Supplementary material

Correcting for spatial heterogeneity in plant breeding experiments with P-splines

María Xosé Rodríguez - Álvarez[1,2], Martin P. Boer[3], Fred A. van Eeuwijk[3], Paul H. C. Eilers[4]

[1] BCAM - Basque Center for Applied Mathematics
Alameda de Mazarredo, 14. E-48009 Bilbao, Basque Country, Spain
mxrodriguez@bcamath.org
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[3] Biometris, Wageningen University & Research, Wageningen, the Netherlands
[4] Erasmus University Medical Centre, Rotterdam, the Netherlands

This document contains supplementary material to the paper "Correcting for spatial heterogeneity in plant breeding experiments with P-splines". In Web Appendix A the nested B-spline bases (Lee et al., 2013) are presented. Web Appendix B describes the estimation procedure implemented in the `SpATS` package that accompany this paper. In Web Appendix C we present some results on hat matrices. Web Appendix D shows the equivalence between the definition given by Cui et al. (2010) of the effective dimension associated to a model's component and that used in this paper. In Web Appendix E the uniformity barley data discussed in Section 2 of the main manuscript are revisited. Finally, Web Appendix F contains a brief usage description of the `SpATS` package.

Web Appendix A Nested bases

With large data sets the computation of $\boldsymbol{Z}_s$ in eqn. (16) of the main manuscript, as well as its inner product, can demand a lot of time, especially for large values of $\hat{m}$ and $\breve{m}$. Lee et al. (2013) propose to speed up computation by using nested bases. The idea is to reduce the dimension of the marginal B-spline bases $\hat{\boldsymbol{B}}$ and $\breve{\boldsymbol{B}}$ (and therefore the associated number of coefficients to be estimated), but only for the smooth-by-smooth interaction term , i.e., $f_{u,v}$. As pointed out by the authors, this simplification can be justified by the fact that the main effects, f_u and f_v, and the smooth varying coefficient terms, h_v and h_u, may in fact explain most of the structure (or spatial trend) presented in the data, and so a less

rich representation of the interaction term would be sufficient.

Let $\hat{\boldsymbol{B}}_N$ and $\breve{\boldsymbol{B}}_N$ be two reduced marginal B-spline bases of dimension $n \times \hat{m}_N$ ($\hat{m}_N < \hat{m}$) and $n \times \breve{m}_N$ ($\breve{m}_N < \breve{m}$) with associated penalty matrices $\hat{\boldsymbol{D}}_N^t \hat{\boldsymbol{D}}_N$ and $\breve{\boldsymbol{D}}_N^t \breve{\boldsymbol{D}}_N$, respectively. Then, the reduced mixed model matrix $\boldsymbol{Z}_s$ for the PS-ANOVA model is constructed as follows

$$\boldsymbol{Z}_s = \left[\boldsymbol{Z}_v \mid \boldsymbol{Z}_u \mid \boldsymbol{Z}_v \square \boldsymbol{u} \mid \boldsymbol{v} \square \boldsymbol{Z}_u \mid \widetilde{\boldsymbol{Z}}_v \square \widetilde{\boldsymbol{Z}}_u \right],$$

where $\widetilde{\boldsymbol{Z}}_v = \hat{\boldsymbol{B}}_N \widetilde{\boldsymbol{U}}_v^N$ and $\widetilde{\boldsymbol{Z}}_u = \breve{\boldsymbol{B}}_N \widetilde{\boldsymbol{U}}_v^N$, with $\widetilde{\boldsymbol{U}}_v^N$ and $\widetilde{\boldsymbol{U}}_v^N$ being the matrices containing the eigenvectors associated to the non-zero eigenvalues of $\hat{\boldsymbol{D}}_N^t \hat{\boldsymbol{D}}_N$ and $\breve{\boldsymbol{D}}_N^t \breve{\boldsymbol{D}}_N$, respectively.

In order to ensure that the reduced model is in fact nested in the model including only the main effects, Lee et al. (2013) showed that the number of segments that define $\hat{\boldsymbol{B}}_N$ and $\breve{\boldsymbol{B}}_N$ should be a divisor of the number of segments used in the original bases $\hat{\boldsymbol{B}}$ and $\breve{\boldsymbol{B}}$. The reasoning behind is graphically illustrated in Web Figure 1, that has been taken from the paper by Lee et al. (2013). The two top plots depict two cubic B-spline bases of dimension 11 and 7, respectively. The squares and triangles denote the breakpoints (knots) that define each segment. As can be observed, the knots of the small basis correspond to a subset of the knots of the large basis. This implies that the space spanned by the small basis is a subset of the space spanned by the large one. This can be seen on the plot at the bottom, where both bases overlap.

Note that the use of nested bases reduces the number of coefficients associated to $f_{u,v}$ from $(\hat{m}-2)(\breve{m}-2)$ to $(\hat{m}_N-2)(\breve{m}_N-2)$. This reduction allows being generous with the number of B-spline basis functions used for the main effects and the smooth varying coefficient terms. Our experience suggests using (a) as many segments for $\hat{\boldsymbol{B}}$ and $\breve{\boldsymbol{B}}$ as number of rows and columns in the field, respectively; and (b) half the number of segments for the nested basis.

Web Appendix B Mixed model estimation procedure implemented in the SpATS package

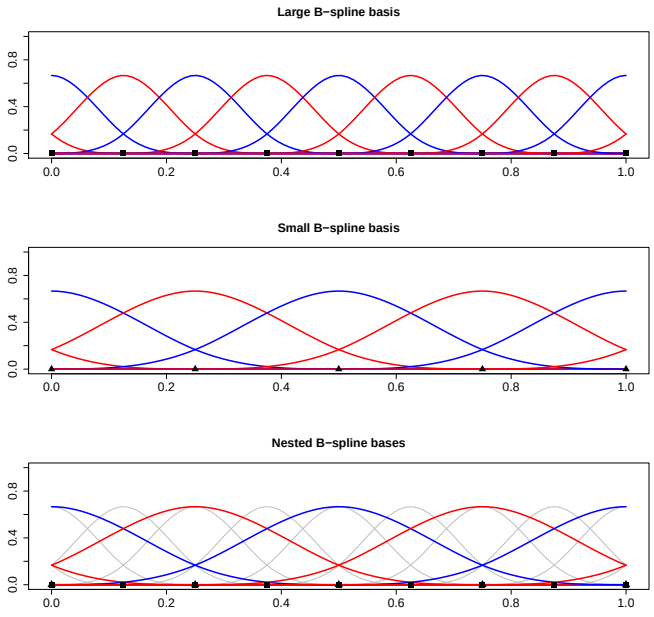
For given values of the variance components (σ^2 and σ_k^2, $k = 1, \ldots, q$), BLUEs for $\boldsymbol{\beta}$ and BLUPs for $\boldsymbol{c}$ can be obtained as the solution to the linear system of equations (Henderson, 1963)

