Journal of Agricultural Science (2015), **153**, 1187–1207. © Cambridge University Press 2014
doi:10.1017/S0021859614000823

CROPS AND SOILS RESEARCH PAPER

Treatment comparisons in agricultural field trials accounting for spatial correlation

C. RICHTER[1], B. KROSCHEWSKI[1], H.-P. PIEPHO[2]* AND J. SPILKE[3]

[1] *Faculty of Life Sciences, Albrecht Daniel Thaer-Institute of Agricultural and Horticultural Sciences, Humboldt-Universität zu Berlin, Berlin 10099, Germany*
[2] *Faculty of Agricultural Sciences, Institute of Crop Science, University of Hohenheim, Stuttgart 70599, Germany*
[3] *Institute of Agricultural and Nutritional Sciences, Martin-Luther-University Halle-Wittenberg, Halle 06099, Germany*

(*Received 22 November 2013; revised 16 May 2014; accepted 5 August 2014;*
first published online 9 September 2014)

SUMMARY

The classical analysis model for agricultural field trials is based on the principles of experimental design – randomization, replication and blocking – and it assumes independent residual effects. Accounting for any existent spatial correlation as an add-on component may be beneficial, but it requires selection of a suitable spatial model and modification of classical tests of treatment contrasts. Using a sugar beet trial laid out in complete blocks for illustration, it is shown that tests obtained with different modifications yield diverging results. Simulations were performed to decide whether different test modifications lead to valid statistical inferences. For the spherical, power and Gaussian models, each with six different values of the range parameter and without a nugget effect, the suitability of the following modifications was studied: a generalization of the Satterthwaite method (1941), the method of Kenward and Roger (1997), and the first-order corrected method described by Kenward and Roger (2009). A second-order method described by Kenward and Roger (2009) is also discussed and detailed results are provided as Supplemental Material (available at: http://journals.cambridge.org/AGS). Simulations were done for experiments with 10 or 30 treatments in complete and incomplete block designs. Model selection was performed using the corrected Akaike information criterion and likelihood-ratio tests. When simulation and analysis models were identical, at least one of the modifications for the *t*-test guaranteed control of the nominal Type I error rate in most cases. When the first-order method of Kenward and Roger was used, control of the *t*-test Type I error rate was poor for 10 treatments but on average very good for 30 treatments, when considering the best-fitting models for a given simulation setting. Results were not satisfactory for the *F*-test. The more pronounced the spatial correlation, the more substantial was the gain in power compared to classical analysis. For experiments with 20 treatments or more, the recommendation is to select the best-fitting model and then use the first-order method for *t*-tests. For *F*-tests, a randomization-based model with independent error effects should be used.

INTRODUCTION

The classical principles underlying the design of agricultural field experiments are replication, randomized allocation of treatments to experimental units (plots) and blocking (Fisher 1935). In the case of one-dimensional spatial trends, designs with complete or incomplete blocks are common, depending on the soil characteristics and the number of treatments tested. The use of incomplete blocks is preferable to complete blocks, when this allows a better control of the experimental error, leading to more precise and accurate treatment effect estimates as well as to a gain in statistical power. It is well known that there is usually a correlation of measurements taken on spatially adjacent plots, the correlation decaying with spatial distance among plots (Richter & Kroschewski 2012). Proper randomization ensures, however, that regardless of any such spatial patterns classical analysis procedures for designs with complete or

* To whom all correspondence should be addressed. Email: piepho@uni-hohenheim.de

incomplete blocks may proceed, assuming independent residual effects (Yates 1939; Grondona & Cressie 1991; Zimmerman & Harville 1991). The model with uncorrelated residual effects will henceforth be referred to as the *baseline model*. To refer to the baseline model, the abbreviation B-RCB is used for complete blocks and B-IB for incomplete blocks. Many authors have reported on studies showing that inclusion of spatial correlations between plots as an add-on component to a baseline model provides efficiency gains (Grondona & Cressie 1991; Zimmerman & Harville 1991; Brownie *et al.* 1993; Gilmour *et al.* 1997; Wu *et al.* 1998; Wu & Dutilleul 1999; Stroup 2002; Pilarczyk 2009; Müller *et al.* 2010; Richter & Kroschewski 2012).

Both the baseline model and the extended model with spatial add-on components can be written in general form as the following mixed linear model:

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{Zu}+\mathbf{e} \tag{1}$$

where $\boldsymbol{\beta}$ is $p\times 1$ vector of fixed effects, $\boldsymbol{u}$ is $q\times 1$ vector of random effects, $\mathbf{e}$ is $n\times 1$ vector of residual effects and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the known design matrices for fixed and random effects, respectively. The following assumptions are made on random effects: $\boldsymbol{u}\sim N(\mathbf{0},\ \boldsymbol{G})$, $\mathbf{e}\sim N(\mathbf{0},\ \boldsymbol{R})$ and $\mathrm{Cov}(\boldsymbol{u},\mathbf{e})=\mathbf{0}$, so that $\mathrm{Var}(\boldsymbol{y})=\boldsymbol{ZGZ'}+\boldsymbol{R}=\boldsymbol{V}.$

For the baseline model the matrix $\boldsymbol{R}$ takes the form $\boldsymbol{R}=\sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix. For the spatial models considered in the present paper, the covariances in $\boldsymbol{R}$ depend on spatial distances among plots. In the present paper, it is assumed, moreover, that spatial correlation does not depend on the direction (isotropy). Observations from two plots with Euclidean distance $d>0$ (measured between plot centres) have covariance $\mathrm{Cov}(d)=\sigma^2\ f(d)$. The spatial covariance models considered here differ only in the form of the function $f(d)$.

The use of spatial models entails two problems:

(i) When analysing a trial, the true underlying covariance model is unknown. Thus, a well-fitting model needs to be selected from a set of candidate models using a suitable selection criterion.

(ii) To obtain valid tests for treatment comparisons, a modification of classical test procedures is required. The following modifications are in common usage: the Satterthwaite method (henceforth abbreviated as SW; Satterthwaite 1941; Giesbrecht & Burns 1985; Fai & Cornelius 1996), the Kenward–Roger method (KR; Kenward & Roger 1997), and the first-order Kenward–Roger method (KR1; described by Kenward & Roger 2009; referred to as the Prasad–Rao estimator in Harville & Jeske (1992)). In addition, Kenward & Roger (2009) described a second-order method, henceforth denoted as KR2.

In the present investigation, the properties of Wald-type *F*-tests will be analysed under the global null hypothesis of no treatment effects and of Wald-type *t*-tests for pairwise comparisons.

The global null hypothesis H_0: $\boldsymbol{L'\beta}=\mathbf{0}$ with $\boldsymbol{L'}$ a matrix of contrasts can be tested by

$$\boldsymbol{F}=\frac{(\boldsymbol{L'\hat{\beta}})'(\boldsymbol{L'}(\boldsymbol{X'\hat{V}}^{-1}\boldsymbol{X})^{-}\boldsymbol{L})^{-1}\boldsymbol{L'\hat{\beta}}}{\mathrm{rank}(\boldsymbol{L})} \tag{2}$$

where $\boldsymbol{\hat{V}}$ is the restricted maximum-likelihood (REML) estimate of $\boldsymbol{V}$. When $\boldsymbol{L'}$ is a single contrast vector, one may also use

$$\boldsymbol{t}=\frac{\boldsymbol{L'\hat{\beta}}}{\sqrt{\boldsymbol{L'}(\boldsymbol{X'\hat{V}}^{-1}\boldsymbol{X})^{-}\boldsymbol{L}}} \tag{3}$$

For a single pairwise contrast, $\boldsymbol{L'}$ is a row vector with zeros everywhere, except for the two treatments being compared, which have coefficients –1 and +1. For testing the global null hypothesis that all *t* treatments have equal mean, $\boldsymbol{L'}$ will have *t*–1 linearly independent rows with pairwise contrasts. In general, the following problems occur with these tests (Kenward & Roger 1997, 2009; Schabenberger & Pierce 2002):

- $(\boldsymbol{X'\hat{V}}^{-1}\boldsymbol{X})^{-}$ is not generally an unbiased estimator of $(\boldsymbol{X'V}^{-1}\boldsymbol{X})^{-}$;
- $(\boldsymbol{X'\hat{V}}^{-1}\boldsymbol{X})^{-}$ underestimates the variance of $\boldsymbol{\hat{\beta}}$, because the variance of the estimator of $\boldsymbol{V}$ is not taken into account; and
- Test statistics in Eqns (2) and (3) follow an exact *F*-distribution and an exact *t*-distribution only in exceptional cases; in the majority of cases the denominator degrees of freedom need to be suitably approximated.

The SW method operates on $(\boldsymbol{X'\hat{V}}^{-1}\boldsymbol{X})^{-}$ where $\boldsymbol{\hat{V}}$ is the REML estimate of $\boldsymbol{V}$. It approximates the denominator degrees of freedom in Eqns (2) or (3) using the method of moments. The KR, KR1 and KR2 methods correct, in slightly different ways, the plug-in estimator of $(\boldsymbol{X'V}^{-1}\boldsymbol{X})^{-}$ and then use the corrected estimator to approximate the denominator degrees of freedom.

Problems (i) and (ii) arise frequently in experiments with repeated measures in time or space (Piepho *et al.* 2004). Problem (ii) also occurs with experiments, in which a more complex covariance structure results from the randomization. For example, in split-plot and strip-plot designs, where treatment contrasts involve several variance components, it may be necessary to approximate the error degrees of freedom, particularly when missing data occur (Fai & Cornelius 1996; Spilke *et al.* 2005). Studies by Kenward & Roger (1997, 2009), Gomez *et al.* (2005) and Schaalje *et al.* (2002) demonstrate that the validity of tests depends on the method chosen, the underlying covariance model and its parameter values, the parameterization of the model and the sample size. When using a spatial add-on component in the analysis of field trials, however, the data structures are not directly comparable, so that results of these studies do not apply. In field trials, the option to fit a spatial correlation structure does not follow from the randomization theory, but it is justified by the commonly observed similarity of adjacent plots across the whole trial field or parts thereof, e.g. within blocks, due to biotic and abiotic effects.

So far, little has been published on the control of the Type I error and the power of the *t*- and *F*-tests using the modifications in (ii) for spatial covariance structures in the context of field trials. For the spherical and exponential models with different parameter values and two trial designs with complete blocks, Hu *et al.* (2006) showed by simulation that the SW method generally yields more valid test results than the KR method. In particular, when no nugget was present, bias in estimates of standard errors of a difference (S.E.D.) obtained by the KR method was substantial. Spilke *et al.* (2010) simulated split-plot designs assuming an exponential covariance model with nugget effect and/or large-scale trend effects. Here, the KR1 method outperformed the SW method. In contrast to Hu *et al.* (2006), Spilke *et al.* (2010) combined each simulated random field with a newly randomized design, so that the results refer to an average over all randomizations. The comparison of the corrected Akaike information criterion (AICC; Hurvich & Tsai 1989) and the Bayesian information criterion (BIC, Schwarz 1978) as a selection criterion did not lead to a clear preference of one criterion over the other. Using AICC, Richter & Kroschewski (2012) showed in simulated experiments with four and ten treatments projected onto uniformity trial data, that trial geometry, position of the trial area and the randomization layout influence the outcome of model selection. Properties of *t*- and *F*-tests were analysed for several analysis models and models best-fitting according to AICC, and using the SW or KR1 methods. As the true underlying covariance model in the uniformity trials was not known, the causes of the observed biases in S.E.D. and the partial lack of Type I error control remained elusive.

