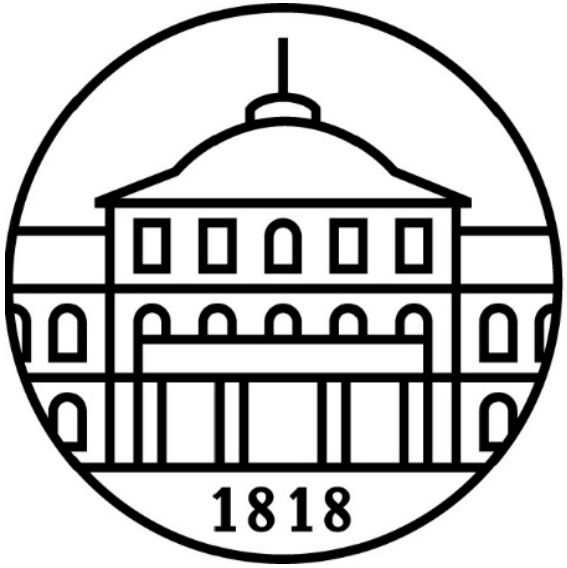
Two-dimensional P-spline smoothing for spatial analysis of field trials

Hans-Peter Piepho

Biostatistics Unit

Institute of Crop Science

Universität Hohenheim



This is joint work with:

Martin Boer

Wageningen University & Research (WUR), Wageningen, NL

Emlyn Williams

Australian National University (ANU), Canberra, AUS

Spatial models

Figure: A single row of plots. Arrow indicates direction of spatial correlation.

D.R. Cox on merits and demerits of spatial design

- "Yates (1939) at the conclusion of the discussion (*between Student and Fisher on the merits or randomized vs systematic designs*), suggested that, while there might sometimes be small gains in precision to be achieved by systematic arrangements, the lack of security in the basis for error estimation in such designs detracted attention from key issues of interpreting the effects under study.

- More recent work in a similar vein, stemming from Bartlett (1978) and Wilkinson et al. (1983) has been based on explicit time series or spatial models of variability, often leading to the so-called neighbourhood balance designs. Again, however, the reality of any apparent gain in precision depends on the adequacy of the assumed model."

Cox, D.R. (2009). International Statistical Review 77, 415-429.

One very common misconception regarding spatial analysis

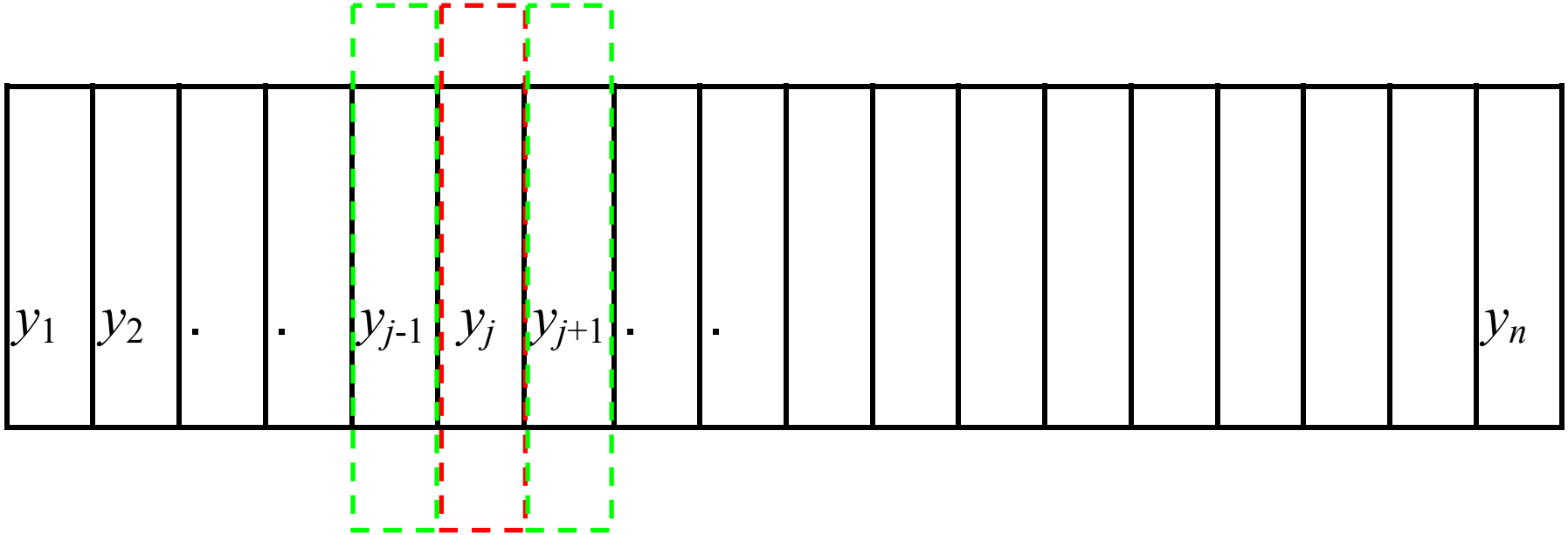
(1) The design was a randomized complete block design

(2) Analysis of variance assumes that errors within blocks are uncorrelated

(3) Inspection of real data reveals, however, that this assumption is incorrect

(4) Thus, it is incorrect to use analysis of variance, and it is better to use a spatial model.

Nearest neighbour analysis (NNA) (Papadakis 1937)



y_j = yield on *j*-th plot

Second differences: $$y_j - \frac{y_{j-1} + y_{j+1}}{2}$$

ΕΠΙΣΤ. ΔΕΛΤΙΟΝ ΑΡ. 23

ΣΤΑΤΙΣΤΙΚΗ ΜΕΘΟΔΟΣ ΔΙΑ ΠΕΙΡΑΜΑΤΑ ΕΠΙ ΤΟΥ ΑΓΡΟΥ

ΥΠΟ

Ι. Σ. ΠΑΠΑΔΑΚΗ Α. Ι. Gx

ΔΙΕΥΘΥΝΤΟΥ ΤΟΥ ΙΝΣΤΙΤΟΥΤΟΥ

INSTITUT D'AMÉLIORATION DES PLANTES A SALONIQUE (GRÈCE)
BULLETIN SCIENTIFIQUE Nº 23

MÉTHODE STATISTIQUE POUR DES EXPÉRIENCES SUR CHAMP

PAR J. S. PAPADAKIS A. I. Gx
DIRECTEUR DE L'INSTITUT

(Papadakis 1937)

Software

- ANOFT (Erik Schwarzbach, Czech Republic)
- AGROBASE (Dieter K. Mulitze, Agromix Software Inc., Canada)
- Self-made implementations (Excel etc.)

