another non-environmental factor at small-scale spatial variation that may affect the performance of the AR and splines models. Tree competition for resources may also bias breeding value estimation from competing individuals by inducing a negative correlation between either individual trees or neighbor plots, and is caused by genetic and environmental sources (e.g., Cappa and Cantet 2008; Costa e Silva and Kerr 2013). In forest genetic trials, both phenomena (i.e., competition and environmental heterogeneity) are dynamic and coexist simultaneously (e.g., Magnussen 1994; Cappa et al. 2015b). We would expect that splines will be less affected than the AR models by a very short-scale disturbance, for instance, due to competition. Further study is required on this topic.

5 Implications for French Douglas fir breeding program and conclusion

Progeny test of the French Douglas-fir breeding program offered a good opportunity to test alternative methods for modeling spatial heterogeneity. Many of these trials are large, with surfaces often being between 5 and 10 ha, and

established on marginal heterogeneous lands, which are good preconditions for the use of efficient modeling approaches. Results suggested that there is a substantial gain in accuracy and precision in switching from classical a priori blocks design to any of the two alternative a posteriori methodologies. Simulations, covering a larger though oversimplified hypothetical setting, seemed to support previous empirical findings. As a result of this, one possibility offered by the use of these alternative methods for future trials would be to trade the extra precision and accuracy granted by these a posteriori methods for some reduction in trial size. In addition, there is a number of recommendations that are fully applicable to our Douglas-fir breeding program, but that could be also of general use for any program relying on progeny testing. These recommendations are the following:

1. In practice, both models (i.e., AR and splines) suitably capture spatial variation. It is usually safe to use any of them. The final choice could be driven solely by operational reasons. Sometimes, it will be more convenient to use AR (e.g., faster), or splines (e.g., irregular arrangement of observations). In most cases, changes in estimated parameters and predicted values will be negligible.
2. It might be worth to assess second-order behavior by visual inspection of the empirical semivariogram of residuals from a non-spatial model and determine whether it is more similar to an AR process or to a splines counterpart. Otherwise, an AR process emerges as a safe default choice. If possible, it is recommended to fit both models and confirm that they yield similar results. If they do not, then investigate possible causes.
3. It might be worth to combine multiple spatial models, notably if they can capture spatial variation at different scales or different sources of variation (i.e., block and splines, or block and AR) to capture spatial variation at different scales or different sources of variation.
4. The empirical semivariogram of residuals from a fitted model with a spatial effect (especially AR models, under MTP designs) can sometimes display a sudden initial drop and a subsequent stabilization at a value below the estimated residual variance (see Fig. 1). This is expected and it does not necessarily mean that the model is mis-specified, nor that the predictions are wrong.

Acknowledgments The authors sincerely acknowledge Jean-Charles Bastien for his help in identifying trials and accessing data. Thanks go to the staff of INRA experimental units (UE GBFOR, INRA Val de Loire) who have established, maintained, and assessed the field trials.

Funding Eduardo P Cappa, F. Muñoz, and L. Sánchez received funding from the European Union's Seventh Framework Program for research, technological development, and demonstration under grant agreement no. 284181 ("Trees4Future"). F. Muñoz is partially funded by research grant MTM2016-77501-P from the Spanish Ministry of Economy and Competitiveness.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans on Automat Contr 19(6):716–723

Anekonda TS, Libby WJ (1996) Effectiveness of nearest neighbor data adjustment in a clonal test of redwood. Silvae Genet 45(1):46–51

Bastien JC, Sánchez L, Michaud D (2013) Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco). In: Ecosystems PLEMF (ed) Forest tree breeding in Europe, vol 24. Springer, New York, pp 325–369

Cappa EP, Muñoz F, Sanchez L (2019) Performance of alternative spatial models in empirical Douglas-fir and simulated datasets. V1. Zenodo. [dataset]. https://doi.org/10.5281/zenodo.2629151

Cappa EP, Yanchuk AD, Cartwright CV (2015a) Estimation of genetic parameters using spatial analysis in *Tsuga heterophylla* full-sibling family trials in British Columbia. Silvae Genet 64:59–73

Cappa EP, Muñoz F, Sanchez L, Cantet RJC (2015b) A novel individual-tree mixed model with competition effects and environmental heterogeneity: a Bayesian approach. Tree Genet Genomes 11:120–135

Cappa EP, Lstiburek M, Yanchuk AD, El-Kassaby YA (2011) Two-dimensional penalized splines via Gibbs sampling to account for spatial variability in forest genetic trials with small amount of information available. Silvae Genet 60:25–35

Cappa EP, Cantet RJC (2008) Direct and competition additive effects in tree breeding: Bayesian estimation from an individual tree mixed model. Silvae Genet 57:45–56

Cappa EP, Cantet RJC (2007) Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. Can J For Res 37:2677–2688

Costa e Silva J, Kerr RJ (2013) Accounting for competition in genetic analysis, with particular emphasis on forest genetic trials. Tree Genet Genomes 9:1–17

Costa e Silva J, Dutkowski GW, Gilmour AR (2001) Analysis of early tree height in forest genetic trials is enhanced by including a spatially correlated residual. Can J For Res 31:1887–1893