The present paper reports results on an experiment with sugar beet that illustrates the pros and cons of using spatial add-on components in the analysis and shows that the various modifications may yield contrasting results, leaving the user with the problem of identifying the best method. For this purpose, simulations are required. Therefore, in the present paper, problems (i) and (ii) were investigated by simulation for experiments with 10 or 30 treatments for typical dimensions of plots and complete and incomplete blocks. The analyses only considered models without nugget effect, because according to simulations by Hu *et al.* (2006) models without nugget may be problematic in terms of bias. Moreover, several authors reported that inclusion of a nugget is often not necessary for field trials (Besag & Kempton 1986; Zimmerman & Harville 1991; Schabenberger & Pierce 2002; Richter & Kroschewski 2012). But other authors reported analyses where addition of a nugget led to a better model fit (Piepho & Williams 2010), suggesting that the need of a nugget effect is data-dependent and possibly dependent on the spatial covariance model. Expanding on previous results in other publications, the current analyses focused on the following issues:

- Assessment of bias of S.E.D. estimates and empirical Type I error rate at a nominal significance level of $\alpha = 0{\cdot}05$ for *t*- and *F*-tests using the SW-, KR- and KR1 methods and considering all combinations of simulation and analysis model.
- Performance of *t*- and *F*-tests for the best-fitting models and discussion in relation to the chosen model selection criterion.
- Power analysis assuming spatial models in comparison with models for designs with complete and incomplete blocks. The latter comparison is of particular interest, because with these designs the model selection problem does not arise, unless spatial model components are used.

Results on the KR2 method are also reported briefly. More detailed results are presented in the Supplemental Material (available at: http://journals.cambridge.org/AGS).

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

EXAMPLE

A three-factorial sugar beet experiment was conducted to assess the effects of *variety* (varieties 1–3), *N-fertilization* (0, 80 and 120 kg N/ha) and a *growth enhancer* (with and without) on corrected sugar yield (t/ha). The experiment was laid out in four randomized complete blocks. The plot size was $3 \times 8\ \text{m}^2$. Only the three central crop rows were harvested. The field layout was a grid of four rows and 18 columns with rows corresponding to blocks. This experiment was analysed using the MIXED and GLIMMIX procedures of the SAS System (Version 9.3 /SAS/STAT 12.1). The following models were fitted (all with fixed block effects): baseline model B-RCB and 12 spatial models (isotropic and anisotropic models, with and without nugget effect, with spatial correlations either confined within blocks or extending across the whole experiment) using the default starting values for the variance parameters. For three models, numerical problems occurred, because the iterative fitting algorithm did not converge. The best-fitting model was the power model without nugget effect and correlation extending across the whole experiment (henceforth referred to as PM; Fig. 1) with estimated parameters $\hat{\sigma}^2 = 0{\cdot}5342$ and $\hat{\rho} = 0{\cdot}7956$. The AICC-value was 133·8, whereas that for B-RCB was 139·9. A likelihood-ratio test (LRT) further indicated a significantly better fit of the power model ($P < 0{\cdot}01$). The power model was analysed with the four approximation methods SW, KR, KR1 and KR2. In the example, the approximated degrees of freedom of KR, KR1 and KR2 were identical and different from SW for *F*-tests. For the *t*-tests, there were no differences in the degrees of freedom between methods (Tables 1 and 2). For all main effects and the interactions, an *F*-test was performed. A *t*-test was only done for varieties and fertilizer levels, because the growth enhancer showed no effects and all interactions were non-significant. The improved fit of the power model compared with the baseline model meant that the *P*-values of *F*- and *t*-tests (Tables 1 and 2) for several effects were reduced. There are notable differences between the four methods, so the user is left with the problem of choice. This is particularly true for the comparison of the second and third N-fertilizer levels (N2 and N3), where the p-value was either below or above the nominal significance level of 0·05. Whether the use of spatial models for analysis can be recommended and if so, which method of approximation is preferable depends on the degree of control of the nominal significance level (often set at 0·05; the same practice is followed in the present paper). Assuming a satisfactory control of the Type I error rate, a smaller *P*-value would correspond to a higher power for a given comparison. Whether or not this is the case cannot be decided based on a single experiment, but requires extensive simulation.

SIMULATIONS

All simulations and analyses were performed with SAS (Version 9.3/STAT 12.1). A complete overview on the simulation and analysis models including the used abbreviations is given in Fig. 1. Designs that are commonly used in field trials were assumed, i.e. trials with $t = 10$ and 30 treatments with $r = 4$ replicates. The plot size was $2 \times 8\ \text{m}^2$. Plots of a complete replicate were located side-by-side in the same row of plots. Replicates were arranged in separate rows, so that the trial size was $20 \times 32\ \text{m}^2$ or $60 \times 32\ \text{m}^2$.

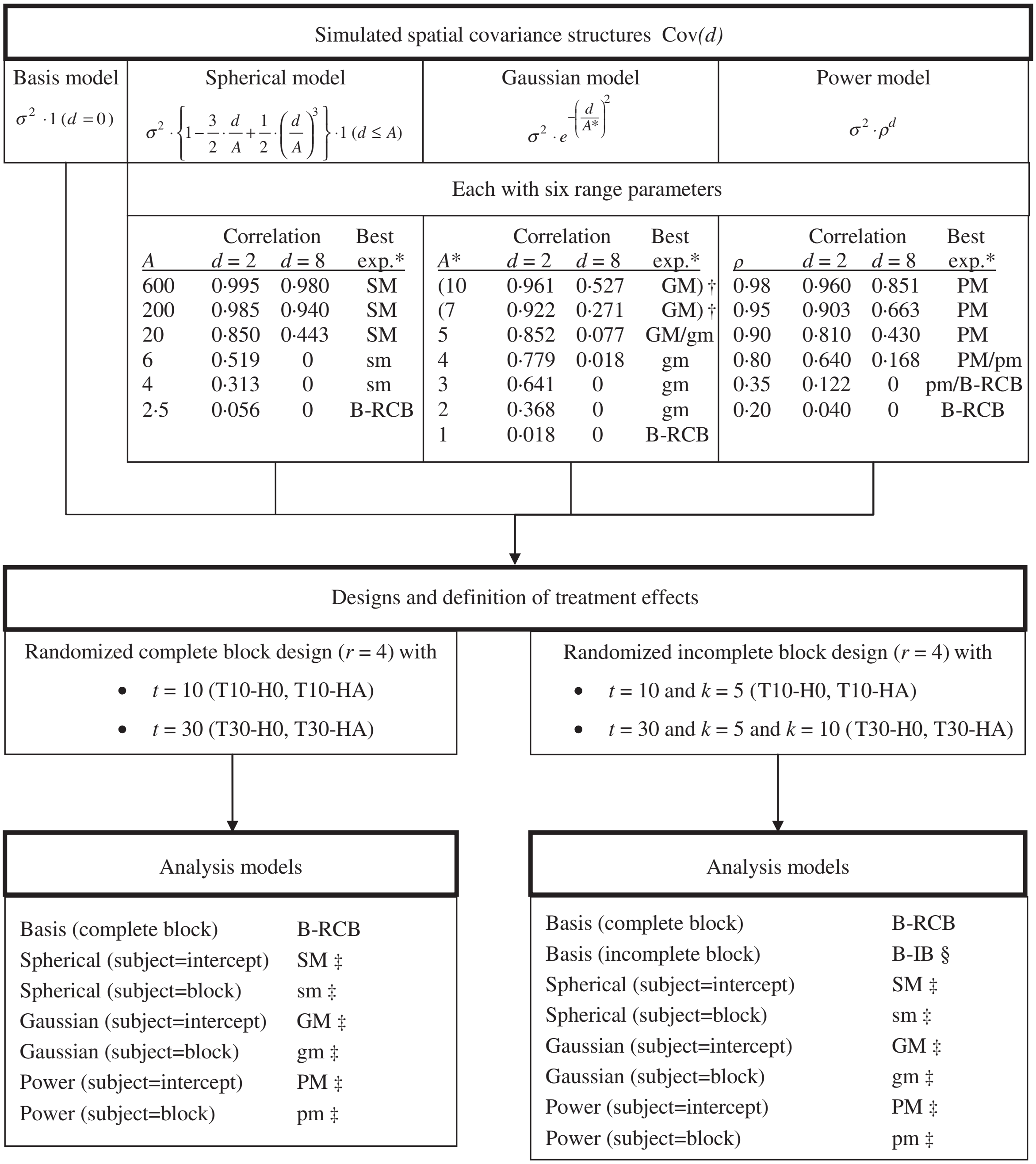
For the whole trial area, 10 000 random fields were simulated according to the baseline model and three spatial models – spherical, Gaussian and power – assuming a residual variance of $\sigma^2 = 200$ throughout. For each of the spatial models, six different values were considered for the range parameters (Fig. 1), spanning a wide range of correlation structures. For the Gaussian model and designs with 10 or 30 treatments, different maximum values of the range parameter ($A^+ = 10$ and $A^+ = 7$, respectively) were selected, because for larger values and given dimensions of the field and blocks the variance–covariance matrix $\boldsymbol{R}$ to be used for simulating the data either became non-positive definite or the parameter estimation was associated with substantial convergence problems.

The power model is equivalent to the exponential model $\text{Cov}(d) = \sigma^2 \exp(-d/A^+)$ with $A^+ = -1/\log(\rho)$, when $0 < \rho < 1$. Better convergence properties were observed with the power model and so this was preferred in simulations.