$$\underbrace{\begin{bmatrix} \boldsymbol{X}^t \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{X}^t \boldsymbol{R}^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^t \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{G}^{-1} + \boldsymbol{Z}^t \boldsymbol{R}^{-1} \boldsymbol{Z} \end{bmatrix}}_{\boldsymbol{C}} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{c}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^t \boldsymbol{R}^{-1} \boldsymbol{y} \\ \boldsymbol{Z}^t \boldsymbol{R}^{-1} \boldsymbol{y} \end{bmatrix} \tag{B1}$$

which gives rise to closed-form expressions

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{y}, \tag{B2}$$

$$\widehat{\boldsymbol{c}} = \boldsymbol{G} \boldsymbol{Z}^t \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X} \widehat{\boldsymbol{\beta}}) = \boldsymbol{G} \boldsymbol{Z}^t \boldsymbol{Q} \boldsymbol{y}, \tag{B3}$$



Web Figure 1: Example of nested B-spline bases.

where $\boldsymbol{Q} = \boldsymbol{V}^{-1} - \boldsymbol{V^{-1}X}\left(\boldsymbol{X}^t\boldsymbol{V^{-1}X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V^{-1}}$ with $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{ZGZ}^t$ and $\boldsymbol{R} = \sigma^2\boldsymbol{I}_n$. As will be seen in Web Appendix C, the last equivalence in (B3) plays an important role in our approach, and it is simply obtained by substituting $\widehat{\boldsymbol{\beta}}$ by (B2).

As far as estimation of variance components is concerned, these can be obtained, as usual, by maximising the REML log-likelihood function

$$l = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X^tV^{-1}X}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^t\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}). \tag{B4}$$

Given that in our approach $\boldsymbol{G}$ is a linear function of variance components, estimation can be accommodated using standard mixed model procedures, as, e.g., those implemented in the R-packages `asreml-R`, `nlme` and `lme4`, or the `PROC MIXED` procedure in SAS®. We now present the numerical procedure implemented in the R-package `SpATS` that accompany this paper. The procedure presents many appealing features, which make it a good candidate for the analysis of field trials: (a) it is fast and stable; (b) it is robust (i.e., it converges from almost any starting values); and (c) it always provides positive estimates of the variance components, although it is possible to obtain values very close to zero. For all these reasons it has been our choice.

As said, REML estimates of the variance components are obtained by maximising (B4). Taking derivatives with respect to the variance components σ^2_k $(k = 1, \ldots, q)$, we obtain (see e.g., Rodríguez-Álvarez et al., 2015; Johnson and Thompson, 1995)

$$\frac{\partial l}{\partial \sigma^2_k} = -\frac{1}{2}trace\left(\boldsymbol{Z}^t\boldsymbol{QZG}\frac{\partial \boldsymbol{G}^{-1}}{\partial \sigma^2_k}\boldsymbol{G}\right) + \frac{1}{2}\widehat{\boldsymbol{c}}^t\frac{\partial \boldsymbol{G}^{-1}}{\partial \sigma^2_k}\widehat{\boldsymbol{c}},$$

where $\boldsymbol{Q} = \boldsymbol{V}^{-1} - \boldsymbol{V^{-1}X}\left(\boldsymbol{X}^t\boldsymbol{V^{-1}X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V^{-1}}$ with $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{ZGZ}^t$ and $\boldsymbol{R} = \sigma^2\boldsymbol{I}_n$. By eqn. (7) of the main manuscript, it is easy to show that the former derivatives can be expressed as

$$2\frac{\partial l}{\partial \sigma^2_k} = -\frac{1}{\sigma^2_k}trace\left(\boldsymbol{Z}^t_k\boldsymbol{QZ}_k\boldsymbol{G}_k\right) + \frac{1}{\sigma^4_k}\widehat{\boldsymbol{c}}^t_k\boldsymbol{\Lambda}^{-1}_k\widehat{\boldsymbol{c}}_k.$$

Then, REML estimates of the variance components are found by equating the former expression to zero, which gives

$$\widehat{\sigma}^2_k = \frac{\widehat{\boldsymbol{c}}^t_k\boldsymbol{\Lambda}^{-1}_k\widehat{\boldsymbol{c}}_k}{\mathrm{ED}_k}, k = 1, \ldots, q, \tag{B5}$$

with

$$\mathrm{ED}_k = \mathrm{trace}\left(\boldsymbol{Z}^t_k\boldsymbol{QZ}_k\boldsymbol{G}_k\right). \tag{B6}$$

An estimate of σ^2 can also be easily obtained following the same reasoning (see Rodríguez-Álvarez et al., 2015, for details). Specifically, in this case we have

$$\widehat{\sigma}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^t\widehat{\boldsymbol{\varepsilon}}}{\mathrm{ED}_{\boldsymbol{\varepsilon}}}, \tag{B7}$$

where $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{c}}$ and

$$\text{ED}_{\boldsymbol{\varepsilon}} = \text{trace}\,(\boldsymbol{RQ}) = n - \text{rank}(\boldsymbol{X}) - \sum_{k=1}^{q} \text{ED}_k. \tag{B8}$$

As can be seen, the right-hand side of eqn. (B6) depends on the unknown variance components. Hence, eqns. (B5) and (B7) need to be solved with an iterative procedure. Given some starting values for the variance components, estimation of model (6) of the main manuscript is thus obtained by iterating, until convergence, among (a) estimating the fixed and random effect coefficients (linear system (B1)); (b) evaluating the right-hand side of eqn. (B6); and (c) updating the variances by means of eqns. (B5) and (B7). In this work, the REML-deviance was used as the convergence criterion.