$\Rightarrow$ quite commonly used in plant breeding

$\Rightarrow$ limited to single-trial analysis

$\Rightarrow$ no model, can't extend

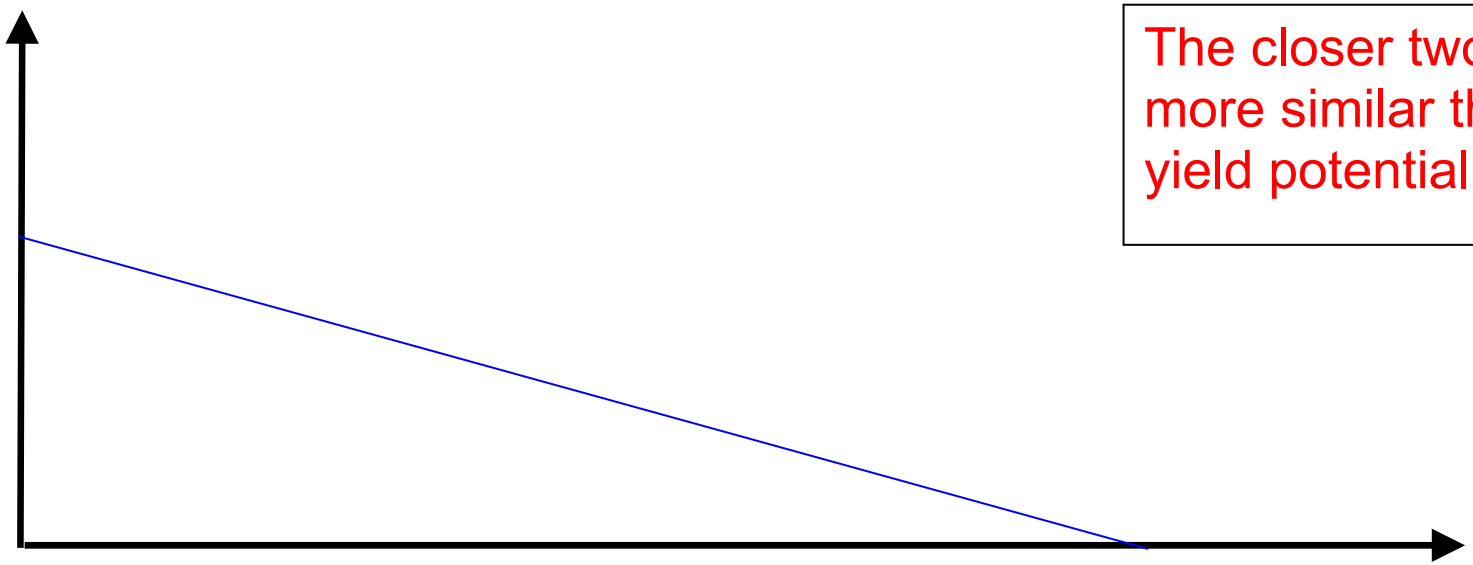
(pedigree/kinship information, genotype-environment interaction)

(Piepho et al. 2008)

An equivalent mixed model (spatial)

Linear Variance (LV) model (Williams 1986)

Covariance ($\Rightarrow$ similarity of plots)



The closer two plots, the more similar they are in yield potential

$\Rightarrow$ Can implement NNA with mixed models

Random walk models (state-space models)

t_1, t_2 = trend values of two adjacent plots

First-order random walk ⇒ first differences:

$$t_1 - t_2 = a_2 \Leftrightarrow t_2 = t_1 + a_2 \quad ; \quad a_2 \sim N\left(0, \sigma_a^2\right)$$ ⇒ equivalent to LV!

Second-order random walk ⇒ second differences:

$$\left(t_1 - t_2\right) - \left(t_2 - t_3\right) = t_1 - 2t_2 + t_3 = a_2 \Leftrightarrow t_3 = 2t_2 - t_1 + a_2 \quad ; \quad a_2 \sim N\left(0, \sigma_a^2\right)$$

First-order autoregressive model [AR(1)]:

$$t_2 = \rho t_1 + a_2 \quad ; \quad a_2 \sim N\left(0, \sigma_a^2\right) \quad 0 < \rho < 1$$

(Piepho et al. 2008)

P-splines

(1) Use effects u_j to model trend values t_i for the plots

(2) Smear out the u_j over adjacent plots

(3) The u_j are regression coefficients for B-spline basis functions $B = \{b_{ij}\}$

Trend value for the *i*-th plot:

$$t_i = b_{i1}u_1 + b_{i2}u_2 + \ldots + b_{im}u_m \qquad ; \qquad b_{i1} + b_{i2} + \ldots + b_{im} = 1$$

P = penalized $\Rightarrow$ penalty on regression coefficients u_j
$\Rightarrow$ first differences or second differences!

(Rodriguez-Alvarez et al. 2018; 'SpATS' package in R)