The following trial designs were projected onto 10 000 simulated fields each:

- Designs with randomized complete blocks for $t = 10$ and 30, each of which was randomized 10 000 times.
- Resolvable incomplete block designs, generated using CycDesigN 4·1 (Whitaker *et al.* 2009) and randomized 10 000 times. For $t = 10$, an incomplete block design with $k =$ five plots per block and an average efficiency factor of $E = 0{\cdot}866$ was generated;

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press



Simulated spatial covariance structures Cov(*d*)

Basis model	Spherical model	Gaussian model	Power model
$\sigma^2 \cdot 1\,(d=0)$	$\sigma^2 \cdot \left\{1-\frac{3}{2}\cdot\frac{d}{A}+\frac{1}{2}\cdot\left(\frac{d}{A}\right)^3\right\}\cdot 1\,(d \le A)$	$\sigma^2 \cdot e^{-\left(\frac{d}{A^*}\right)^2}$	$\sigma^2 \cdot \rho^d$

Each with six range parameters

	Correlation		Best
A	$d=2$	$d=8$	exp.*
600	0·995	0·980	SM
200	0·985	0·940	SM
20	0·850	0·443	SM
6	0·519	0	sm
4	0·313	0	sm
2·5	0·056	0	B-RCB

	Correlation		Best
A^*	$d=2$	$d=8$	exp.*
(10	0·961	0·527	GM) †
(7	0·922	0·271	GM) †
5	0·852	0·077	GM/gm
4	0·779	0·018	gm
3	0·641	0	gm
2	0·368	0	gm
1	0·018	0	B-RCB

	Correlation		Best
ρ	$d=2$	$d=8$	exp.*
0·98	0·960	0·851	PM
0·95	0·903	0·663	PM
0·90	0·810	0·430	PM
0·80	0·640	0·168	PM/pm
0·35	0·122	0	pm/B-RCB
0·20	0·040	0	B-RCB

Designs and definition of treatment effects

Randomized complete block design ($r = 4$) with

- $t = 10$ (T10-H0, T10-HA)
- $t = 30$ (T30-H0, T30-HA)

Randomized incomplete block design ($r = 4$) with

- $t = 10$ and $k = 5$ (T10-H0, T10-HA)
- $t = 30$ and $k = 5$ and $k = 10$ (T30-H0, T30-HA)

Analysis models

Basis (complete block)	B-RCB
Spherical (subject=intercept)	SM ‡
Spherical (subject=block)	sm ‡
Gaussian (subject=intercept)	GM ‡
Gaussian (subject=block)	gm ‡
Power (subject=intercept)	PM ‡
Power (subject=block)	pm ‡

Analysis models

Basis (complete block)	B-RCB
Basis (incomplete block)	B-IB §
Spherical (subject=intercept)	SM ‡
Spherical (subject=block)	sm ‡
Gaussian (subject=intercept)	GM ‡
Gaussian (subject=block)	gm ‡
Power (subject=intercept)	PM ‡
Power (subject=block)	pm ‡

Fig. 1. Overview of simulation models and analysis models. r = number of replicates, t = number of treatments, k = number of treatments per incomplete block. T10-H0, T10-HA, T30-H0, T30-HA definition of treatment effects (see text). * Expected best-fitted analysis model. †$A^+ = 10$ for 10 treatments; $A^+ = 7$ for 30 treatments. ‡Analysed with SW, KR, KR1 and KR2. § Analysed with SW and KR = KR1 = KR2.

for $t = 30$ a design with $k = 5$ ($E = 0{\cdot}805$) and a design with $k = 10$ ($E = 0{\cdot}915$) was constructed.

Every randomized plan was combined with a random field, yielding 10 000 simulated trials.

In all simulations, effects for the four complete replicates were set to 8, 24, 40 and 56. For the treatment effects, four variations to the simulation scheme were considered, corresponding to the following objectives:

- (T10-H0) for $t = 10$: all 45 treatment differences were zero to analyse bias of S.E.D. estimates and control of the nominal Type I error rate of t- and F-tests;
- (T10-HA) for $t = 10$: three treatment effects were zero and the others were 1·2, …, 8·4 (1·2), so that

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

treatment differences were staggered according to Δ=0, 1·2, …, 8·4 (1·2) for analysis of power;

- (T30-H0) for $t=30$: all 435 treatment differences were zero with the same analysis aim as for (T10-H0);
- (T30-HA) for $t=30$: with the chosen treatment effects (7 times 0, 2 times 1·2, 2·4, 3·6 and 4·8, 3 times 6·0, 5 times 7·2 and 7 times 8·4), so that 59 treatment differences were equal to zero and 376 differences were staggered according to Δ=1·2, …, 8·4 (1·2) for assessments analogous to (T10-HA).

Table 1. *Results of the analysis of the sugar beet trial. F-tests for the analysis corresponding to B-RCB and the best-fitting model (PM) with four correction methods*

		Denominator D.F.			*F* value					Probability > *F*				
			Power model			Power model					Power model			
Effect	Numerator D.F.	B-RCB	SW	KR = KR1 = KR2	B-RCB	SW	KR	KR1	KR2	B-RCB	SW	KR	KR1	KR2
Variety (V)	2	51	41·90	41·97	7·35	17·84	15·83	17·15	17·42	0·0016	$2·49\times10^{-6}$	$7·54\times10^{-6}$	$3·61\times10^{-6}$	$3·10\times10^{-6}$
N fertilizer (N)	2	51	37·18	37·38	34·95	60·43	54·18	58·90	60·13	$2·76\times10^{-10}$	$2·08\times10^{-12}$	$9·03\times10^{-12}$	$2·80\times10^{-12}$	$2·08\times10^{-12}$
Growth enhancer (GE)	1	51	44·94	44·94	0·39	0·05	0·05	0·05	0·05	0·5312	0·8187	0·8291	0·8216	0·8217
V × N	4	51	39·41	39·49	0·40	1·35	1·20	1·31	1·32	0·8145	0·2670	0·3253	0·2845	0·2794
V × GE	2	51	45·32	45·28	1·48	2·24	1·98	2·12	2·14	0·2374	0·1182	0·1521	0·1320	0·1292
N × GE	2	51	38·41	38·48	0·31	0·35	0·31	0·33	0·34	0·7324	0·7099	0·7370	0·7183	0·7145
V × N × GE	4	51	39·29	39·20	0·40	0·54	0·48	0·52	0·53	0·8107	0·7100	0·7523	0·7234	0·7166
Block	3	51	8·98	9·32	0·58	0·39	0·41	0·38	0·38	0·6300	0·7631	0·7527	0·7689	0·7681

Each simulated trial with complete blocks was analysed using a linear model with fixed effects for treatment and replicate and random residual effects following the baseline model as well as the Gaussian, power and spherical models. For spatial models, two alternatives were considered: correlation expanded across the whole field (option subject=intercept in SAS, yielding a single subject) or only across plots in the same block (option subject=block, yielding r=four subjects). Each of these analyses was done by the SW, KR, KR1 and KR2 methods. Simulated trials with incomplete blocks were analysed with the same models. In addition, these data were analysed using a model with fixed replicate and treatment effects and a random effect for incomplete blocks (Fig. 1, analysis model B-IB). This approach implied that observations of the same block had a constant covariance not depending on spatial distance, while observations from different blocks were independent. In this case, the covariance structure was intrinsically linear (Kenward & Roger 2009), so that the KR, KR1 and KR2 methods yielded the same results. Analyses with the SW, KR and KR1 methods were performed using the MIXED procedure of the SAS System, employing the REML-method without provision of starting values or bounds for the variance parameters. These three methods were also implemented in the GLIMMIX procedure. The KR2 method was only implemented in the GLIMMIX procedure. GLIMMIX and MIXED employ different fitting algorithms (GLIMMIX uses a quasi-Newton algorithm and MIXED a ridge-stabilized Newton–Raphson algorithm) and convergence criteria. The different algorithms often yield identical results (for example in the sugar beet experiment), but in some cases convergence properties and results may differ. For a given approximation method, neither a variation of convergence criteria nor the selection of different fitting algorithms led to identical results for both procedures

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

Table 2. *Results of the analysis of the sugar beet trial. t-tests for three varieties (V1, V2, V3) and three N fertilizer levels (N1, N2, N3) corresponding to B-RCB and the best-fitting model (PM) with four correction methods*

	Estimated differences		Standard error of differences					D.F.		Probability > |*t*|				
				Power model					Power model		Power model			
Comparison	B-RCB	Power model	B-RCB	SW	KR	KR1	KR2	B-RCB	SW = KR = KR1 = KR2	B-RCB	SW	KR	KR1	KR2
V1–V2	0·6713	0·8924	0·1967	0·1636	0·1736	0·1668	0·1654	51	39·95	0·0013	$2{\cdot}76\times10^{-6}$	$7{\cdot}56\times10^{-6}$	$3{\cdot}88\times10^{-6}$	$3{\cdot}33\times10^{-6}$
V1–V3	0·6333	0·8114	0·1967	0·1732	0·1844	0·1774	0·1763	51	43·16	0·0022	$2{\cdot}80\times10^{-5}$	$7{\cdot}00\times10^{-5}$	$4{\cdot}00\times10^{-5}$	$3{\cdot}66\times10^{-5}$
V2–V3	−0·0381	−0·0811	0·1967	0·1746	0·1861	0·1792	0·1779	51	43·21	0·8473	0·6447	0·6653	0·6532	0·6509
N1–N2	−1·2487	−1·2937	0·1967	0·1471	0·1547	0·1484	0·1464	51	33·75	$5{\cdot}80\times10^{-8}$	$3{\cdot}00\times10^{-10}$	$9{\cdot}74\times10^{-10}$	$3{\cdot}72\times10^{-10}$	$2{\cdot}66\times10^{-10}$
N1–N3	−1·5508	−1·6401	0·1967	0·1658	0·1759	0·1688	0·1676	51	40·08	$2{\cdot}20\times10^{-10}$	$2{\cdot}56\times10^{-12}$	$1{\cdot}36\times10^{-11}$	$4{\cdot}29\times10^{-10}$	$3{\cdot}50\times10^{-12}$
N2–N3	−0·3021	−0·3464	0·1967	0·1636	0·1736	0·1668	0·1653	51	39·74	0·1308	0·0405	0·0528	0·0443	0·0429

in all simulation runs. The main focus of the present paper will be on the results obtained with MIXED, because this procedure yielded slightly better results in terms of bias of S.E.D. estimates and the control of the nominal Type I error rate for *t*-tests (checked for KR1). The direct comparison of the KR2 and KR1 methods, controlling for the optimization methods, was only possible in GLIMMIX, which implements both methods. The results of that comparison are reported here only briefly. More details can be found in the Supplemental Material (available at: http://journals.cambridge.org/AGS).

For each simulated trial, the best-fitting model was selected by AICC because Burnham & Anderson (1998) recommended this criterion for trials with a ratio between sample size and number of variance parameters smaller than 40. For a larger ratio, the AICC approaches the usual Akaike information criterion (AIC; Akaike 1973). If the best-fitting model was a spatial model, it was subsequently compared to B-RCB by a LRT. When the best spatial model was not significantly better than B-RCB according to a LRT, the B-RCB model was selected as the best model. Similarly, the B-RCB model was selected when no spatial model converged. When several spatial models had the same smallest AICC and were significantly better than the baseline model by a LRT, then the simulation model was virtually always among the tied best models. In these cases, the simulation model was declared as the best one. This approach was modified only in case of a simulated power model and a tied first rank for the power and spherical models. Here, the spherical model was selected as the best model, because the power model frequently caused numerical problems in the subsequent significance tests (for details see the Results section). Thus, for every simulated experiment there was a selected best analysis model that is independent of the approximation method and hence of the resulting parameter estimates and significance tests. The number of times out of 10 000 runs an analysis model was selected as the best one for a given simulation model under T10-H0 and T30-H0 was determined.

Depending on the combination of simulation and analysis model as well as on the approximation method, there was a varying number of simulation runs with convergence problems, which were excluded from final summaries. In addition, simulation runs meeting at least one of the following conditions were

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

excluded: (i) The *F*-value for the global null hypothesis tended to infinity, corresponding to a numerically exact *P*-value of zero; (ii) for at least one treatment comparison, the *t*-value was reported as plus or minus infinity, meaning the S.E.D. was reported as an exact zero; (iii) in the analysis according to PM or pm, the estimate of ρ was close to ±1 ($1 - |\hat{\rho}| < 10^{-4}$); and (iv) the *P*-value of the *F*-test was nearly 1 ($1 – P < 5 \times 10^{-8}$). In most of the excluded cases, more than one of these conditions was met. The exclusion of these cases meant that the assessment of bias of S.E.D. estimates as well as of the empirical Type I error rates were based on 10 000 simulation runs or less. The total number of runs among the 10 000 that were used for final assessment will be referred to as *effective number of runs*.