To the best of our knowledge, this iterative algorithm was originally proposed by Henderson in an unpublished manuscript, and discussed in detail by Harville (1977) and Engel (1990), among others. Schall (1991) further extended the algorithm for the estimation of generalised linear mixed models. In the P-spline literature, the algorithm has been also successfully used (e.g., Schnabel and Eilers, 2009; Lee et al., 2013), and usually referred to as Schall's algorithm.

From a computational point of view, the traces in (B6) may involve the computation and manipulations of several large matrices. However, there are several ways this computation can be relaxed. For instance, using some results on mixed models (see, e.g., Johnson and Thompson, 1995) we have that

$$\boldsymbol{Z}_k^t \boldsymbol{Q} \boldsymbol{Z}_k \boldsymbol{G}_k = \boldsymbol{I}_{m_k} - \boldsymbol{G}_k^{-1} \boldsymbol{C}_{kk}^* = \boldsymbol{I}_{m_k} - \frac{1}{\sigma_k^2} \boldsymbol{\Lambda}_k^{-1} \boldsymbol{C}_{kk}^*, \tag{B9}$$

where m_k is the number of random coefficients in the vector $\boldsymbol{c}_k$, $\boldsymbol{C}^*$ is the inverse of $\boldsymbol{C}$ in (B1), and $\boldsymbol{C}_{kk}^*$ is that partition of $\boldsymbol{C}^*$ corresponding to $\boldsymbol{c}_k$. An alternative approach would be to use the result given in eqn. (5.3) in Harville (1977)

$$\boldsymbol{Z}^t \boldsymbol{Q} \boldsymbol{Z} = \boldsymbol{G}^{-1} \boldsymbol{C}_m^* \left[\boldsymbol{X} \mid \boldsymbol{Z}\right]^t \boldsymbol{R}^{-1} \boldsymbol{Z},$$

where $\boldsymbol{C}_m^*$ denotes the matrix formed by the last m rows of $\boldsymbol{C}^*$ (with $m = \sum_{k=1}^{q} m_k$). Here, the block-diagonal elements of $\boldsymbol{Z}^t\boldsymbol{Q}\boldsymbol{Z}$ correspond to $\boldsymbol{Z}_k^t\boldsymbol{Q}\boldsymbol{Z}_k$. Despite the apparent complexity of this expression, in order to compute (B6) only the diagonal of $\boldsymbol{Z}^t\boldsymbol{Q}\boldsymbol{Z}$ needs to be explicitly obtained, since $\boldsymbol{G}$ is a diagonal matrix (see Rodríguez-Álvarez et al., 2015, for further details).

Web Appendix C Some results on hat matrices

Let's denote as $\boldsymbol{H}$ the "hat" matrix, defined as $\boldsymbol{H}\boldsymbol{y} = \widehat{\boldsymbol{y}}$. Expressions (B2) and (B3) reveal that we can define two separate hat matrices: one for the fixed part of the model and one

for the random part. Specifically

$$\boldsymbol{H}\boldsymbol{y} = \widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{c}} = \boldsymbol{H}_F\boldsymbol{y} + \boldsymbol{H}_R\boldsymbol{y},$$

where $\boldsymbol{H}_F = \boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}$ and $\boldsymbol{H}_R = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{Q}$.

However, we can go one step further. It is worth remembering that in this paper we assume that $\boldsymbol{Z} = [\boldsymbol{Z}_1 \mid \ldots \mid \boldsymbol{Z}_q]$, where each $\boldsymbol{Z}_k$ represents the design matrix associated to the k-th random component $\boldsymbol{c}_k$, with $\boldsymbol{c} = \left(\boldsymbol{c}_1^t, \ldots, \boldsymbol{c}_q^t\right)^t$, and that the variance-covariance $\boldsymbol{G} = \bigoplus_{k=1}^{q}\boldsymbol{G}_k$. As a consequence (see (B3))

$$\begin{aligned}\widehat{\boldsymbol{c}} = \begin{bmatrix}\widehat{\boldsymbol{c}}_1\\ \widehat{\boldsymbol{c}}_2\\ \vdots\\ \widehat{\boldsymbol{c}}_q\end{bmatrix} &= \begin{bmatrix}\boldsymbol{G}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{G}_2 & \cdots & \boldsymbol{0}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{G}_q\end{bmatrix}\begin{bmatrix}\boldsymbol{Z}_1^t\\ \boldsymbol{Z}_2^t\\ \vdots\\ \boldsymbol{Z}_q^t\end{bmatrix}\boldsymbol{Q}\boldsymbol{y}\\ &= \begin{bmatrix}\boldsymbol{G}_1\boldsymbol{Z}_1^t\boldsymbol{Q}\boldsymbol{y}\\ \boldsymbol{G}_2\boldsymbol{Z}_2^t\boldsymbol{Q}\boldsymbol{y}\\ \vdots\\ \boldsymbol{G}_q\boldsymbol{Z}_q^t\boldsymbol{Q}\boldsymbol{y},\end{bmatrix}.\end{aligned}$$

This result implies that

$$\boldsymbol{H}_R\boldsymbol{y} = \boldsymbol{Z}\widehat{\boldsymbol{c}} = \sum_{k=1}^{q}\boldsymbol{Z}_k\widehat{\boldsymbol{c}}_k = \sum_{k=1}^{q}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{Z}_k^t\boldsymbol{Q}\boldsymbol{y} = \sum_{k=1}^{q}\boldsymbol{H}_k\boldsymbol{y},$$

where $\boldsymbol{H}_k = \boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{Z}_k^t\boldsymbol{Q}$. Note that $\boldsymbol{H}_k$ $(k = 1, \cdots, q)$ is the "hat" matrix corresponding to the k-th random component in our SpATS model (eqn. (6) of the main manuscript), i.e.,

$$\widehat{f}_v(\boldsymbol{v}) = \boldsymbol{Z}_1\widehat{\boldsymbol{c}}_1 = \boldsymbol{H}_1\boldsymbol{y} \qquad \widehat{f}_u(\boldsymbol{u}) = \boldsymbol{Z}_2\widehat{\boldsymbol{c}}_2 = \boldsymbol{H}_2\boldsymbol{y} \qquad \boldsymbol{u}\odot\widehat{h}_v(\boldsymbol{v}) = \boldsymbol{Z}_3\widehat{\boldsymbol{c}}_3 = \boldsymbol{H}_3\boldsymbol{y},$$

$$\boldsymbol{v}\odot\widehat{h}_u(\boldsymbol{u}) = \boldsymbol{Z}_4\widehat{\boldsymbol{c}}_4 = \boldsymbol{H}_4\boldsymbol{y} \qquad \widehat{f}_{u,v}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{Z}_5\widehat{\boldsymbol{c}}_5 = \boldsymbol{H}_5\boldsymbol{y}, \qquad \boldsymbol{Z}_k\widehat{\boldsymbol{c}}_k = \boldsymbol{H}_k\boldsymbol{y} \;\; (k = 6, \ldots, q).$$