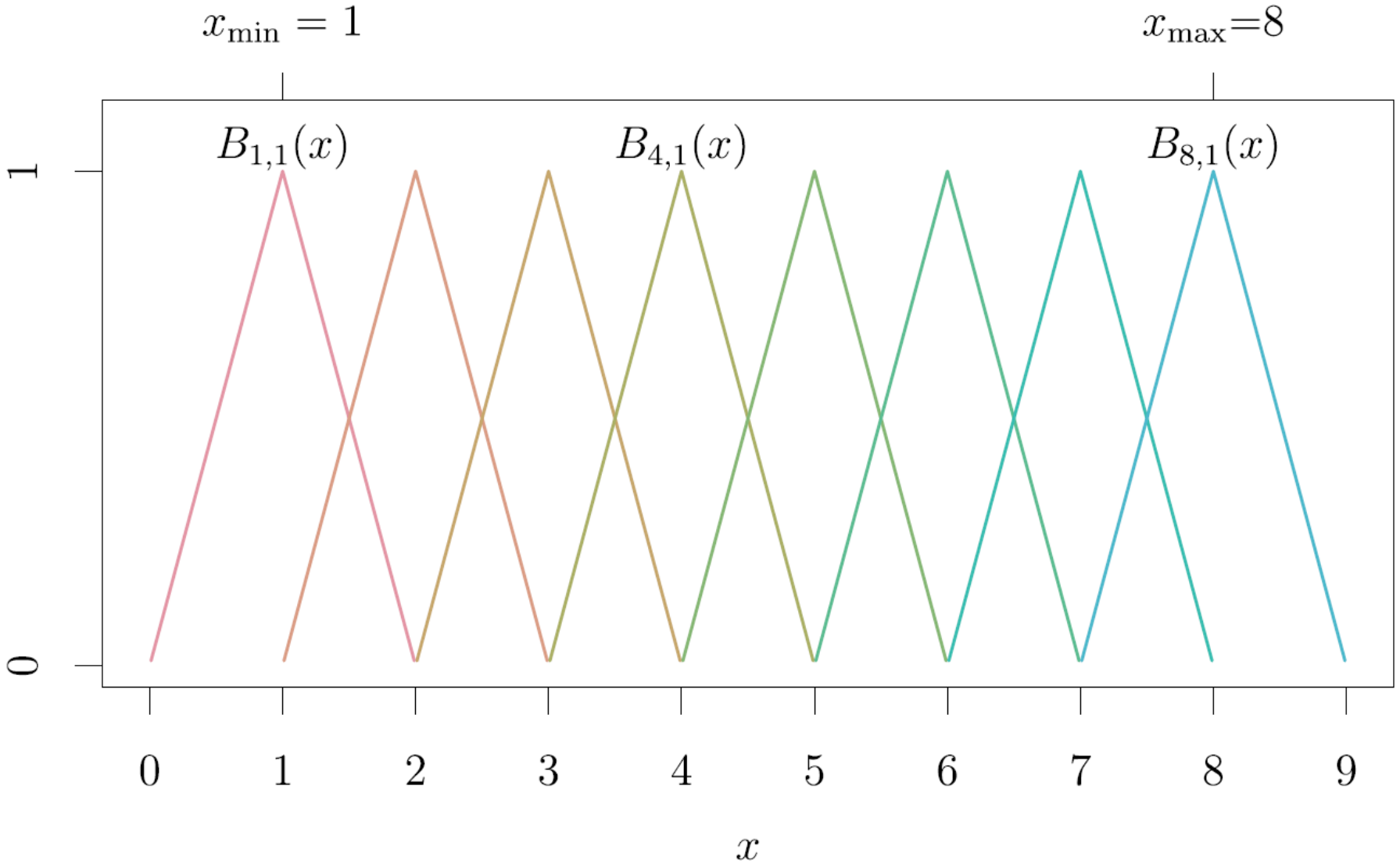


Figure 1. Example of a first-degree B-spline basis for a continuous coordinate x.

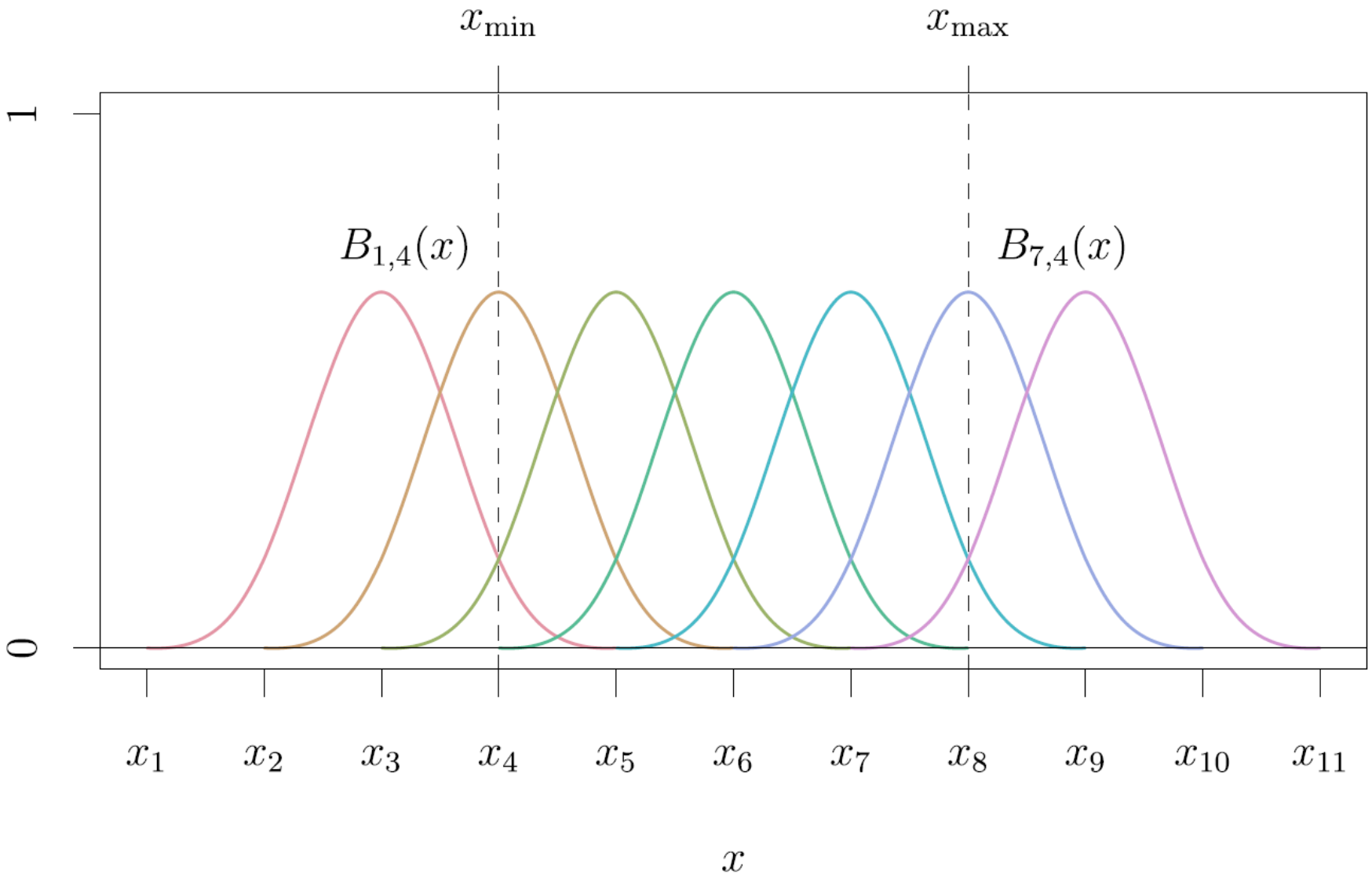


Figure 2. Cubical B-spline basis for a continuous coordinate x.