The bias of S.E.D. estimates was assessed according to

$$\text{bias of s.e.d. } (\%) = (\overline{\text{S.E.D.}}_{\text{est}} - \overline{\text{S.E.D.}}_{\text{obs}})/\overline{\text{S.E.D.}}_{\text{obs}} \times 100$$

where $\overline{\text{S.E.D.}}_{\text{obs}}$ is the square root of the mean variance of all estimated treatment differences and $\overline{\text{S.E.D.}}_{\text{est}}$ the square root of the mean-estimated variances of the treatment differences, in both cases averaged over all effective number of runs and all treatment comparisons.

In addition, for every simulation model the bias of the S.E.D. and the empirical Type I error rate for the best-fitting model were assessed. Again, the effective number of runs was 10 000 or less.

For the *t*-test, the power analyses were conducted under T10-HA and T30-HA, but are meaningful only if the simulations under the corresponding null hypothesis yielded control of the nominal Type I error rate according to a binomial test. Thus, when the empirical rejection rate was contained in the interval $\left\langle 0{\cdot}05 \pm 1{\cdot}96 \times \sqrt{\frac{0{\cdot}05 \times (1 - 0{\cdot}05)}{\text{effective number of runs}}} \right\rangle$, the control was considered to be satisfactory.

RESULTS

For trials laid out in complete or incomplete blocks, analysis according to B-RCB or by spatial models yielded similar results. Therefore, only results for complete blocks are reported here in detail, while for incomplete blocks only results for the B-IB model are reported. The following results were obtained with the MIXED procedure.

Agreement between simulation model and best-fitting model

A model selection strategy is successful if the true (simulated) model is correctly identified. Relative merits of models with correlation extending across the whole field and models with correlations restricted within blocks were expected to depend on the underlying range parameter and thus on the strength of correlation. In particular, a transition to the baseline model was expected for weak correlations for $d = 2$ (Fig. 1). Table 3 shows the relative frequency of selected models based on simulations for T10-H0 and T30-H0 for trials with complete blocks. For simulations under T10-HA and T30-HA, the results were essentially identical (the maximum deviation compared to Table 3 was 1·5%). Results for trials with incomplete blocks are included in aggregated form.

For the Gaussian model the agreement between simulated and best-fitting model for $t = 10$ and $t = 30$ was good. When the true model was the power model, then with $t = 10$ models SM and sm were preferentially selected, whereas for $t = 30$ the PM and pm models were selected more often. For the spherical model with $t = 10$ and $A \geqslant 20$, the frequency of correct selections was modest. When the range parameter was smaller, the model B-RCB was most frequently selected. With $t = 30$, correct selection occurred only when $A \geqslant 200$, while for a smaller range models PM and gm and for $A \leqslant 2{\cdot}5$ B-RCB were preferred in model selection. All results agree in that with decreasing covariance there was an increasing tendency to favour the baseline model B-RCB. The baseline model itself was correctly selected in 96% ($t = 10$) and 97% ($t = 30$) of the cases. Overall, the Gaussian model was more often selected correctly than the power and spherical models.

For incomplete blocks and a model choice between B-RCB, B-IB and spatial models, the B-IB model was rarely selected. For $t = 30$ the proportion was somewhat higher for $k = 5$ than for $k = 10$, because a constant covariance is a more realistic assumption for smaller blocks.

When there was only a choice between B-RCB and B-IB and simulation was with incomplete blocks, then model fit with B-IB was generally better when correlations are strong. This finding was particularly true for $t = 30$ and was more pronounced with $k = 5$ than with $k = 10$ (results not shown).

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

Bias of estimated standard error of a difference

Table 4 shows that the bias of S.E.D. estimates depended on the combination of simulation and analysis model, the value of the range parameter, the number of treatments and the method for adjusting *t*- and *F*-tests (SW, KR and KR1). Results for the range parameter values not shown in Table 4 followed the same trend.

When simulation and analysis model agreed, a relatively small bias was expected. But the current results show that this was not generally the case:

- For the spherical model and analysis according to SM and sm, for all $A \geqslant 20$ the bias was minimal. As *A* decreased, bias became increasingly negative for SM. For $t=30$ and SM results were less favourable than for $t=10$. When analysing according to sm with $t=10$, the negative tendencies were milder than for SM, and for $t=30$ starting from $A \leqslant 4$ a bias was practically non-existent. For SW, KR and KR1 biases were comparable in magnitude.
- For the Gaussian model, biases for $t=30$ were considerably smaller than for $t=10$ most of the time. With $t=10$, SW entailed slight negative biases, while for KR1 the bias was slightly positive. The results for GM and gm were very similar.
- The power model in conjunction with KR showed extreme biases, more severely so for $t=10$ than for $t=30$. For $t=30$, the differences between SW and KR1 were minor.

Some noteworthy results when simulation and analysis model did not coincide are now discussed briefly:

- In general, only a small bias was observed for $t=10$ with KR1 and for $t=30$ with all three methods in case of weak correlation. This was not the case for analysis according to SM ($t=10$ and 30) and sm ($t=10$).
- When the true model was spherical or the power model and analysis was done by GM, results were very similar. The negative bias with strong correlations became smaller with decreasing correlation for all three methods, and with KR turned positive or vanished altogether, as it did with KR1. With large correlations, bias for $t=30$ was more pronounced than with $t=10$. Analysis by gm yielded nearly identical results as GM.
- When analysing the spherical model with models PM or pm, biases showed similar values as with the simulated power model.
- For the Gaussian model, severe positive biases occurred for range parameter values $A^{+} \geqslant 4$ and analysis by SM, sm, PM or pm with all methods.
- For incomplete block designs, analyses by spatial models yielded similar results as for complete block designs. When analysing by B-IB combined with SW, the bias ranged from –2·2 to –0·26% and with KR (= KR1) it ranged from –0·4 to 0·9% (results not shown).
- For the analysis according to B-RCB, the bias was between –0·5 and +0·4% for all simulation models (results not shown).

Control of the nominal Type I error rate for *t*- and *F*-tests

The biases reported in Table 4 are marked when the nominal Type I error rate was controlled according to the binomial test. The maximum of the effective number of runs on which the binomial test was based was 10 000 and for T10-H0 the minimum number was 1911 (SW and KR1) and 1139 (KR), respectively, and for T30-H0 the minimum was 6859 (SW and KR1) and 1691 (KR), respectively. The largest numbers of runs that did not converge or were excluded according to the criteria (i)–(iv) occurred when the analysis was done according to PM or pm, regardless of the simulation model.

When the bias of the S.E.D. estimate was small, the empirical Type I error rate tended to be controlled at the nominal level. Negative bias in S.E.D. tended to be associated with liberal Type I error rates, and similarly, positive bias in S.E.D. was associated with conservative Type I error rates. The association between bias in S.E.D. and Type I error rates is apparent from Table 4. When simulation and analysis models coincided, there was almost always (the spherical model was an exception) one approximation that guaranteed control of the Type I error rate of *t*-test. For the *F*-test, error control was obviously more problematic than with the *t*-test (Table 4). There was no case where the *F*-test controlled the nominal level, but the *t*-test did not.

In Figs 2 and 3, empirical Type I error rates are depicted for $t=30$ for the *t*-and *F*-tests based on trials with complete blocks. The *F*-test consistently showed stronger departures from the nominal level than the *t*-test. Results for KR1 are now considered in detail. For $t=30$ the *t*-test was most conservative with a rejection rate of 0·001 (bias of S.E.D. 66·7%), and most liberal

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

Table 3. *Best-fitting models according to AICC and LRT (%) in 10 000 runs of each simulation model. Bold faced = largest observed percentage, grey = largest expected percentage according to Fig. 1. Empty cells represent a percentage of 0%*

			Complete block designs							Incomplete block designs					
										k = 5			*k* = 10		
			Analysis models												
Simulation model	*t*	Range	B-RCB	SM	sm	GM	gm	PM	pm	B-RCB	B-IB	Σ spatial models	B-RCB	B-IB	Σ spatial models
Spherical	10	600	4·8	**53·7**	14·8	6·2	2·6	12·4	5·5	3·9	3·0	**93·1**			
		200	4·6	**53·4**	14·9	6·2	2·7	12·8	5·4	4·0	3·1	**92·9**			
		20	5·2	**47·9**	18·3	6·9	3·2	12·7	5·8	3·9	3·2	**92·9**			
		6	**28·7**	4·9	18·7	12·2	19·2	7·9	8·4	25·8	5·2	**69·0**			
		4	**67·5**	1·0	6·3	5·2	11·7	4·6	3·7	**62·4**	6·6	31·0			
		2·5	**94·0**	0·2	1·2	0·8	1·9	1·0	0·9	**90·0**	4·5	5·5			
	30	600		**53·4**	3·7	0·1		38·9	3·9			**100·0**			**100·0**
		200		**51·9**	3·7	0·1		40·5	3·8			**100·0**			**100·0**
		20		8·0	16·3	0·2	0·1	**65·8**	9·6			**100·0**			**100·0**
		6	0·3		7·8	15·1	**37·3**	18·0	21·5	0·2	0·6	**99·2**	0·4	0·4	**99·2**
		4	21·9		0·2	2·8	**50·4**	12·6	12·1	19·4	5·5	**75·1**	19·8	3·6	**76·6**
		2·5	**92·2**				4·0	1·7	2·1	**85·4**	7·7	6·9	**86·3**	6·2	7·5
Gaussian	10	10				**92·1**	7·9					**100·0**			
		5		0·1		**55·4**	44·5					**100·0**			
		4	0·1	1·5	1·6	**48·8**	47·8	0·1	0·1	0·1	0·1	**99·8**			
		3	4·0	5·2	9·6	35·7	**40·9**	2·0	2·6	3·6	1·1	**95·3**			
		2	**56·8**	1·5	8·5	7·7	15·2	5·6	4·7	**52·3**	6·5	41·2			
		1	**95·5**	0·1	0·9	0·6	1·4	0·8	0·7	**91·9**	3·9	4·2			
	30	7				**87·1**	12·9					**100·0**			**100·0**
		5				**61·2**	38·8					**100·0**			**100·0**
		4				**52·8**	47·2					**100·0**			**100·0**
		3			1·4	48·1	**49·1**	0·6	0·8			**100·0**			**100·0**
		2	9·7		0·7	6·2	**57·6**	13·3	12·5	8·4	3·2	**88·4**	8·7	2·1	**89·2**
		1	**95·8**				2·2	0·9	1·1	**89·8**	6·4	3·8	**90·3**	5·5	4·2

Power	10	0·98	5·7	**49·2**	15·9	6·6	3·1	13·9	5·6	4·5	3·4	**92·1**			
		0·95	7·4	**41·6**	17·1	7·7	3·6	15·6	7·0	5·5	4·1	**90·4**			
		0·9	10·7	**29·4**	18·3	9·2	4·8	18·2	9·4	8·8	5·4	**85·8**			
		0·8	**25·0**	12·8	17·6	11·1	6·3	16·3	10·9	20·7	8·0	**71·3**			
		0·35	**90·6**	0·3	1·9	1·2	2·9	1·6	1·5	**85·3**	5·8	8·9			
		0·2	**94·6**	0·2	1·1	0·7	1·7	0·9	0·8	**90·9**	4·2	4·9			
	30	0·98		37·0	4·9	0·1	0·1	**52·8**	5·1			**100·0**			**100·0**
		0·95		16·1	5·8	0·3	0·1	**69·6**	8·1			**100·0**			**100·0**
		0·9		1·7	6·7	0·8	0·5	**75·8**	14·5		0·2	**99·8**		0·1	**99·9**
		0·8	0·1		5·6	5·6	2·1	**59·9**	26·7		1·2	**98·8**	0·1	0·5	**99·4**
		0·35	**81·9**			0·1	8·6	4·5	4·9	**73·2**	10·7	16·1	**74·8**	8·0	17·2
		0·2	**94·1**				2·9	1·3	1·7	**87·4**	7·3	5·3	**88·2**	6·0	5·8
Basis	10		**96·1**	0·1	0·7	0·4	1·3	0·8	0·6	**92·5**	3·7	3·8			
	30		**96·9**				1·7	0·6	0·8	**91·4**	5·7	2·9	**91·9**	5·1	3·0

with a rejection rate of 0·131 (bias of S.E.D. –21·6%). For the *F*-test, the rejection rate ranged between 0 and 0·6292.