Accordingly, the hat matrix associated to the random part of the SpATS model can be decomposed as a sum of independent hat matrices, each related to a specific random component, i.e.,

$$\boldsymbol{H}_R = \sum_{k=1}^{q}\boldsymbol{H}_k.$$

Finally, it can also be shown (see, e.g., eqn. (9d) and Appendix 1 in Johnson and Thompson, 1995) that

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{c}} = \boldsymbol{R}\boldsymbol{Q}\boldsymbol{y}, \tag{C10}$$

and the residuals' hat matrix is thus $\boldsymbol{H}_{\boldsymbol{\varepsilon}} = \boldsymbol{R}\boldsymbol{Q}$.

Web Appendix D Equivalence between effective dimensions definitions

This section presents the equivalence between the definition given by Cui et al. (2010) of the effective dimension associated to a model's component and ED_k as defined in (B6). Specifically, Cui et al. (2010) define the effective dimension of a random component $\boldsymbol{c}_k$ as

$$\begin{aligned} \text{ED}\left(\boldsymbol{Z}_k\right) &= \lim_{\varsigma \to +\infty} \text{trace}\left(\boldsymbol{Z}_k \boldsymbol{G}_k \boldsymbol{Z}_k^t \left(\boldsymbol{V} + \varsigma \boldsymbol{X} \boldsymbol{W} \boldsymbol{X}^t\right)^+\right) && \text{(D11)} \\ &= \text{trace}\left(\boldsymbol{Z}_k \boldsymbol{G}_k \boldsymbol{Z}_k^t \left[\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right) \boldsymbol{V} \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right)\right]^+\right), && \text{(D12)} \end{aligned}$$

where $\boldsymbol{\Gamma}^+$ denotes the Moore-Penrose pseudoinverse of $\boldsymbol{\Gamma}$, $\boldsymbol{W}$ is a positive definite matrix, ς is a positive scalar, and $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t$. The expression given above defines the effective dimension of a model's component as the trace of the ratio of that "component's modelled variance matrix" ($\boldsymbol{Z}_k \boldsymbol{G}_k \boldsymbol{Z}_k^t$) to "total variance matrix" ($\boldsymbol{V} + \varsigma \boldsymbol{X} \boldsymbol{W} \boldsymbol{X}^t$). Note that this definition treats $\boldsymbol{\beta}$ as a random vector with variance-covariance $\varsigma \boldsymbol{W}$. However, as pointed out by the authors, a fixed effect can be viewed as the limiting case of a random effect for which the variance-covariance matrix goes to infinity (i.e., when $\varsigma \to +\infty$). Similarly to $\text{ED}\left(\boldsymbol{Z}_k\right)$, Cui et al. (2010) define the effective dimension for the error term as

$$\begin{aligned} \text{ED}\left(\boldsymbol{\varepsilon}\right) &= \lim_{\varsigma \to +\infty} \text{trace}\left(\boldsymbol{R}\left(\boldsymbol{V} + \varsigma \boldsymbol{X} \boldsymbol{W} \boldsymbol{X}^t\right)^+\right) && \text{(D13)} \\ &= \text{trace}\left(\boldsymbol{R}\left[\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right) \boldsymbol{V} \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right)\right]^+\right), \end{aligned}$$

and show that

$$\begin{aligned} n &= \text{ED}(\boldsymbol{X}) + \sum_{k=1}^{q} \text{ED}(\boldsymbol{Z}_k) + \text{ED}\left(\boldsymbol{\varepsilon}\right) \\ &= \text{rank}\left(\boldsymbol{X}\right) + \sum_{k=1}^{q} \text{ED}(\boldsymbol{Z}_k) + \text{ED}\left(\boldsymbol{\varepsilon}\right). \end{aligned}$$

The definitions given in Cui et al. (2010) thus partition n (the number of observations) into independent effective dimensions for the model's components and error. In authors' words, this result jointly with (D11) and (D13), suggests interpreting the effective dimension of a model's component $\text{ED}\left(\boldsymbol{Z}_k\right)$ as the fraction of response variation attributed to that individual effect, and the same applies to the error term. Besides, it allows explaining how components compete with one another to explain that variation.

To show that $\text{ED}_k = \text{ED}\left(\boldsymbol{Z}_k\right)$, we use results derived in the paper by Hoog et al. (1990). Given that (see identity (1) in Hoog et al., 1990)

$$\boldsymbol{Q} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1} = \left[\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right)\boldsymbol{V}\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right)\right]^+,$$

we have

$$\begin{aligned}
\text{ED}_k &= \text{trace}\left(\boldsymbol{Z}_k^t \boldsymbol{Q} \boldsymbol{Z}_k \boldsymbol{G}\right) \\
&= \text{trace}\left(\boldsymbol{Z}_k \boldsymbol{G} \boldsymbol{Z}_k^t \boldsymbol{Q}\right) \\
&= \text{trace}\left(\boldsymbol{Z}_k \boldsymbol{G} \boldsymbol{Z}_k^t \left[\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right) \boldsymbol{V} \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right)\right]^+\right) \\
&= \text{ED}\left(\boldsymbol{Z}_k\right).
\end{aligned}$$

We would like to note that the equivalence between (D11) and (D12) can also be proved using results of Hoog et al. (1990). Furthermore, we also have that

$$\begin{aligned}
\text{ED}_{\boldsymbol{\varepsilon}} &= \text{trace}\left(\boldsymbol{R} \boldsymbol{Q}\right) \\
&= \text{trace}\left(\boldsymbol{R} \left[\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right) \boldsymbol{V} \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right)\right]^+\right) \\
&= \text{ED}\left(\boldsymbol{\varepsilon}\right),
\end{aligned}$$

where $\text{ED}_{\boldsymbol{\varepsilon}}$ has been presented in eqns. (B7) and (B8).