Cressie N (1993) *Statistics for Spatial Data*. Wiley series in probability and statistics. Wiley, New York

Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. J Agric Biol Environ Stat 11:381–393

Dutkowski GW, Costa e Silva J, Gilmour AR, Lopez GA (2002) Spatial analysis methods for forest genetic trials. Can J For Res 32:2201–2214

Dutkowski GW, Costa e Silva J, Gilmour AR, Wellendorf H, Aguiar A (2006) Spatial analysis enhances modeling of a wide variety of traits in forest genetic trials. Can J For Res 36:1851–1870

Eilers PHC, Marx BD (2003) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. Chemometr Intell Lab Syst 66:159–174

Ericsson T (1997) Enhanced heritabilities and best linear unbiased predictors through appropriate blocking of progeny trials. Can J For Res 27:2097–2101

Federer WT (1998) Recovery of interblock, intergradient, and intervarietal information in incomplete block and lattice rectangle designed experiments. Biometrics 54:471–481

Fu YB, Yanchuk AD, Namkoong G (1999) Incomplete block designs for genetic testing: some practical considerations. Can J For Res 29: 1871–1878

Gezan SA, White TL, Huber DA (2010) Accounting for spatial variability in breeding trials: a simulation study. Agron J 102:1562–1571

Gezan SA, Huber DA, White TL (2006) Post hoc blocking to improve heritability and precision of best linear unbiased genetic predictions. Can J For Res 36:2141–2147. https://doi.org/10.1139/X06-112

Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. J Agric Biol Environ Stat 2:269–293

Grondona MO, Crossa J, Fox PN, Pfeiffer WH (1996) Analysis of variety yield trials using two-dimensional separable ARIMA processes. Biometrics 52:763–770

Hamann A, Koshy M, Namkoong G (2002) Improving precision of breeding values by removing spatially autocorrelated variation in forestry field experiments. Silvae Genet 51:210–215

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph, Ont, Canada

Joyce D, Ford R, Fu YB (2002) Spatial patterns of tree height variations in a black spruce farm-field progeny test and neighbors-adjusted estimations of genetic parameters. Silvae Genet 51:13–18

Kroon J, Andersson B, Mullin TJ (2008) Genetic variation in the diameter-height relationship in scots pine (*Pinus sylvestris*). Can J For Res 38:1493–1503

Lopez GA, Potts BM, Dutkowski GW, Apiolaza LA, Gelid P (2002) Genetic variation and inter-trait correlations in *Eucalyptus globulus* base population trials in Argentina. For Genet 9:223–237

Manly BFJ (1991) Randomization, bootstrap and Monte Carlo methods in biology, 2nd edn. Chapman and Hall/CRC, New York

Magnussen S (1993) Bias in genetic variance estimates due to spatial autocorrelation. Theor Appl Genet 86:349–355

Magnussen S (1994) A method to adjust simultaneously for spatial microsite and competition effects. Can J For Res 24:985–995

Misztal I (1999) Complex models, more data: simpler programming. Proc Inter Workshop Comput Cattle Breed '99, March 18-20, Tuusala, Finland. Interbull Bul. 20:33-42

Muñoz F, Sanchez L (2015) breedR: statistical methods for forest genetic resources analysts. R package version 0.7–16. https://github.com/famuvie/breedR

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554

Qiao CG, Basford KE, DeLacy IH, Cooper M (2000) Evaluation of experimental designs and spatial analyses in wheat breeding trials. Theor Appl Genet 100:9–16

Rodríguez-Álvarez MX, Boer MP, van Eeuwijk FA, Eilers PHC (2018) Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spatial Statistics 23:52–71

Saenz-Romero C, Nordheim EV, Guries RP, Crump PM (2001) A case study of a provenance/progeny test using trend analysis with correlated errors and SAS PROC MIXED. Silvae Genet 50:127–135

Smith AB, Cullis BR, Gilmour A (2001) The analysis of crop variety evaluation data in Australia. Aust N Z J Stat 43:129–145

Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer-Verlag, New York

Sørbye SH, Rue H (2014) Scaling intrinsic Gaussian Markov random field priors in spatial modelling. Spat Stat 8:39–51

Thomson AJ, El-Kassaby YA (1988) Trend surface analysis of provenance-progeny transfer data. Can J For Res 18: 515–520

Velazco JG, Rodríguez-Álvarez MX, Boer MP, Jordan DR, Eilers PH, Malosetti M, van Eeuwijk FA (2017) Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. Theor Appl Genet 130:1375–1392. https://doi.org/10.1007/s00122-017-2894-4

Verbyla AP, Cullis BR, Kenward MG, Welham SJ (1999) The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). Appl Stat 48:269–311

Ye TZ, Jayawickrama KJS (2008) Efficiency of using spatial analysis in first-generation coastal Douglas-fir progeny tests in the US Pacific Northwest. Tree Genet Genomics 4:677–692

Zas R (2006) Iterative kriging for removing spatial autocorrelation in analysis of forest genetic trials. Tree Genet Genomics 2:177–185