For *t*=10 the empirical Type I error rates of the *t*-test ranged from 0·001 (bias equal to 87·6%) to 0·095 (bias equal to –14·4%) and for the *F*-test from 0 to 0·198.

Independently of the simulation model, analysis by the B-RCB model using a *t*-test controlled the Type I error for both *t*=10 and 30. The same was true for the *F*-test and *t*=30, while for *t*=10, this was only the case for weaker correlations (Table 5). For higher correlation the tests were conservative. Analysis of incomplete block designs according to B-IB generally resulted in good control of the nominal Type I error rate for the *t*-test with KR (=KR1) (Table 5). For the *F*-test with *t*=10, this tended to be the case for intermediate correlation, while good control was generally achieved for *t*=30.

The following section discusses to what degree the best-fitting model controlled the Type I error rate and will also investigate whether the same adjustment method (SW, KR or KR1) can be used for all analysis models. If one considers the three analysis models SM (sm), GM (gm) and PM (pm) and the three methods SW, KR and KR1, then there are $3^3=27$ possible choices for analysis. Checking all 27 possibilities, it was found that with nearly all simulated cases the Type I error rate was controlled well by the best-fitting model, if for all analysis models the KR1 method was used. In these analyses, the number of effective runs under KR1 was equal or close to 10 000 in all cases (minimal effective number of runs was 9982 for *t*=30 and the power model with $\rho=0·98$). With *t*=10 and simulation of the spherical or power model, the best-fitting models showed slightly excessive Type I error rates for both the *t*- and *F*-tests (Table 5). The maximum error rates were obtained for intermediate correlations. With the Gaussian model, better results were achieved for the *t*-test, which, again, tended to become liberal for intermediate correlations. The *F*-test showed both liberal and conservative behaviour, and in no case was the Type I error rate controlled at the nominal level. For *t*=30, Type I error control of the *t*-test was very good, and only slightly liberal behaviour was observed for spherical and power models when correlation was intermediate, but this was not as pronounced as for *t*=10. For the *F*-test, results for *t*=30 with the spherical and power models were better than for *t*=10, but again not fully satisfactory; results for the Gaussian model were often worse.

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

Table 4. *Bias of estimates of mean standard error of a treatment difference* (S.E.D.) (%) *in designs with complete blocks for all combinations of simulation and analysis model for selected values of range parameters. Control of the nominal Type I error rate with approximated denominator* D.F. *for t-test* □ *and for t- and F-tests* ⧈. *Approximation based on Satterthwaite (1941) – SW, on Kenward & Roger (1997) – KR and on Kenward & Roger (2009) – KR1*

		Simulation model																			
		Spherical							Gaussian							Power					
		t = 10			*t* = 30				*t* = 10			*t* = 30				*t* = 10			*t* = 30		
Analysis model	Range	SW	KR	KR1	SW	KR	KR1	Range	SW	KR	KR1	SW	KR	KR1	Range	SW	KR	KR1	SW	KR	KR1
SM	600	$\boxed{-1·2}$	$\boxed{-1·1}$	$\boxed{\boxed{-0·8}}$	$\boxed{\boxed{-0·5}}$	$\boxed{\boxed{-0·4}}$	$\boxed{\boxed{-0·4}}$	10/7*	86·3	87·2	87·6	66·2	65·9	66·4	0·98	$\boxed{-1·9}$	−1·7	$\boxed{-1·4}$	$\boxed{-0·9}$	$\boxed{-0·9}$	$\boxed{-0·9}$
	20	$\boxed{-1·2}$	$\boxed{-0·8}$	$\boxed{\boxed{0·2}}$	$\boxed{0·6}$	$\boxed{\boxed{0·5}}$	$\boxed{0·6}$	5	30·4	31·4	32·0	36·5	36·2	36·5	0·90	−4·5	−4·3	−4·0	−3·8	−3·7	−3·8
	6	−5·6	−5·6	−5·3	−6·5	−6·4	−6·5	3	2·4	2·8	3·2	3·1	3·4	3·1	0·80	−7·3	−7·1	−6·8	−7·6	−7·4	−7·6
	4	−9·3	−9·6	−9·2	−12·5	−12·5	−12·5	2	−8·3	−8·5	−8·2	−11·0	−11·0	−11·0	0·35	−12·8	−13·2	−12·8	−17·0	−17·0	−17·0
	Basis	−14·5	−14·9	−14·4	−18·6	−18·7	−18·6	Basis	−14·5	−14·9	−14·4	−18·6	−18·7	−18·6	Basis	−14·5	−14·9	−14·4	−18·6	−18·7	−18·6
sm	600	$\boxed{-0·5}$	−2·0	$\boxed{\boxed{0·8}}$	$\boxed{-0·5}$	$\boxed{-0·9}$	$\boxed{\boxed{-0·3}}$	10/7	88·6	85·9	89·4	64·7	64·8	65·4	0·98	$\boxed{-1·6}$	−3·1	$\boxed{-0·3}$	−1·7	−2·0	$\boxed{-1·4}$
	20	$\boxed{-1·5}$	−2·8	$\boxed{\boxed{0·0}}$	$\boxed{-1·6}$	$\boxed{-1·6}$	$\boxed{-1·1}$	5	31·4	30·9	33·4	34·0	34·1	34·9	0·90	−5·4	−6·3	−3·6	−5·7	−5·6	−5·1
	6	−7·0	−6·8	−4·9	−5·6	−5·5	−5·3	3	1·9	$\boxed{\boxed{2·1}}$	4·5	$\boxed{0·9}$	$\boxed{1·1}$	1·6	0·80	−8·5	−8·7	−6·6	−8·9	−8·7	−8·3
	4	−7·1	−7·0	−6·0	$\boxed{-0·7}$	$\boxed{-0·7}$	$\boxed{-0·7}$	2	−7·5	−7·5	−6·2	−1·5	−1·4	−1·4	0·35	−5·3	−5·4	−4·7	$\boxed{\boxed{-0·2}}$	$\boxed{\boxed{-0·2}}$	$\boxed{\boxed{-0·2}}$
	Basis	−4·1	−4·1	−3·7	$\boxed{\boxed{-0·2}}$	$\boxed{\boxed{-0·2}}$	$\boxed{\boxed{-0·2}}$	Basis	−4·1	−4·1	−3·7	$\boxed{\boxed{-0·2}}$	$\boxed{\boxed{-0·2}}$	$\boxed{\boxed{-0·2}}$	Basis	−4·1	−4·1	−3·7	$\boxed{\boxed{-0·2}}$	$\boxed{\boxed{-0·2}}$	$\boxed{\boxed{-0·2}}$
GM	600	−15·8	−11·5	−10·2	−22·8	−22·2	−21·6	10/7	$\boxed{\boxed{-5·6}}$	$\boxed{\boxed{-2·6}}$	0·7	$\boxed{-1·2}$	$\boxed{-1·9}$	$\boxed{0·0}$	0·98	−14·9	−10·4	−9·1	−20·2	−19·5	−19·0
	20	−14·8	−10·3	−9·0	−16·1	−15·2	−14·7	5	$\boxed{-2·7}$	$\boxed{\boxed{-2·0}}$	2·6	$\boxed{-0·7}$	$\boxed{-1·3}$	$\boxed{0·4}$	0·90	−12·3	−6·9	−6·0	−11·9	−10·7	−10·2
	6	−8·9	−1·9	−1·7	−4·3	−2·3	−1·9	3	−5·8	$\boxed{-0·3}$	1·9	−2·2	$\boxed{-0·7}$	$\boxed{0·2}$	0·80	−9·9	−3·5	−3·3	−6·2	−4·5	−4·1
	4	−7·1	$\boxed{\boxed{1·6}}$	$\boxed{0·1}$	−2·6	$\boxed{\boxed{0·0}}$	$\boxed{0·1}$	2	−7·4	$\boxed{1·1}$	$\boxed{0·1}$	−2·7	$\boxed{-0·2}$	$\boxed{-0·1}$	0·35	−5·4	$\boxed{3·8}$	$\boxed{\boxed{-0·1}}$	−2·1	$\boxed{0·6}$	$\boxed{\boxed{-0·2}}$
	Basis	−4·0	$\boxed{4·3}$	$\boxed{\boxed{0·1}}$	$\boxed{-1·3}$	$\boxed{0·8}$	$\boxed{\boxed{-0·1}}$	Basis	−4·0	$\boxed{4·3}$	$\boxed{\boxed{0·1}}$	$\boxed{-1·3}$	$\boxed{0·8}$	$\boxed{\boxed{-0·1}}$	Basis	−4·0	$\boxed{4·3}$	$\boxed{\boxed{0·1}}$	$\boxed{-1·3}$	$\boxed{0·8}$	$\boxed{\boxed{-0·1}}$
gm	600	−15·6	−11·3	−10·0	−22·4	−21·8	−21·2	10/7	−6·1	−3·8	$\boxed{\boxed{0·4}}$	$\boxed{-1·0}$	$\boxed{-1·8}$	$\boxed{0·1}$	0·98	−14·8	−10·3	−9·0	−19·9	−19·2	−18·7
	20	−14·7	−10·2	−8·9	−16·0	−15·1	−14·6	5	$\boxed{-2·8}$	$\boxed{\boxed{-2·2}}$	2·5	$\boxed{-0·7}$	$\boxed{-1·4}$	$\boxed{0·3}$	0·90	−12·2	−6·9	−6·0	−11·8	−10·6	−10·1
	6	−8·9	−1·9	−1·7	−4·3	−2·3	−1·9	3	−5·8	$\boxed{-0·3}$	1·9	−2·2	$\boxed{-0·7}$	$\boxed{0·2}$	0·80	−9·9	−3·5	−3·2	−6·2	−4·5	−4·1
	4	−7·1	$\boxed{\boxed{1·6}}$	$\boxed{0·1}$	−2·6	$\boxed{\boxed{0·0}}$	$\boxed{-0·1}$	2	−7·4	$\boxed{\boxed{1·1}}$	$\boxed{0·1}$	−2·7	$\boxed{-0·2}$	$\boxed{-0·1}$	0·35	−5·4	$\boxed{3·8}$	$\boxed{\boxed{-0·1}}$	−2·1	$\boxed{0·6}$	$\boxed{\boxed{-0·2}}$
	Basis	−4·0	$\boxed{4·3}$	$\boxed{\boxed{0·1}}$	$\boxed{-1·3}$	$\boxed{0·8}$	$\boxed{\boxed{-0·1}}$	Basis	−4·0	$\boxed{4·3}$	$\boxed{\boxed{0·1}}$	$\boxed{-1·3}$	$\boxed{0·8}$	$\boxed{\boxed{-0·1}}$	Basis	−4·0	$\boxed{4·3}$	$\boxed{\boxed{0·1}}$	$\boxed{-1·3}$	$\boxed{0·8}$	$\boxed{\boxed{-0·1}}$
PM	600	$\boxed{1·8}$	x†	3·9	$\boxed{1·3}$	54·5	$\boxed{1·4}$	10/7	78·6	x	78·9	66·5	98·6	66·7	0·98	1·9	x	4·1	$\boxed{1·1}$	51·1	$\boxed{1·2}$
	20	2·8	x	4·7	2·9	35·4	3·1	5	31·2	x	32·4	37·4	63·9	37·7	0·90	$\boxed{0·1}$	94·4	2·8	$\boxed{\boxed{0·0}}$	14·5	$\boxed{\boxed{0·4}}$
	6	$\boxed{-0·6}$	22·5	3·7	$\boxed{1·7}$	4·7	2·9	3	6·3	44·1	9·4	8·4	12·3	9·3	0·80	$\boxed{\boxed{-2·1}}$	38·7	$\boxed{\boxed{1·6}}$	$\boxed{\boxed{-0·5}}$	3.4	$\boxed{\boxed{0·2}}$
	4	−3·8	10·6	$\boxed{\boxed{1·3}}$	$\boxed{\boxed{-1·2}}$	$\boxed{1·9}$	$\boxed{0·8}$	2	−3·1	12·4	$\boxed{2·0}$	$\boxed{-0·6}$	2·5	$\boxed{1·2}$	0·35	−4·1	8·2	$\boxed{0·2}$	−1·9	$\boxed{1·2}$	$\boxed{\boxed{-0·1}}$
	Basis	−3·3	7·4	$\boxed{0·0}$	$\boxed{-1·2}$	$\boxed{1·1}$	$\boxed{\boxed{-0·1}}$	Basis	−3·3	7·4	$\boxed{0·0}$	$\boxed{-1·2}$	$\boxed{1·1}$	$\boxed{\boxed{-0·1}}$	Basis	−3·3	7·4	$\boxed{0·0}$	$\boxed{-1·2}$	$\boxed{1·1}$	$\boxed{\boxed{-0·1}}$
pm	600	$\boxed{2·0}$	x	4·8	$\boxed{1·5}$	39·7	$\boxed{1·7}$	10/7	65·4	x	66·0	64·5	91·5	64·7	0·98	$\boxed{1·7}$	x	4·5	$\boxed{1·3}$	35·9	$\boxed{1·5}$
	20	2·5	x	5·0	2·8	20·0	3·0	5	30·0	x	31·4	35·6	51·3	35·9	0·90	$\boxed{-0·2}$	90·7	2·9	$\boxed{\boxed{-0·1}}$	9·4	$\boxed{\boxed{0·3}}$
	6	$\boxed{-1·0}$	23·7	3·2	$\boxed{1·5}$	4·4	2·7	3	5·5	50·9	8·7	8·0	11·5	8·8	0·80	$\boxed{\boxed{-2·4}}$	43·2	$\boxed{1·5}$	$\boxed{\boxed{-0·7}}$	2·9	$\boxed{\boxed{0·1}}$
	4	−3·6	10·3	$\boxed{\boxed{1·2}}$	$\boxed{\boxed{-1·0}}$	$\boxed{1·8}$	$\boxed{0·7}$	2	−3·1	13·2	$\boxed{1·7}$	$\boxed{-0·4}$	2·3	$\boxed{1·1}$	0·35	−4·0	8·2	$\boxed{0·2}$	−1·8	$\boxed{1·1}$	$\boxed{\boxed{-0·1}}$
	Basis	−3·2	7·3	$\boxed{0·0}$	$\boxed{-1·2}$	$\boxed{1·1}$	$\boxed{\boxed{-0·1}}$	Basis	−3·2	7·3	$\boxed{0·0}$	$\boxed{-1·2}$	$\boxed{1·1}$	$\boxed{\boxed{-0·1}}$	Basis	−3·2	7·3	$\boxed{0·0}$	$\boxed{-1·2}$	$\boxed{1·1}$	$\boxed{\boxed{-0·1}}$