Web Appendix E Barley uniformity trial revisited

This section revisits the uniformity barley data discussed in Section 2 of the main manuscript. Web Appendix E.1 presents and discusses the results of the analysis of this field trial in terms of effective dimensions, and in Web Appendix E.2, the trial is used to introduce the proposal by Gilmour et al. (1997), and comparisons between both approaches when including genotypic effects are reported.

Web Appendix E.1 Results in terms of effective dimensions

For fitting model (2) of the main manuscript, we used cubic B-spline bases of dimension $\hat{m} = 18$ and $\breve{m} = 51$, jointly with a nested basis for the columns with $\breve{m}_N = 27$. Table 1 shows the model (i.e., the number of coefficients), and effective dimensions associated to the row and column random factors, the smooth spatial field $f(u, v)$, and each of the PS-ANOVA components. If we focus on the random effects for the rows and columns, we have that the estimated effective dimensions are 5.5 (out of 14) and 33.6 (out of 47) respectively. This result suggests that the column effect is stronger than the row effect, and that these two components are probably needed. For the PS-ANOVA spatial field (excluding the intercept), the total effective dimension is 79.8, with the smooth-by-smooth interaction trend being responsible for the strongest contribution, with an effective dimension of 53.3. As can be observed on the graphical results depicted in Section 2 of the main manuscript, the smooth trend along the rows, $f_u(u)$, is more complex (or rougher) than the one along the columns, $f_v(v)$, and this fact is made evident on the effective dimension related to each of these components, 6.2 and 4.6 respectively. Besides, as could have also been expected, the effective dimension associated to $vh_u(u)$ is also larger than that associated $uh_v(v)$ (8.2

Web Table 1: Model and effective dimension of the smooth spatial component, and the ANOVA-type decomposition components for the barley uniformity trial.

Dimensions	Random		Spatial components Smooth					
	$\boldsymbol{c}_r$	$\boldsymbol{c}_c$	Global - $f(u,v)$	$f_u(u)$	$f_v(v)$	$vh_u(u)$	$uh_v(v)$	$f_{u,v}(u,v)$
Model (m_k)	48	15	533	16	49	16	49	400
Effective (ED_k)	33.6	5.5	79.8	6.2	4.6	8.2	4.5	53.3

and 4.5 respectively). These results suggest and emphasise the need of modelling spatial trends by means of bivariate surfaces. Here the additive assumption – only based on main smooth effects – would have not been flexible enough to recover the spatial trend variation present in the data.

Web Appendix E.2 Simulation study

The barley uniformity trial was also analysed using the AR×AR model proposed by Gilmour et al. (1997). Model selection was performed by means of the sample variogram and plots of the residuals as suggested by Stefanova et al. (2009). Starting with the simplest model, including only the separable Gaussian AR process of order 1, we further evaluated the need of extra model components. In total, 9 different models were considered, with the best being

$$\boldsymbol{y} = \mathbf{1}_n\beta_0 + \boldsymbol{u}\beta_1 + \boldsymbol{v}\beta_2 + \boldsymbol{x}_d\beta_d + f_u(\boldsymbol{u}) + f_v(\boldsymbol{v}) + \boldsymbol{Z}_c\boldsymbol{c}_c + \boldsymbol{\xi} + \boldsymbol{\varepsilon}. \tag{E14}$$

Here $\boldsymbol{\xi}$ is a (720×1) spatially dependent random vector, for which a separable Gaussian AR process of order 1 in the row and column directions is assumed. Accordingly, $cov\left(\xi_l, \xi_p\right) = \sigma_s^2\rho_r^{|u_l-u_p|}\rho_c^{|v_l-v_p|}$, where ρ_r and ρ_c are the autocorrelation parameters for row and column, respectively. As in our approach, $f_u\left(\cdot\right)$ and $f_v\left(\cdot\right)$ represent smooth-trend functions over the row and column direction respectively. We note that in the geostatistics literature, $\boldsymbol{\varepsilon}$ is usually referred to as measurement error or nugget effect.

As postulated by Gilmour et al. (1997), model (E14) accounts for three sources of spatial variation: the global trend variation, the local trend variation; and the so-called extraneous variation. Here, (a) the global trend variation is modelled by the linear effect along the rows (β_1) and columns (β_2) as well as by the smooth-effect functions $f_u\left(\cdot\right)$ and $f_v\left(\cdot\right)$; and (b) the local trend variation by means of the spatially dependent random error $\boldsymbol{\xi}$. The extraneous variation, related to the experimental procedure, is accounted for by the column random factor $\boldsymbol{c}_c$ and the correction for the three-column cycle pattern β_d (discussed in the main manuscript). Under this framework, we can see our SpATS model as that based on aggregating both the local and global trend variation in one component, and

modelling it by means of a smooth bivariate surface. Table 2 shows the REML estimates of the variance components based on model (2) of the main manuscript and model (E14). Based on this table, we find it difficult to compare both approaches. Thus, to gain more insights in the performance of these two models, we designed a simulation study in which, on top of the uniformity data, genotypic effects were included, i.e,

$$\boldsymbol{y}^* = \boldsymbol{y} + \boldsymbol{Z}_g \boldsymbol{c}_g,$$

where $\boldsymbol{c}_g$ denotes the genotypic effects, with $\boldsymbol{c}_g \sim N(\boldsymbol{0}, \sigma^2_g \boldsymbol{I}_{m_g})$. For the results reported here, we considered $\sigma^2_g = 144$, and a total of $m_g = 360$ genotypes, each replicated twice. The 360 genotypes were allocated to the plots following an alpha design, in blocks of size 15 (the number of rows in the field).