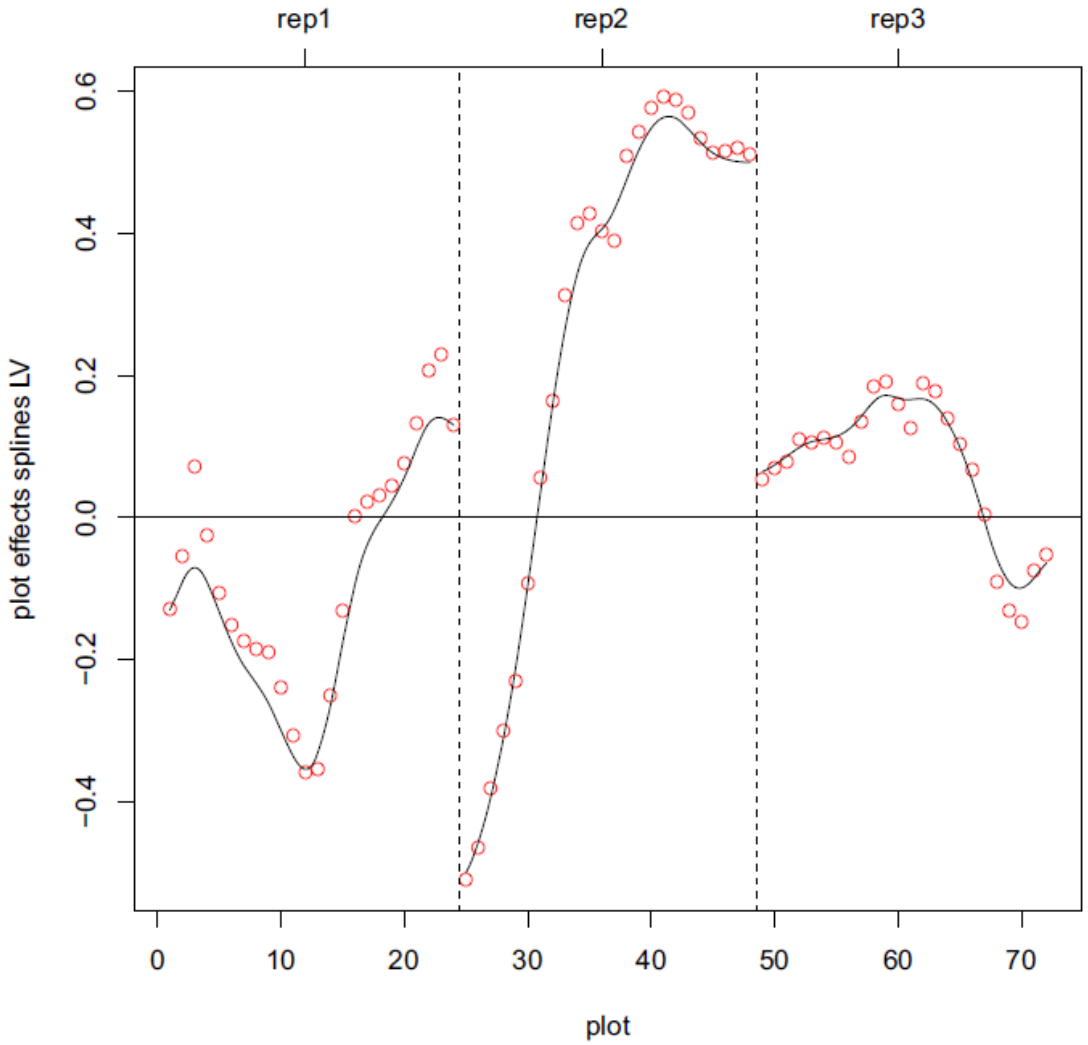


Figure 4. Comparison of plot effects u for the oats data. The red points are the estimates for the LV model and equivalent to first-order P-splines. The black curve is based on P-splines, using a third degree B-splines with 12 segments, so half the number of plots per replicates.

Trend LV model

Trend 3rd degree P-spline

(Boer et al., 2020)

Examples for 5 plots

Cubical = third-degree, 7 knots at plots (including "outer" plots!):

$$B = \begin{pmatrix} 0.167 & 0.666 & 0.167 & 0 & 0 & 0 & 0 \\ 0 & 0.167 & 0.666 & 0.167 & 0 & 0 & 0 \\ 0 & 0 & 0.167 & 0.666 & 0.167 & 0 & 0 \\ 0 & 0 & 0 & 0.167 & 0.666 & 0.167 & 0 \\ 0 & 0 & 0 & 0 & 0.167 & 0.666 & 0.167 \end{pmatrix}$$

$\Rightarrow$ need $u_1, \ldots, u_7$

$\Rightarrow t_1 = 0.167 \times u_1 + 0.666 \times u_2 + 0.167 \times u_3 + 0 \times u_4 + 0 \times u_5 + 0 \times u_6 + 0 \times u_7$

$\Rightarrow t_2 = 0 \times u_1 + 0.167 \times u_2 + 0.666 \times u_3 + 0.167 \times u_4 + 0 \times u_5 + 0 \times u_6 + 0 \times u_7$

etc.

Examples for 5 plots

Cubical = third-degree, just 5 knots:

$$B = \begin{pmatrix} 0.167 & 0.666 & 0.167 & 0 & 0 \\ 0.021 & 0.479 & 0.479 & 0.021 & 0 \\ 0 & 0.167 & 0.666 & 0.167 & 0 \\ 0 & 0.021 & 0.479 & 0.479 & 0.021 \\ 0 & 0 & 0.167 & 0.666 & 0.167 \end{pmatrix}$$

$\Rightarrow$ need u_1, ..., u_5

$\Rightarrow t_1 = 0.167 \times u_1 + 0.666 \times u_2 + 0.167 \times u_3 + 0 \times u_4 + 0 \times u_5$

$\Rightarrow t_2 = 0.021 \times u_1 + 0.479 \times u_2 + 0.479 \times u_3 + 0.021 \times u_4 + 0 \times u_5$

etc.

Examples for 5 plots

First-degree, just 3 knots:

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}$$

$\Rightarrow$ need u_1, ..., u_3

$\Rightarrow t_1 = 1 \times u_1 + 0 \times u_2 + 0 \times u_3$

$\Rightarrow t_2 = 0.5 \times u_1 + 0.5 \times u_2 + 0 \times u_3$

etc.

Examples for 5 plots

First-degree, 5 knots at plots:

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$\Rightarrow$ need $u_1, \ldots, u_5$

$\Rightarrow t_1 = 1 \times u_1 + 0 \times u_2 + 0 \times u_3 + 0 \times u_4 + 0 \times u_5 = u_1$

$\Rightarrow t_2 = 0 \times u_1 + 1 \times u_2 + 0 \times u_3 + 0 \times u_4 + 0 \times u_5 = u_2$

etc.