* Range = 10 for *t* = 10; range = 7 for *t* = 30.

† x: bias > 100%.

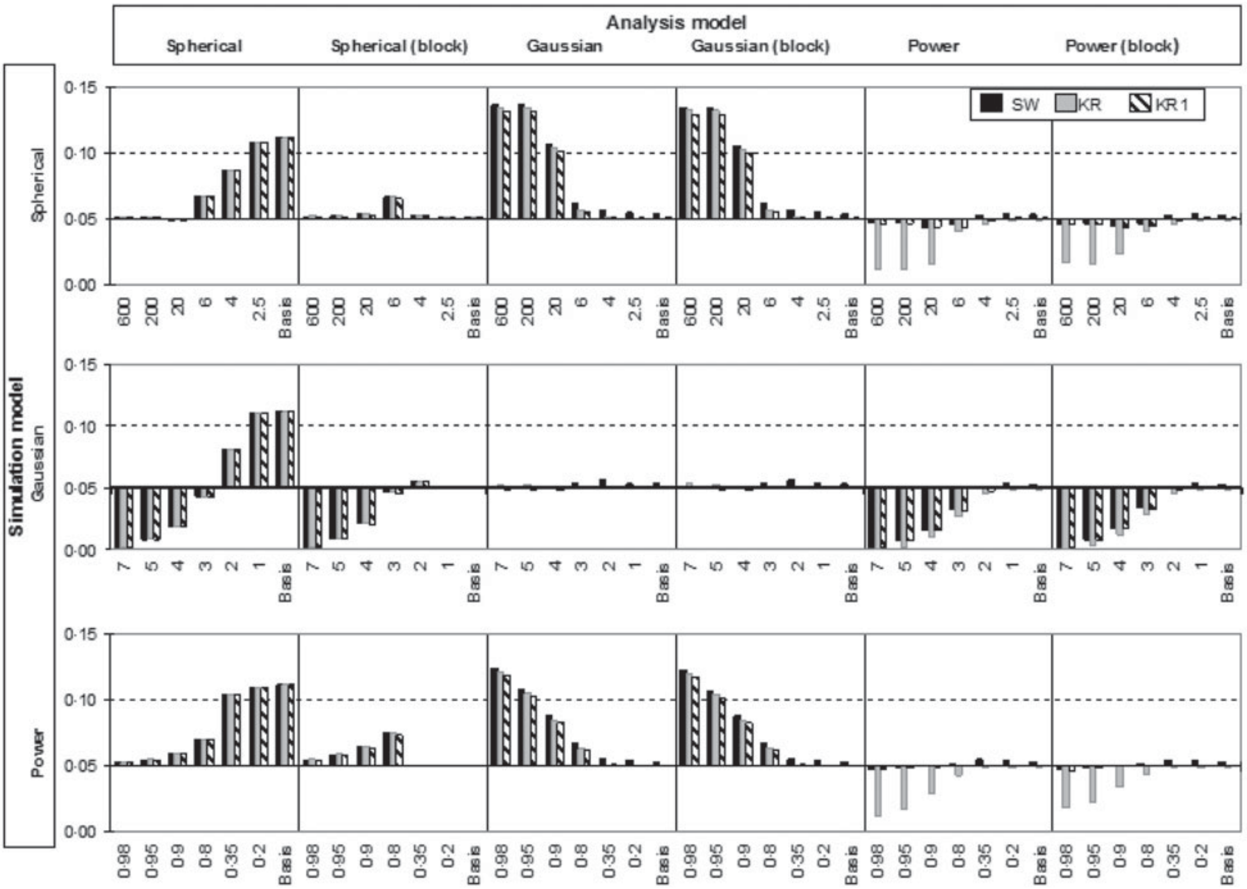


Fig. 2. Empirical Type I error rates of *t*-test with three adjusting methods (SW, KR and KR1) for $t=30$ treatments.

Power analysis for *t*-test

Power analysis is reported in detail for $t=30$ and KR1, because in this case there was usually satisfactory Type I error rate control when using the best-fitting model (Fig. 4). When there was not a sufficient control, power was computed using significance thresholds derived from the simulations under the null hypothesis (these cases are marked as 'corrected' in Fig. 4).

For B-RCB the power was generally the smallest and it was the highest for the best-fitting model. Analysis by B-IB was intermediate between these two extremes, and power was higher for $k=5$ than for $k=10$. The larger the spatial correlation, the more marked was the power gain from spatial analysis and also of an analysis by B-IB. The most pronounced gains were found when data were simulated by the Gaussian model, followed by the power and spherical models. For $t=10$, in most cases only the corrected power could be interpreted, because control of the Type I error rate was not acceptable. The same ranking of analysis models was found, but the differences were smaller: the spatial models had slightly lower power and B-RCB a slightly higher power than for $t=30$.

DISCUSSION

Role of rejection rules for simulation runs

The rejection rules used in the current analysis became effective with varying frequency for the different combinations of simulation and analysis model as well as the approximation methods SW, KR and KR1. In some cases, they had a pronounced effect on the reduction of bias of S.E.D. estimates and the Type I error control. It is useful to check for an obtained model fit whether any of the criteria (i)–(iv) is met, because these identify situations that are statistically questionable. This can be illustrated for criterion (iv), which played an important role for analyses by PM and pm, when the estimate of ρ fell close to the boundary of the parameter space. When $\hat{\rho}$ is close to 1 in the power model, this implies a strong correlation, while $\hat{\rho} = 1$ implies a confounding of the spatial component with the block effect, corresponding to a lack of spatial