For each data set simulated as described above, we fitted our SpATS model (see eqn. (2) of the main manuscript), including the genetic random factor. For Gilmour et al. (1997)'s approach we considered model (E14) (with and without the nugget) plus the genetic random factor. For comparison purposes, we also fitted a model only with the correction for rows, columns and the three-column cycle pattern (eqn. (1) of the main manuscript). The procedure was repeated a total of $R = 500$ times. For SpATS, we used cubic B-spline bases of dimension $\hat{m} = 18$ and $\breve{m} = 51$, jointly with a nested basis for the columns with $\breve{m}_N = 27$. As for the simulation reported in Section 5 of the main manuscript, models' performance was measured in terms of the RMSE (for the genotypic BLUPs), and the bias for the REML estimates of the genetic variance σ^2_g.

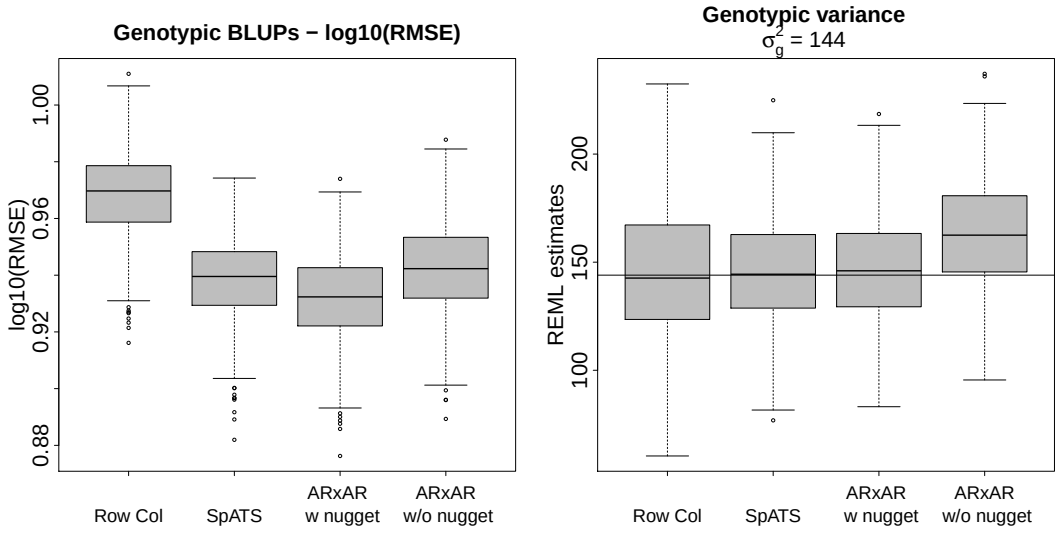
Web Figure 2 shows the boxplots of log10(RMSE) associated to the genotypic BLUPs and the REML estimates of σ^2_g, for each of the four models considered. In terms of the log10(RMSE), and as could have been expected, the worst performance corresponds to the model including only the correction for rows and columns. The remaining three models present a similar behaviour, with the best approach being the AR×AR model including the nugget, followed by our SpATS model. However, if we focus on the REML estimates of σ^2_g, we observe that the AR×AR model excluding the nugget tends to overestimate the genetic signal. This behaviour can explain the larger heritability (on average) provided by this model in comparison with SpATS or the AR×AR model including the nugget (see Table 3). Surprisingly, the model including only the correction for rows and columns provides (on average) good estimates of σ^2_g. However, it also produces the lowest heritabilities. Note that this is the expected behaviour, as the variance associated to the measurement error will be inflated in this case.

Web Appendix F `SpATS` package

This section contains a brief description of the developed R-package associated to this paper. The package can be freely downloaded from https://cran.r-project.org/package=SpATS, where a more detailed depiction of it use can be found. The main function of the package

Web Table 2: REML estimates of the variance parameters for the barley uniformity trial based on both AR×AR and SpATS approaches.

Model	Parameter					
	σ_s^2	σ_r^2	σ_c^2	ρ_r	ρ_c	σ^2
AR×AR	265.94	-	78.13	0.383	0.834	173.10
SpATS	-	20.38	80.64	-	-	238.80



Web Figure 2: For the barley uniformity data simulation study: Boxplots, based on 500 simulated data sets, of the log10(RMSE) associated to the genotypic random factor and the REML estimates of σ_g^2. "Row Col" corresponds to the model including only the correction for rows, columns and the three-column cycle, "SpATS" to our proposal, "AR×AR w nugget" and "AR×AR w/o nugget" to the AR×AR model with and without the nugget respectively.

Web Table 3: Numerical results associated to the simulation study based on the barley uniformity data. For the genotypic BLUPs, the $\log_{10}$(RMSE) is shown. For the REML estimates of σ_g^2 the results show the bias. In all cases, averages and standard deviations over 500 simulated data sets are presented. "Row Col" corresponds to the model including only the correction for rows, columns and the three-column cycle, "SpATS" to our proposal, "AR×AR w nugget" and "AR×AR w/o nugget" to the AR×AR model with and without the nugget respectively.

	Model			
	Row Col	SpATS	AR×AR w nugget	AR×AR w/o nugget
Genotypic RMSE	0.969 (0.016)	0.938 (0.016)	0.932 (0.016)	0.942 (0.017)
$\sigma_g^2 = 144$	0.618 (29.846)	1.717 (24.425)	2.167 (23.809)	19.179 (24.246)
H_g^2	0.405 (0.061)	0.488 (0.053)	0.502 (0.051)	0.541 (0.049)

is `SpATS()`, which fits the spatial model presented in Section 4 of the main manuscript. Numerical and graphical summaries of the fitted spatial model can be obtained, as usual in `R`, by using `summary.SpATS()`, `variogram.SpATS()`, `predict.SpATS()` and `plot.SpATS()`. In the implementation of the package, the sparse structure of the design matrix associated with the genotype has been taken into account, which, in combination with the estimation procedure presented in Web Appendix B and the possible use of nested B-spline bases, makes the package computational efficient, allowing the analysis of very large datasets.