$\Rightarrow$ This is LV when first differences are used to penalize u_i!

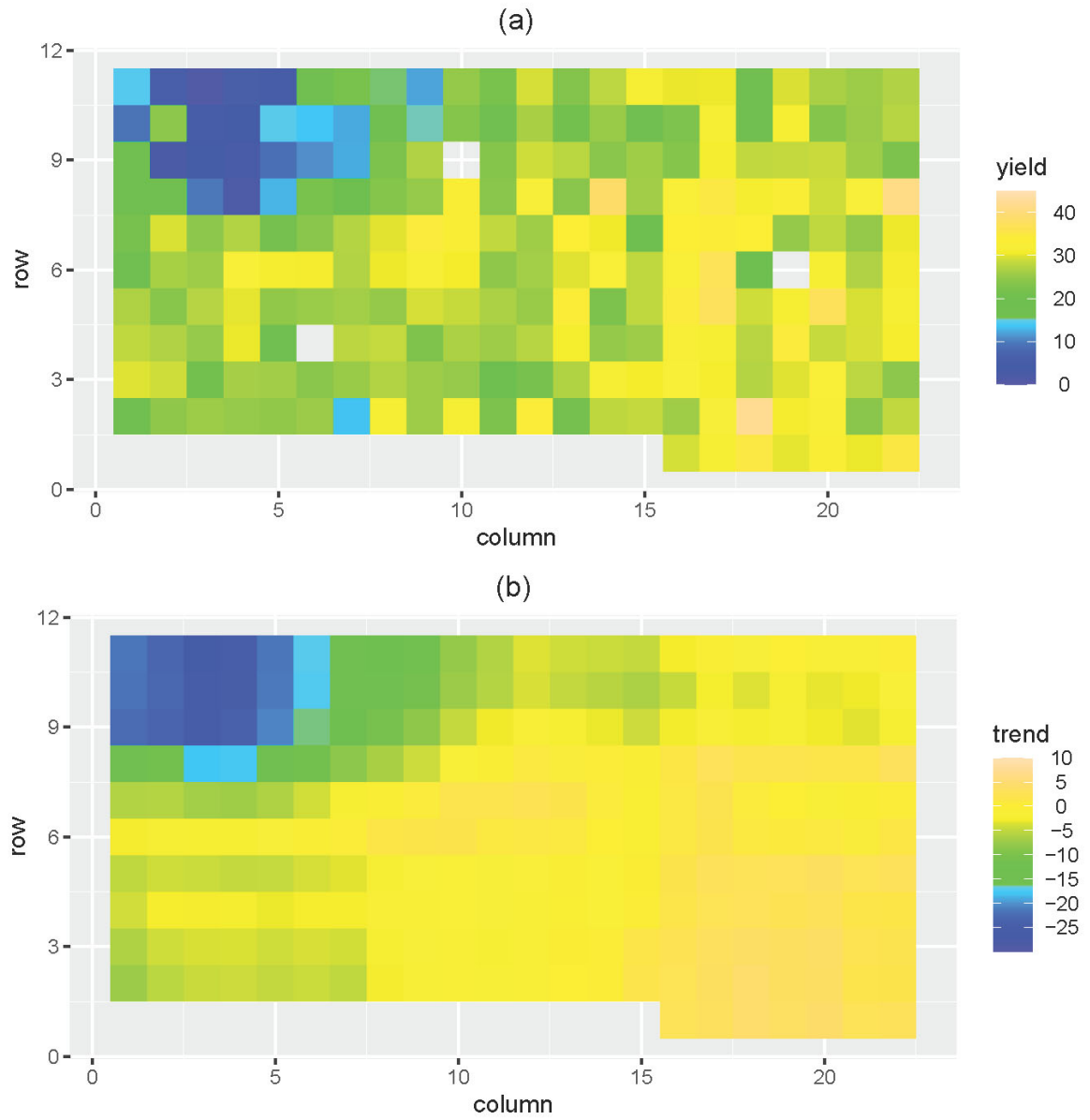


Figure Heatmap of wheat data of Stroup et al. (1994).

(a) Raw data

(b) Smooth trend (P-spline)

(Piepho et al., 2022)

Intermediate summary on P-splines

- Modelling smooth trend
- Just another spatial covariance model
- All covariance structures linear in the variance components
- LV and random walk are special cases when using first differences
- Many knobs (degree of B-spline basis, number of knots, difference penalty)
- SpATS uses second differences, but often first differences sufficient

(Boer et al. 2020; Piepho et al. 2022)

Fitting P-splines as mixed models: one column of plots

$$t = Bu$$

t = vector of trend values for k plots

B = $\{b_{ij}\}$ = $k \times m$ matrix of m B-spline bases of q-th degree

u = m-vector of coefficients

Penalty for coefficients u:

$$\theta u^T D^T D u$$

θ = penalty parameter

D = $(m-p) \times m$ matrix of p-th differences

The penalty

$\theta u^T P u$

$P = D^T D$ = penalty matrix

$\Rightarrow$ equivalent to quadratic form for random effect in likelihood for random effects u

$\sigma^{-2} u^T P u$

where

σ^2 = variance

P = precision matrix

Next question

What is the variance-covariance matrix for u?

$\Rightarrow$ needed for mixed model package

If P were positive-definite, we could just use $\mathrm{var}(u) = P^{-1}\sigma^2$

But P is singular!

$\Rightarrow$ use spectral decomposition of P

Spectral decomposition of P

$$P = U\ diag(d)\, U^T$$

U = eigenvectors of $P = (U_+ \mid U_0)$

d = eigenvalues of P

where

U_+ = $(m - p)$ eigenvectors corresponding to d_+

U_0 = p eigenvectors corresponding to zero eigenvalues

d_+ = subvector of $(m - p)$ positive eigenvalues in d

Decomposing $\boldsymbol{Bu}$

$$Bu = BUU^T u = BU_0 U_0^T u + BU_+ U_+^T u = X\beta + Zw$$

where

$X = BU_0$ and $\beta = U_0^T u$ with precision zero! $\Rightarrow$ fixed effect!