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

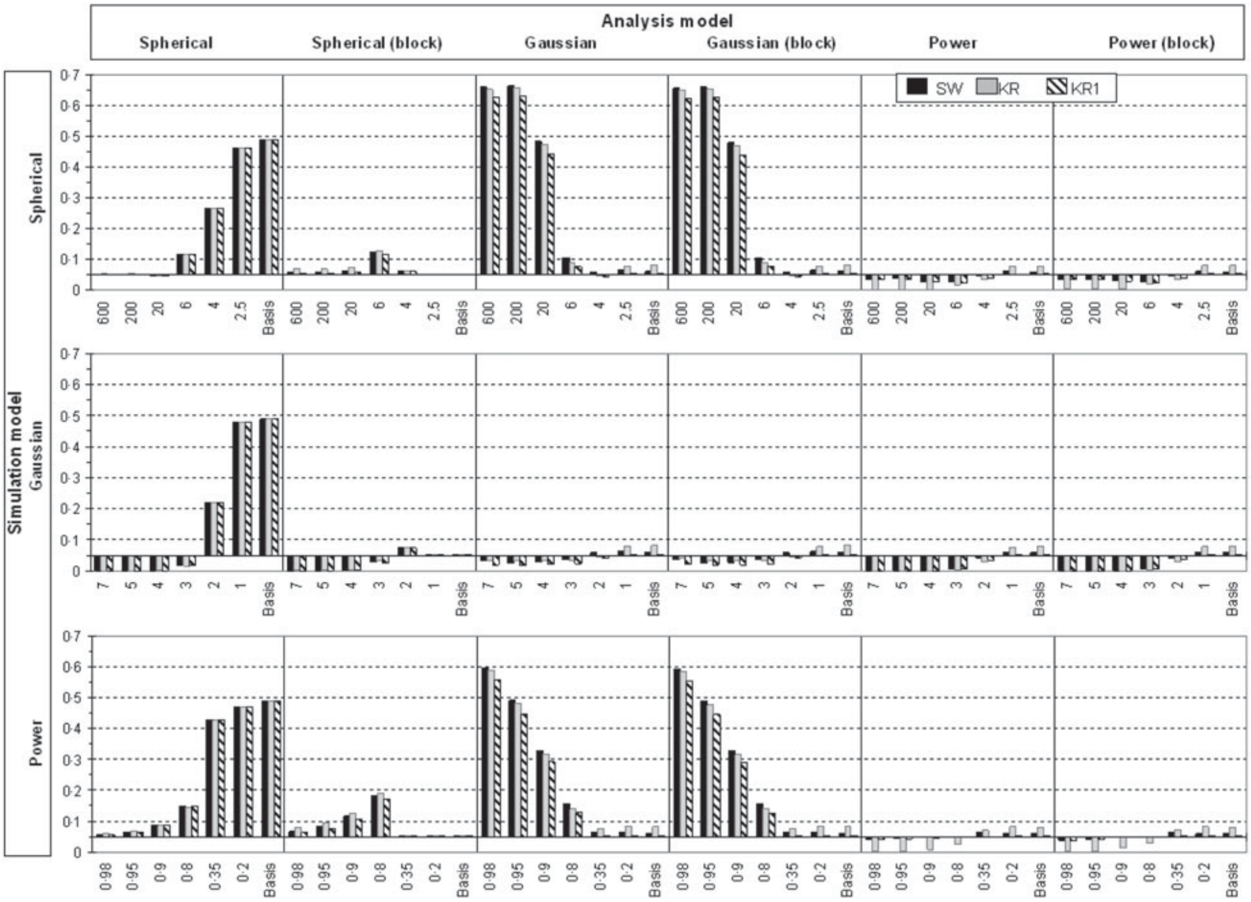


Fig. 3. Empirical Type I error rates of *F*-test with three adjusting methods (SW, KR and KR1) for $t = 30$ treatments.

correlation within blocks (Piepho *et al.* 2008). The decision in favour of one of these two opposing possibilities depends on the numerical precision, but has contrasting interpretations. The effect of this criterion was particularly important, when stronger correlations were present and analysis was by PM or pm with KR: For all simulation models the originally observed bias was larger than 10^4% with the maximal range, but dropped to values between 130 and 450% for t =10 and 36 to 92% for $t = 30$. Also with KR1 substantial bias reductions were observed. These observations were the motivation to give preference to sm or SM in case of a tied first rank between PM and SM or between pm and sm according to AICC and LRT (see the Section on *simulations*). These ties occurred mostly when $\hat{\rho} = 1$. The current simulations did not deviate from the default settings of the optimization routines in order not to rule out *a priori* any potentially critical cases that would otherwise have gone unnoticed. When analysing by PM or pm, then independently of the simulation model the criterion (iv) was often in agreement with criteria (i)–(iii). Based on the current results it can be recommended to impose the boundary constraint $|\hat{\rho}| < 0{\cdot}9999$ during optimization of the restricted log-likelihood.

In the current assessment of KR1, the effect of replacing a best-fitting model that was excluded according to our criteria with the best-fitting model that passed all criteria was checked. Since there were only 18 rejections, their replacement had virtually no effect on the distribution of the best-fitting models.

Bias of estimates of standard errors of a difference and of parameters of the covariance structure

The degree of positive or negative bias for S.E.D. depends on the combination of simulation and analysis model and on the value of the range parameter. The most extreme biases occur when the simulation model has specific features that cannot be captured by the analysis model. Two such features will be considered in more detail.

(1) While differences in bias are relatively small between PM and pm as well as between GM and gm

Table 5. *Empirical Type I error rates for the best-fitting models using the KR1 method compared with analysis according to B-RCB and B-IB for all simulation models. Bold faced: nominal Type I error rate of 0·05 not controlled according to binomial test*

		t = 10						*t* = 30							
		Complete block designs				Incomplete block designs		Complete block designs				Incomplete block designs			
		Analysis model													
		B-RCB		Best-fitting model*		B-IB *k* = 5		B-RCB		Best-fitting model*		B-IB *k* = 5		B-IB *k* = 10	
Simulation model	Range	*t*-test	*F*-test	*t*-test	*F*-test	*t*-test†	*F*-test†	*t*-test	*F*-test	*t*-test	*F*-test	*t*-test†	*F*-test†	*t*-test†	*F*-test†
Spherical	600	0·049	**0·042**	**0·062**	**0·072**	0·049	**0·044**	0·050	0·049	0·050	0·048	0·050	0·051	0·050	0·050
	200	0·049	**0·042**	**0·062**	**0·073**	0·049	**0·045**	0·049	0·050	0·050	0·049	0·050	0·050	0·051	0·050
	20	0·049	**0·043**	**0·064**	**0·081**	0·049	0·047	0·050	0·047	0·047	**0·042**	0·050	0·050	0·050	0·050
	6	0·049	0·049	**0·071**	**0·088**	0·049	0·053	0·050	0·048	**0·056**	**0·076**	0·050	0·051	0·050	0·050
	4	0·049	0·049	**0·065**	**0·079**	0·050	**0·055**	0·050	0·047	**0·055**	**0·063**	0·051	0·052	0·050	0·049
	2·5	0·050	0·050	**0·055**	**0·066**	0·050	**0·055**	0·050	0·050	0·052	**0·061**	0·050	0·052	0·050	0·050
Gaussian	10/7‡	0·046	**0·039**	0·047	**0·042**	0·047	**0·042**	0·050	0·051	0·050	**0·021**	0·051	0·048	0·050	**0·044**
	5	0·048	**0·041**	0·046	**0·032**	0·048	**0·042**	0·050	0·049	0·050	**0·022**	0·051	0·050	0·049	**0·043**
	4	0·049	**0·045**	0·046	**0·028**	0·047	**0·044**	0·050	0·050	0·049	**0·021**	0·050	0·048	0·050	0·049
	3	0·049	0·047	0·054	**0·044**	0·049	0·046	0·050	0·051	0·050	**0·024**	0·050	0·048	0·050	0·051
	2	0·049	0·050	**0·067**	**0·082**	0·050	0·053	0·050	0·047	0·054	**0·056**	0·051	0·051	0·050	0·055
	1	0·050	0·050	0·054	**0·063**	0·050	**0·055**	0·050	0·050	0·051	**0·058**	0·050	0·052	0·050	0·050
Power	0·98	0·049	**0·042**	**0·064**	**0·080**	0·049	**0·045**	0·049	0·051	0·050	0·052	0·050	0·050	0·051	0·048
	0·95	0·049	**0·043**	**0·067**	**0·085**	0·049	0·047	0·050	0·049	0·051	**0·055**	0·050	0·049	0·050	0·050
	0·9	0·049	0·046	**0·070**	**0·093**	0·049	0·047	0·050	0·047	0·053	**0·063**	0·050	0·049	0·050	0·051
	0·8	0·049	0·047	**0·073**	**0·100**	0·050	0·049	0·050	0·051	**0·056**	**0·075**	0·050	0·048	0·050	0·050
	0·35	0·050	0·050	**0·058**	**0·070**	0·050	**0·055**	0·050	0·050	0·054	**0·067**	0·050	0·052	0·050	0·050
	0·2	0·050	0·050	**0·055**	**0·064**	0·050	**0·055**	0·050	0·050	0·052	**0·060**	0·050	0·052	0·050	0·049
Basis		0·050	0·051	0·054	**0·062**	0·050	**0·055**	0·050	0·050	0·051	**0·057**	0·050	0·053	0·050	0·050

* Selected from the models B-RCB, SM, sm, GM, gm, PM and pm.

† Using the KR method (= KR1 method).

‡ Range = 10 for *t* = 10; range = 7 for *t* = 30.

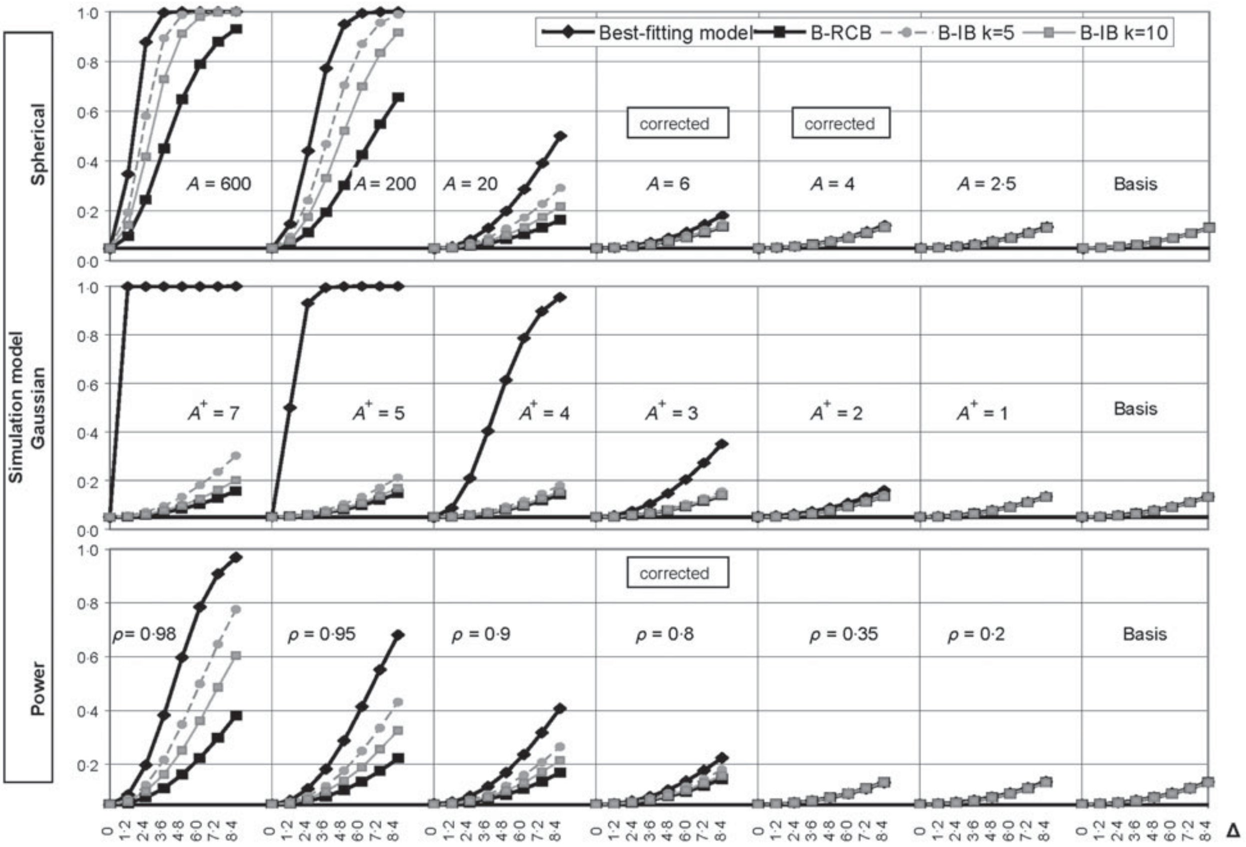


Fig. 4. Power of the *t*-test as a function of treatment difference Δ for $t=30$ treatments. Analysis of simulated models according to B-RCB, B-IB and best-fitting model using the KR1 method.