By way of example, we present here the syntax for the Australian wheat trial example discussed in the paper by Gilmour et al. (1997). The aim of this trial was the evaluation of advanced breeding lines and commercial varieties. The trial consisted of 107 varieties, which were sown in three replicates, each replicate being a complete block. Each block comprised 5 columns and 22 rows, yielding a total of 330 ($5 \times 22 \times 3$) plots on the field. To meet the 110 plots per replicate, from the 107 varieties, three were sown twice in each of these. For more details about the trial, we refer the readers to the cited paper. On the basis of the results shown in Gilmour et al. (1997), the following statistical model was assumed

$$\boldsymbol{y} = \boldsymbol{X}_g \boldsymbol{\beta}_g + f(\boldsymbol{u}, \boldsymbol{v}) + \boldsymbol{Z}_r \boldsymbol{c}_r + \boldsymbol{Z}_c \boldsymbol{c}_c + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}_g$ is a (106×1) vector of fixed variety (genetic) effects, and $\boldsymbol{X}_g$ is the corresponding (330×106) design matrix. Note that the dimension of the genetic effect is $m_g - 1$ (where $m_g = 107$) since the intercept is included in $f(\boldsymbol{u}, \boldsymbol{v})$. Here, $\boldsymbol{c}_r \sim N\left(\boldsymbol{0}, \sigma_r^2 \boldsymbol{I}_{22}\right)$ and $\boldsymbol{c}_c \sim N\left(\boldsymbol{0}, \sigma_c^2 \boldsymbol{I}_{15}\right)$ are vectors of row and column random effects respectively, and $\boldsymbol{\varepsilon} \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{330}\right)$.

The dataset can be found in the `R`-package `agridat`, under the name `gilmour.serpentine`. Here there is a brief summary of the data

```
> library(agridat)
> GS <- gilmour.serpentine
> summary(GS)
      col              row             rep                  gen           yield
 Min.   : 1   Min.   : 1.0   R1:110   TINCURRIN   :  6   Min.   :194.0
 1st Qu.: 4   1st Qu.: 6.0   R2:110   VF655       :  6   1st Qu.:469.0
 Median : 8   Median :11.5   R3:110   WW1477      :  6   Median :617.5
 Mean   : 8   Mean   :11.5            (WWH*MM)*WR*:  3   Mean   :591.8
 3rd Qu.:12   3rd Qu.:17.0            (WqKPWmH*3Ag:  3   3rd Qu.:713.5
 Max.   :15   Max.   :22.0            AMERY       :  3   Max.   :925.0
                                      (Other)     :303
```

The dataset contains the column and row positions (`col` and `row` variables), the complete block (variable `rep`), the variety (`gen`) and the yield (`yield`). In order to incorporate in the model the random factors of rows and columns, we need first to create the corresponding factor variables, that we denote as `col_f` and `row_f`, and we then fit the model

```
>  GS$col_f = factor(GS$col)
>  GS$row_f = factor(GS$row)
```

```
> fit.SpATS <- SpATS(response = "yield", genotype = "gen", genotype.as.random = FALSE,
+ spatial = ~ PSANOVA(col, row, nseg = c(16,20), degree = 3, nest.div = 2),
+ fixed = NULL, random = ~ row_f + col_f,
+ data = GS, control =  list(tolerance = 1e-03, monitoring = 1))

Timings:
SpATS 0.38 seconds
All process 0.57 seconds
```

Through `response` and `genotype` arguments, users specify the name of the variables in the dataset that contains, respectively, the response variable (phenotype) of interest and the genotype or variety. The genotype can be included in the model either as fixed (default) or random (`genotype.as.random = TRUE`). For modelling the spatial trend, argument `spatial`, we consider 16 segments (`nseg`) for the column position and 20 for the row. This, jointly with the fact we use cubic B-splines, `degree = 3`, gives rise to B-spline bases of dimension $\check{m} = 16+3 = 19$ and $\hat{m} = 20+3 = 23$ for the columns and rows, respectively. By specifying the argument `nest.div = 2`, we indicate the use of nested bases, with half the number of segments of the original ones (see Web Appendix A). The fixed and random effects to be included in the model are indicated in `fixed` and `random`, and argument `control` allows to modify some default parameters that control the fitting process. For instance, the tolerance for the convergence criterion for the variance components can be altered using this argument, as well as the maximum number of iterations. Under this representation,

the model has a total of 322 coefficients, but it took less than 1 seconds to be fitted. A numerical summary of the fitted model can be obtained by calling the function `summary()`. By indicating the argument `which = "all"`, we obtain both the estimates of the variance components and the effective dimensions

```
> summary(fit.SpATS, which = "all")
[...]

Variance components:
                   Variance            SD     log10(lambda)
row_f             4.397e+02     2.097e+01           0.67320
col_f             4.442e+03     6.665e+01          -0.33128
f(col)            1.245e+04     1.116e+02          -0.77895
f(row)            7.240e+01     8.509e+00           1.45657
f(col):row        7.847e+02     2.801e+01           0.42159
col:f(row)        6.490e-06     2.548e-03           8.50408
f(col):f(row)     2.530e+03     5.030e+01          -0.08684

Residual          2.072e+03     4.552e+01

Dimensions:
                  Effective     Model     Nominal     Ratio     Type
gen                   106.0       106         106      1.00        F
Intercept               1.0         1           1      1.00        F
row_f                  12.6        22          21      0.60        R
col_f                  10.3        15          14      0.74        R
col                     1.0         1           1      1.00        S
row                     1.0         1           1      1.00        S
row:col                 1.0         1           1      1.00        S
f(col)                  2.3        17          17      0.14        S
f(row)                  1.0        21          21      0.05        S
f(col):row              2.6        17          17      0.15        S
col:f(row)              0.0        21          21      0.00        S
f(col):f(row)           7.5        99          99      0.08        S

Total                 146.3       322         320      0.46
Residual              183.7
Nobs                    330

Type codes: F 'Fixed'    R 'Random'    S 'Smooth/Semiparametric'
```

In this example, there are seven variance components, five associated to the spatial trend,