$Z = BU_+$ and $w = U_+^T u$ with precision $\theta\, diag(d_+)$

The penalty

Using

$$P = U\ diag(d)U^T = U_+ diag(d_+)U_+^T$$

and

$$w = U_+^T u$$

we find

$$\theta u^T P u = \theta u^T U_+ diag(d_+) U_+^T u = \theta w^T diag(d_+) w$$

$\Rightarrow$ can assume $\mathrm{var}(w) = \theta^{-1} diag(d_+^{-1}) = \sigma^2 diag(d_+^{-1})$

The variance-covariance matrix for u

The penalty again:

$$\theta u^T P u = \theta u^T U_+ diag\left(d_+\right) U_+^T u = \theta w^T diag\left(d_+\right) w$$

Observing that

$$P^+ = U_+ diag\left(d_+^{-1}\right) U_+^T$$

is the Moore-Penrose inverse of P, we can also just fit the random effect Bu

assuming that

$$\mathrm{var}\left(u\right) = \sigma^2 P^+$$

But: Must also fit fixed effects $X\beta = BU_0^T u$ (Lee et al. 2021)

What are the fixed effects? (What's in the null space?)

p = 1:

$X = 1$ (just an intercept)

p = 2:

$X = \begin{pmatrix} 1 & h \end{pmatrix}$

h = vector of plot numbers

$\Rightarrow$ linear regression on plot numbers

Extending the model to two dimensions

Plots of a field trial on a regular grid with k rows and s columns

Model for spatial trend:

$$t = B_{rc} u_{rc}$$

where

$$B_{rc} = B_r \otimes B_c$$

B_r = matrix of m_r B-spline bases for k rows

B_c = matrix of m_c B-spline bases for s columns

A separable two-dimensional penalty

Consider differences $D_{rc}u_{rc}$ where

$$D_{rc} = D_r \otimes D_c$$

D_r = matrix of p-th differences for k rows

D_c = matrix of p-th differences for s columns

$\Rightarrow$ penalty $\theta_{rc} u_{rc}^T P_{rc} u_{rc}$

where

$$P_{rc} = P_r \otimes P_c = D_r^T D_r \otimes D_c^T D_c$$

What are the fixed effects?

$$(B_r \otimes B_c)(U_r \otimes U_c)\left(U_r^T \otimes U_c^T\right) u_{rc} = X_{00}\beta_{00} + X_{r0}\beta_{r0} + X_{0c}\beta_{0c} + Z_{rc}w_{rc},$$

$$\begin{aligned}
\rightarrow\ X_{00} &= (B_r \otimes B_c)(U_{0r} \otimes U_{0c}) = X_r \otimes X_c, \\
\rightarrow\ X_{r0} &= (B_r \otimes B_c)(U_{+r} \otimes U_{0c}) = Z_r \otimes X_c, \\
\rightarrow\ X_{0c} &= (B_r \otimes B_c)(U_{0r} \otimes U_{+c}) = X_r \otimes Z_c, \text{ and} \\
Z_{rc} &= (B_r \otimes B_c)(U_{+r} \otimes U_{+c}) = Z_r \otimes Z_c
\end{aligned}$$

Fixed effects (null space) have dimension

$$m_r m_c - (m_r - p)(m_c - p)$$

$\Rightarrow$ need extra smoothing terms for row and column main effects

An important detail when $p = 2$

Have design matrices

$X_r = \begin{pmatrix} 1 & h_r \end{pmatrix}$ and $X_c = \begin{pmatrix} 1 & h_c \end{pmatrix}$

When smoothing these using P-splines, have a random coefficient regression

$\Rightarrow$ need to allow for a covariance between intercept and slope

$\Rightarrow$ this is not nice but necessary to ensure invariance

$\Rightarrow$ to our knowledge this fact has been ignored in literature on P-splines

$\Rightarrow$ this complication is a good reason to favor $p = 1$

Special cases

Several older approaches turn out to be special cases when knots are at the plots and first differences are used (p = 1).

One particular one occurs when first-degree B-spline bases are used (q = 1). In this case the model is closely related to the

LV ⊗ LV

model (Piepho and Williams 2010; Boer et al. 2020).

This model, in turn, is a limiting case of the very popular

AR(1) ⊗ AR(1)

model (Gilmour et al. 1997).

An alternative penalty derived from sum of Kronecker products

$$u_{rc}^T \left\{ \theta_{rc1} P_r \otimes I_{m_r} + \theta_{rc2} I_{m_c} \otimes P_c \right\} u_{rc}$$

(Lee and Durbin 2011; Rodriguez-Alvarez et al. 2018; 'SpATS'; IAR model)

Advantage:

Null space is $X_r \otimes X_c$ ($\Rightarrow$ fixed effects) only has dimension p^2

Two Examples

Barley data or Durban et al. (2003)

Wheat data of Stroup et al. (1994)

⇒ Fitted randomization-based baseline model and added spatial components

TABLE 5 Analysis of barley data of Durbán et al. (2003) and wheat data of Stroup et al. (1994) using other common models. All models have fixed effects for replicates, genotypes, row numbers h_r, column numbers h_c, and the product of row and column numbers

		Barley data		Wheat data	
Model	**Description**	**Deviance**	***AIC***	**Deviance**	***AIC***
M53	Baseline[a]	410.19	412.19	1101.53	1103.53
M54	Baseline + row & column[b]	352.40	358.40	1083.52	1089.52
M55	AR1⊗AR1	299.60	305.60	1067.32	1073.32
M56	AR1⊗AR1 + nugget	278.20	286.20	1050.34	1058.34
M57	LV⊗LV	283.71	291.71	1051.33	1059.33

[a]Model with fixed effects for genotype, replicate, linear regression on row and column numbers as well as their product, and i.i.d. residual error.
[b]Baseline, adding random effects for rows and columns nested within replicates.

The separable penalty ($p = 1$)

TABLE 3 Analysis of barley data of Durbán et al. (2003) and wheat data of Stroup et al. (1994) using P-spline approach with $p = 1$ (first differences) and the penalty in (3) for B_{rc}. All models have fixed effects for replicates, genotypes, row numbers h_r, column numbers h_c, and the product of row and column numbers. The marginal smooths use $\text{var}(u_r) = P_r^+ \otimes \sigma_r^2$, $X_c = 1_s$, $\text{var}(u_c) = \sigma_c^2 \otimes P_c^+$ and $X_r = 1_k$ for all models