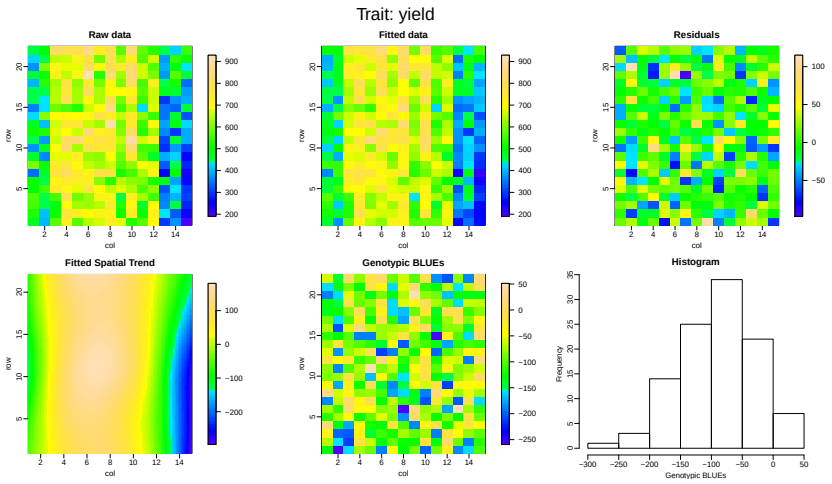
one associated to the row random effects, one to the column random effects, and the residual variance σ^2. The column `log10(lambda)` shows the logarithm of base 10 of the smoothing parameters, i.e., the ratio between the residual variance and the variance component. As far as the dimensions are concerned, for each component in the model (either fixed, random or spatial), the function returns (a) the `effective` dimension or effective degrees of freedom, (b) the `model` dimension, i.e., the number of parameters to be estimated, (c) the `nominal` dimension, which, for the random components is the model dimension minus one, lost due to the constraint of zero-mean imposed to them; and (d) the `ratio` between the `effective` and the `nominal` dimension. It is worth remembering that, if the variety had been included as random, the `ratio` for the variety would have provided an estimate of the so-called generalised heritability as proposed by Oakey et al. (2006). If we focus on the random effects for the rows and columns, we have that the effective dimensions are, respectively, about 60% and 74% of the nominal dimensions. As discussed in Section 2 of the main manuscript, for the spatial trend we have, in total, 8 components (excluding the intercept). The linear effects for the rows and the columns (`row` and `column`), as well as the linear interaction (`row:col`), represent the fixed or unpenalised part of the tensor-product P-spline. The remaining five components, i.e., the main effects (`f(row)` and `f(col)`), the smooth varying coefficient terms (`f(col):row` and `row:f(col)`); and the smooth-by-smooth interaction component (`f(col):f(row)`) correspond to the penalised or random part, and have been extensively discussed in Sections 2 and 4.2 of the main manuscript. On the basis of the effective dimensions associated to each of these five components, we may infer that most of the trend has been captured by the main effect and the smooth varying coefficient term along the column position, but also by the smooth-by-smooth interaction term, for which we have an effective dimension of 7.6.

To complement those numerical results, the `SpATS` package furnishes different graphical results that can be used to further explore the fitted model. Specifically, the sample variogram can be obtained using the function `variogram()`, which can also be plotted; and the function `plot()` depicts six different graphics: the raw data, the fitted values, the residuals, the estimated spatial trend (excluding the intercept), the genotypic BLUEs (or BLUPs) and their histogram. Except for the histogram, the plots are depicted in terms of the spatial coordinates (e.g., the rows and columns of the field).

```
> plot(fit.SpATS)

> plot(variogram(fit.SpATS))
```

The result of the above code is shown in Web Figures 3 and 4. The spatial plots of the residuals and the genotypic BLUEs ($\widehat{\boldsymbol{\beta}}_g$) do not suggest the presence of any extra spatial pattern that should have been taken into account. A similar conclusion can be drawn from the sample variogram of the residuals shown in Web Figure 4. Finally, note that the fitted spatial trend takes values between -300 and 200, whereas the residuals vary between -100



Web Figure 3: Graphical results provided by the `SpATS` package for the Australian wheat trial.

and 100. This result highlights what could have been expected based on the raw data, that spatial (plot-to-plot) variation is larger than random (plot-to-plot) variation.

Acknowledgements

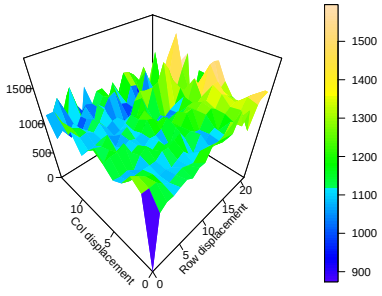
This research was supported by the Spanish Ministry of Economy and Competitiveness MINECO grant MTM2014-55966-P and BCAM Severo Ochoa excellence accreditation SEV-2013-0323, and by the Basque Government through the BERC 360 2014-2017. We thank SESVanderHave for providing the sugar beet data. We are grateful to Cajo ter Braak, María Durbán, Dae-Jin Lee and Julio Velazco for useful discussions.

References

Cui, Y., J. S. Hodges, X. Kong, and B. P. Carlin (2010). Partitioning degrees of freedom in hierarchical and other richly-parameterized models. *Technometrics 52*, 124–136.

Engel, B. (1990). The analysis of unbalanced linear models with variance components. *Statistica Neerlandica 44*, 195–219.

Gilmour, A. R., B. R. Cullis, and A. P. Verbyla (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological and Environmental Statistics 2*, 269–293.



Web Figure 4: Sample variogram of the residuals provided by the SpATS package for the Australian wheat trial.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association 72*(358), 320–338.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistical Genetics and Plant Breeding 982*, 141–163.

Hoog, F. R., T. P. Speed, and E. R. Willians (1990). On a matrix identity associated with generalized least squares. *Linear Algebra and its Applications 127*, 449–456.

Johnson, D. L. and R. Thompson (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of Dairy Science 78*, 449–456.

Lee, D.-J., M. Durban, and P. H. C. Eilers (2013). Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested basis. *Computational Statistics and Data Analysis 61*, 22–37.

Oakey, H., A. Verbyla, W. Pitchford, B. Cullis, and H. Kuchel (2006). Joint modeling of additive and non-additive genetic line effects in single field trials. *Theoretical and Applied Genetics 113*(5), 809–819.

Rodríguez-Álvarez, M. X., D.-J. Lee, T. Kneib, M. Durban, and P. H. C. Eilers (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the sap algorithm. *Statistics and Computing 25*, 941–957.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Schnabel, S. K. and P. H. C. Eilers (2009). Optimal expectile smoothing. *Computational Statistics and Data Analysis 52*, 4168–4177.

Stefanova, K. T., A. B. Smith, and B. R. Cullis (2009). Enhanced diagnostics for the spatial analysis of field trials. *Journal of Agricultural, Biological, and Environmental Statistics 14*(4), 392–410.