					Barley data		Wheat data	
Model	q	i_r	i_c	$\text{var}(u_{rc})$	**Deviance**	***AIC***	**Deviance**	***AIC***
M21	3	k	s	–	293.37	299.37	1072.47	1078.47
M22	3	k	s	$\sigma_{rc}^2 P_r^+ \otimes P_c^+$	279.28	287.28	1046.40	1054.40
M23	2	k	s	–	293.56	299.56	1071.91	1077.91
M24	2	k	s	$\sigma_{rc}^2 P_r^+ \otimes P_c^+$	279.18	287.18	1046.06	1054.06
M25	1	k	s	–	295.78	301.78	1075.14	1081.14
M26	1	k	s	$\sigma_{rc}^2 P_r^+ \otimes P_c^+$	278.45	286.45	1047.12	1055.12
M27	3	10	20	–	292.92	298.92	1073.38	1079.38
M28	3	10	20	$\sigma_{rc}^2 P_r^+ \otimes P_c^+$	281.18	289.18	1056.41	1064.41
M29	2	10	20	–	291.61	297.61	1073.27	1079.27
M30	2	10	20	$\sigma_{rc}^2 P_r^+ \otimes P_c^+$	279.74	287.74	1057.09	1065.09
M31	1	10	20	–	296.75	302.75	1075.43	1081.43
M32	1	10	20	$\sigma_{rc}^2 P_r^+ \otimes P_c^+$	281.31	289.31	1052.41	1060.41

M26 is doing very well $\Rightarrow$ mathematically close to LV $\otimes$ LV

The separable penalty ($p = 2$)

TABLE 2 Analysis of barley data of Durbán et al. (2003) and wheat data of Stroup et al. (1994) using P-spline approach with $p = 2$ (second differences), $i_r = k$, $i_c = s$, and the penalty in (3) for B_{rc}. All models have fixed effects for replicates, genotypes, row numbers h_r, column numbers h_c, and the product of row and column numbers

						Barley data				Wheat data			
						$q = 1$		$q = 3$		$q = 1$		$q = 3$	
Model	var(u_r)[a]	X_c[b]	var(u_c)[a]	X_r[b]	var(u_{rc})	Deviance	*AIC*	Deviance	*AIC*	Deviance	*AIC*	Deviance	*AIC*
M1	1	1_s	1	1_k	–	295.22	301.22	295.70	301.70	1076.03	1082.03	1075.93	1081.93
M2	1	1_s	1	1_k	$\sigma^2_{rc} P^+_r \otimes P^+_c$	292.20	300.20	293.09	301.09	1075.78	1083.78	1071.88	1079.88
M3	Σ	$\ddot{X}_c$	–	–	–	378.69	386.69	378.56	386.56	1075.45	1083.45	1075.51	1083.51
M4	Σ	$1_s : h_s$	–	–	–	378.69	386.69	378.56	386.56	%	%	%	%
M5	Σ	$a_s : b_s$	–	–	–	378.69	386.69	378.56	386.56	%	%	%	%
M6	Ω	$\tilde{X}_c$	–	–	–	378.72	384.72	378.62	384.62	1078.49	1084.49	1075.15	1079.15
M7	Ω	$1_s : h_s$	–	–	–	381.95	387.95	383.41	389.41	1082.32	1088.32	1081.92	1087.92
M8	Ω	$a_s : b_s$	–	–	–	378.87	384.87	378.72	384.72	1081.76	1087.76	1081.08	1087.08
M9	I_2	$\tilde{X}_c$	I_2	$\tilde{X}_r$	–	296.46	302.46	296.56	302.56	1058.02	1064.02	1060.05	1066.05
M10	I_2	$1_s : h_s$	I_2	$1_k : h_k$	–	317.58	323.58	316.77	322.77	1087.29	1093.29	1082.64	1088.64
M11	I_2	$a_s : b_s$	I_2	$a_k : b_k$	–	296.45	302.45	297.13	303.13	1058.02	1064.02	1059.84	1065.84
M12	Ω	$\tilde{X}_c$	Ω	$\tilde{X}_r$	–	296.41	306.41	295.23	305.23	1049.91	1059.91	1057.51	1067.51
M13	Ω	$1_s : h_s$	Ω	$1_k : h_k$	–	291.15	301.15	291.52	301.52	1057.10	1067.10	1057.34	1067.34
M14	Ω	$a_s : b_s$	Ω	$a_k : b_k$	–	287.88	297.88	288.27	298.27	1057.45	1067.45	1059.24	1069.24
M15	Σ	$\ddot{X}_c$	Σ	$\ddot{X}_r$	–	283.65	297.65	284.12	298.12	1043.37	1057.37	1044.22	1058.22
M16	Σ	$1_s : h_s$	Σ	$1_k : h_k$	–	%	%	%	%	%	%	%	%
M17	Σ	$a_s : b_s$	Σ	$a_k : b_k$	–	%	%	%	%	%	%	%	%
M18	I_2	$\tilde{X}_c$	I_2	$\tilde{X}_r$	$\sigma^2_{rc} P^+_r \otimes P^+_c$	293.23	301.23	293.60	301.60	1058.02	1066.02	1060.05	1068.05
M19	Ω	$\tilde{X}_c$	Ω	$\tilde{X}_r$	$\sigma^2_{rc} P^+_r \otimes P^+_c$	293.16	305.16	292.28	304.28	1049.91	1061.91	1057.49	1067.49
M20	Σ	$\ddot{X}_c$	Σ	$\ddot{X}_r$	$\sigma^2_{rc} P^+_r \otimes P^+_c$	280.12	296.12	280.27	296.27	1042.80	1058.80	1044.22	1058.22

[a]1: $P^+_r \otimes \sigma^2_r$ or $\sigma^2_c \otimes P^+_c$; I_2: $P^+_r \otimes \sigma^2_r I_2$ or $\sigma^2_c I_2 \otimes P^+_c$; Ω: $P^+_r \otimes \Omega_r$ or $\Omega_c \otimes P^+_c$; Σ: $P^+_r \otimes \Sigma_r$ or $\Sigma_c \otimes P^+_c$.

[b]Representation of X_r or X_c in the marginal smooth; $\tilde{X}$ obtained using U_0 of spectral decomposition of P as obtained by software (IML procedure of SAS); $\ddot{X}$ obtained using near optimal orthogonal rotation (Figures 2 and 4).

% Did not converge due to poor scaling of X_r and/or X_c.

$\Rightarrow$ Second differences not doing so well

Brief summary

- P-splines are a very versatile option for spatial modelling in field trials
- Several choices to make (many knobs to turn)
- First differences simpler than second differences; work quite well
- Okay to place knots at plots and use first-degree B-spline bases
- Close ties with older approaches [NN-analysis, random walk, LV, AR(1)]
- Our philosophy: always start with a randomization-based model as baseline; any spatial modelling is just an add-on
- Important disclaimer: A sophisticated spatial analysis is no substitute for good experimental design

References

Boer, M., Piepho, H.P., Williams, E.R. (2020): Linear variance, P-splines and neighbour differences for spatial adjustment in field trials: How are they related? *Journal of Agricultural Biological and Environmental Statistics* **25**, 676-698.

Lee, W., Piepho, H.P., Lee, Y. (2021): Resolving the ambiguity of random-effects models with singular precision matrix. *Statistica Neerlandica* **75**, 482-499.

Piepho, H.P., Boer, M., Williams, E.R. (2022): Two-dimensional P-spline smoothing for spatial analysis of field trials. *Biometrical Journal* **64**, 835-857.

Piepho, H.P., Ogutu, J.O. (2007): Simple state-space models in a mixed model framework. *The American Statistician* **61**, 224-232.

Piepho, H.P., Richter, C., Williams, E.R. (2008): Nearest neighbour adjustment and linear variance models in plant breeding trials. *Biometrical Journal* **50**, 164-189.

Piepho, H.P., Williams, E.R. (2010): Linear variance models for plant breeding trials. *Plant Breeding* **129**, 1-8.

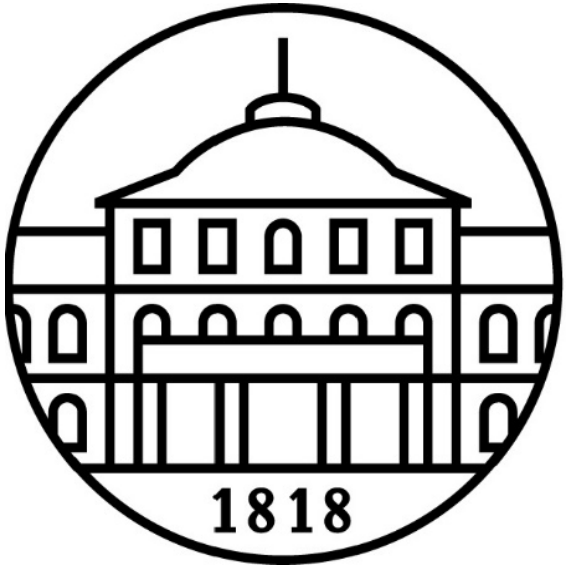
A two-stage approach to recovery of inter-block information and shrinkage of block effect estimates

Hans-Peter Piepho

Biostatistics Unit

Institute of Crop Science

Universität Hohenheim



The linear mixed model

$$y = X\beta + Zu + e$$

X , Z known design matrices

β , u fixed and random effects

$$E(u) = 0,\ E(e) = 0$$

$$\text{var}(u) = G,\ \text{var}(e) = \sigma_e^2 I,\ \text{cov}(u, e) = 0$$

$$\text{var}(y) = V = ZGZ^T + \sigma_e^2 I$$

Partitioned model

$$X\beta = X_1\beta_1 + X_2\beta_2$$

$$Zu = Z_1u_1 + Z_2u_2$$

$$\text{var}(u_1) = G_1,\ \text{var}(u_2) = G_2,\ \text{cov}(u_1, u_2) = 0$$

Effects of principal interest: β_1, u_1

Nuisance effects: β_2, u_2

The solution of the mixed model equations pertaining to β

$$\hat{\beta} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} y = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} C_{11} X_1^T + C_{12} X_2^T \\ C_{21} X_1^T + C_{22} X_2^T \end{pmatrix} V^{-1} y$$

where

$$\left(X^T V^{-1} X\right)^{-1} = \begin{pmatrix} X_1^T V^{-1} X_1 & X_1^T V^{-1} X_2 \\ X_2^T V^{-1} X_1 & X_2^T V^{-1} X_2 \end{pmatrix}^{-1} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

$\Rightarrow$ BLUE

The solution of the mixed model equations pertaining to u

$$\hat{u} = GZ^T V^{-1}\left(y - X\hat{\beta}\right) = \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \end{pmatrix} = \begin{pmatrix} G_1 Z_1^T \\ G_2 Z_2^T \end{pmatrix} V^{-1}\left(y - X\hat{\beta}\right)$$

$\Rightarrow$ BLUP

A two-stage representation of the BLUE and BLUP

If we knew β_2 and u_2, we could compute the corrected data

$$y_c^* = y - X_2\beta_2 - Z_2u_2$$

and fit the reduced model

$$y_c^* = X_1\beta_1 + Z_1u_1 + e$$

with $\text{var}\left(y_c^*\right) = V_c = Z_1G_1Z_1^T + \sigma_e^2 I$

In practice, need to replace the unknown effects β_2 and u_2 with their estimators:

$$y_c = y - X_2\hat{\beta}_2 - Z_2\hat{u}_2$$

where the estimates $\hat{\beta}_2$ and $\hat{u}_2$ are obtained from full model

Proposition 1: One may analyse y_c as if its variance-covariance structure were the same as that of y_c^* under the reduced model. Thus, we may fit the model

$$y_c = X_1\beta_1 + Z_1u_1 + e_c$$

as if $e_c = e$

Proposition 1 (cont'd): The estimators of β_1 and u_1 under this naïve model yield the BLUE of β_1 and the BLUP of u_1 under the full model, i.e.,

$$\tilde{\beta}_1 = \left(X_1^T V_c^{-1} X_1\right)^{-1} X_1^T V_c^{-1} y_c = \hat{\beta}_1$$

$$\tilde{u}_1 = G_1 Z_1^T V_c^{-1}\left(y_c - X_1 \tilde{\beta}_1\right) = \hat{u}_1$$

But:

While these equations yield correct point estimates of effects, they do not lend themselves for further statistical inference (significance tests or confidence intervals). For example, the variance of $\widetilde{\beta}_1$ is not equal to $\left(X_1^T V_c^{-1} X_1\right)^{-1}$.

Application: Removing spatial trend from raw data

u_2 = spatial trend t modelled using P-splines or any other geostatistical method

Correct raw data for u_2 and proceed with corrected data as if they were i.i.d.

$\Rightarrow$ Downstream point estimates okay

Problem: Can't get downstream inference (standard errors etc.) right

Solution: Need proper weights $\Rightarrow$ meta-analysis

Application: Recovery of inter-block information

u_2 = random incomplete block effects

$$\mathrm{var}(u_2) = I\sigma^2_{u2}$$

Two interesting limiting cases:

$\sigma^2_{u2} \to \infty$: BLUP of u_2 behaves like a fixed effect,

correction maximal, inter-block information absent

$\sigma^2_{u2} \to 0$: BLUP of u_2 shrinks to zero,

correction absent, inter-block information maximal

Reference

Piepho, H.P., Williams, E.R., Ogutu, J.O. (2013): A two-stage approach to recovery of inter-block information and shrinkage of block effect estimates. *Communications in Biometry and Crop Science* **8**, 10-22.