with all simulation models, the differences were more marked between SM and sm in case of modest correlation. In addition, when analysing by PM, pm, GM, gm and sm, the bias tended to decrease with decreasing correlation. This was not the case when analysing by SM. These observed differences can be explained as follows: Estimation by REML requires that the first two partial derivates of the matrix $\boldsymbol{V}$ with respect to the variance parameters exist. Kenward & Roger (1997, 2009) pointed this out explicitly for KR, KR1 and KR2. Differentiability holds for the Gaussian and power models for distances $d>0$, but not for the spherical model at $d=A$. Thus, biases and/or convergence problems are to be expected for the spherical model whenever the range A is smaller than the maximal distance within a block or the whole experiment. As is well known for the exponential model (Zhang & Zimmerman 2005), the parameters of the spherical model cannot be estimated without bias and they cannot be meaningfully interpreted. Irvine *et al.* (2007) showed for the exponential model that the ratio of REML-estimators of the two parameters is nearly unbiased with respect to the true ratio. The current results showed the same for the spherical model in the case of strong spatial correlation and analysis according to SM or sm. This means that for the exponential and spherical models the estimated covariance function may be expanded or compressed along the distance axis relative to the true covariance function. When analysing by SM, a pronounced expansion was observed, meaning that the critical point $d=\mathrm{A}$ tended to be larger than the largest distance within the trial area. But this did not allow representation of the weak correlations present for $\mathrm{A}\leqslant 6$. When analysing by sm, decreasing values of A led to stronger compression, such that at $\mathrm{A}\leqslant 4$ in more than 61% ($t=10$) and more than 97% ($t=30$) of simulation runs, the range parameter was estimated as zero and the estimate of σ^2 was equal to that for B-RCB. Since there is no bias for B-RCB, bias was also reduced for sm compared with SM and bias vanishes for $t=30$. This behaviour of SM and sm occurred irrespective of the simulation model and the approximation method. In order to also allow a transition to the baseline model when using SM, it is useful to add a nugget effect. To explore this point, a nugget for $t=10$ was exemplarily

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

added, simulating from the spherical and Gaussian models with weak correlation and analysing according to SM. Addition of the nugget substantially reduced the bias for S.E.D., and with KR1 the *t*-test controlled the nominal Type I error rate. When simulating from the baseline model, for the *F*-test the empirical rejection rate went down from 0·198 to 0·057. This positive effect of inclusion of a nugget effect could be demonstrated for analysis by SM only, and the improvement seems to depend on the critical point d=A and the dimensions of the experimental field. In principle, one can generally include a nugget effect σ_N^2 in case of weak (or zero) correlations. The analysis model would then be $\mathrm{Cov}(d)=\sigma_N^2+\sigma^2 \bullet f(d)$ and would coincide with the baseline model in case of $\sigma^2=0$. When correlation is high, presence of a nugget effect often causes convergence problems.

The consistent estimation of the ratio of variance and range parameters observed for the exponential model by Irvine *et al.* (2007) could not be found for the power model, because estimates $\hat{\rho} \leqslant 0$ and $\hat{\rho}=1$ frequently occurred, meaning that the estimated covariance model did not always coincide with that for the exponential model.

(2) In contrast to the other two models, the Gaussian model has a point of inflexion, located at the spatial distance $d_{\mathrm{IP}}=A^+/\surd 2$. When analysing by either GM or gm, the mean of observed estimates of A^+ and σ^2 and their ratio were close to the true values when the correlation was high, whereas A^+ was slightly overestimated for weak correlations.

When data were generated from the power or spherical models, both of which do not have a point of inflexion, analysis by GM or gm shifted the inflexion point in such a way, that d_{IP} was within the smallest distance between two plots, or nearly so. Even for strong correlation the maximal estimated ranges were smaller than the shortest distance to the neighbouring block (8 m), such that results for GM and gm were nearly identical. Estimates of σ^2 were slightly above the value obtained for the analysis according to B-RCB, applying the equation $\hat{\sigma}^2_{\mathrm{B-RCB}}=\sigma^2-\overline{\mathrm{Cov}(d)}_{\mathrm{CB}}$, where $\overline{\mathrm{Cov}(d)}_{\mathrm{CB}}$ is the average covariance among plots in the same complete block according to the simulation model. The excess covariance compared to this average covariance was captured by GM and gm, leading to the negative bias observed.

Conversely, when data were simulated by the Gaussian model, then with stronger correlation, the models SM, sm, PM or pm were not able to represent the sigmoidal shape of the covariance function. With larger correlation this led to a high positive bias. With $A^+<2{\cdot}8$ the value of d_{IP} is smaller than 2. In this case, biases were of the same order of magnitude as with analysis of the spherical model by SM or sm and of the power model by PM or pm when correlations were comparable.

In comparison to the general tendencies, the differences between the three approximations were minor. Hu *et al.* (2006) compared the SW- and KR-methods for specific contrasts exploiting the correlation across the whole experimental field. In accordance with the present results for the power model, Hu *et al.* (2006) found that the exponential model without nugget, combined with the KR method, entailed biases of S.E.D.>100%. Similar results were obtained for the spherical model without nugget, which does not agree with the current results. Based on these findings, Hu *et al.* (2006) recommended the SW method for models with and without nugget. Closer scrutiny revealed that this discrepancy is due to the fact that Hu *et al.* (2006) used the expected information matrix (Fisher scoring), while in the current work the observed information matrix (Newton–Raphson algorithm) was used for maximizing the residual log-likelihood. Hu *et al.* (2006) preferred Fisher scoring because this led to better convergence behaviour than the Newton–Raphson algorithm when their simulation and analysis models contained a nugget effect.

For the analysis by B-RCB and B-IB, a bias is virtually non-existent, due to the averaging over 10 000 randomizations, with the KR (=KR1=KR2 in this case) showing slight advantages compared to SW.

Relation of model selection and control of Type I error rate

Except for the spherical model where a nugget was recommended, the Type I error rate of the *t*-test was controlled by at least one of the three methods when simulation and analysis model were the same. The fact that this was also the case when the spatial correlation was low and simulation and analysis model did not agree can be explained by a frequent convergence of these models to the baseline model. Scrutiny of the best-fitting model showed in all cases, that the empirical Type I error was largest for intermediate values of the spatial correlation, sometimes even significantly larger than the nominal rate of 0·05. In view of this finding, the LRT was also considered for model selection when models were nested, which

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

favours the baseline model somewhat more often than AICC. The expectation was that this would improve Type I error control, but Table 5 shows that this was not always the case. This raises the question if AICC should be replaced with AIC or BIC. The penalty term of BIC is defined as $q \times \log(n)$, where q is the number of variance parameters and n is the 'sample size'. In mixed models it is not obvious, however, what the sample size is, because observations are not stochastically independent (Pauler 1998), and implementations in software often use rather *ad hoc* definitions. For example, when data are modelled to be independent between blocks, then the MIXED procedure of SAS sets $n = r$ for the spatial models, where r is the number of replicates, while for B-RCB it sets $n = (t-1)(r-1)$, where t is the number of treatments, as well as for spatial models where correlation extends across blocks. The current simulations show that with these definitions, in almost all cases one of the models sm, pm or gm was selected when using BIC for model selection. Even when simulating from the baseline model, this was not generally identified as the best-fitting model, which adversely affected the empirical Type I errors of the best-fitting model. The AIC tended to favour spatial models more often than AICC.

As already reported by Richter & Kroschewski (2012) for uniformity trials with $t = 10$, the *F*-test did not control the nominal level well.

Power analyses

With analysis according to B-RCB, the average covariance within blocks is larger for $t = 10$ than for $t = 30$ for all simulated spatial models, meaning that $\hat{\sigma}^2_{\mathrm{B-RCB}}$ was smaller, on average, for $t = 10$. Although the error degrees of freedom for $t = 10$ are smaller, the power was larger than for $t = 30$. Only when the spatial correlation is weak (spherical model for $A \leqslant 4$, Gaussian model for $A^+ \leqslant 2$, power model for $\rho \leqslant 0{\cdot}35$), do the larger degrees of freedom for $t = 30$ lead to a slightly elevated power (+0·05) compared with $t = 10$.

Estimates of σ^2 based on B-IB were approximately equal to $\sigma^2 - \overline{\mathrm{Cov}}(d)_{\mathrm{IB}}$ as expected, where $\sigma^2 = 200$ and $\overline{\mathrm{Cov}}(d)_{\mathrm{IB}}$ is the mean covariance of plots in the same incomplete block under the assumed simulation model. From this it emerges that for $t = 30$ power is expected to be larger with $k = 5$ than with $k = 10$, and that for B-IB power is generally larger than for B-RCB. Comparing the ratio $\hat{\sigma}^2_{\mathrm{B-IB}}/\hat{\sigma}^2_{\mathrm{B-RCB}}$ with the efficiency of the incomplete block design under study, one can determine the point starting from which value of the range parameter, for given σ^2 and fixed plots and block sizes and arrangements, a power gain is to be expected from incomplete blocking as compared to complete blocking. With the three incomplete block designs considered in the present study, about the same results were obtained, because efficiency factors do not differ much. The correlations between neighbouring plots ($d = 2$) should be at least 0·52, 0·53 and 0·42, which is relatively high. This corresponds to range parameter values $A \geqslant 6$, $A^+ \geqslant 2{\cdot}5$ and $\rho \geqslant 0{\cdot}65$, respectively. These values agree roughly with those depicted in Fig. 4. When correlations become slightly smaller, the spatial analyses are still at an advantage compared with B-RCB and B-IB, but the edge is marginal.

Standard errors of treatment means

Just as for treatment differences, treatment means are also estimated without bias (Harville & Jeske 1992). But there are often severe biases in the estimated S.E.M. that are much larger than for S.E.D. Based on the current simulation results, it may be assumed that while estimates of S.E.D. benefit from the fact that estimates of the ratio of covariance parameters are approximately unbiased, for S.E.M. the absolute values of the parameter estimates are crucial, and these tend to be more strongly biased. Only with simulation according to the Gaussian model and analysis by GM or gm was the bias small for S.E.M., because covariance-parameter estimators are nearly unbiased. This means that computation of confidence limits for treatment means, as often found in the literature, is not to be recommended, while confidence limits and tests for differences are reliable.

Preliminary assessment of GLIMMIX compared with MIXED and second-order Kenward–Roger method compared with first-order Kenward–Roger method

Detailed results can be found in the Supplemental Material (available at: http://http://journals.cambridge.org/AGS): only a brief overview is given here. The comparison of MIXED and GLIMMIX for KR1 showed the influence of the numerical optimization method on convergence behaviour and results. The largest difference was observed with simulation by the Gaussian model for $t = 10$ ($A^+ = 10$) and $t = 30$ ($A^+ = 7$) and analysis by GM and gm. With MIXED and analysis

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

by GM, a maximum of three runs did not converge, while with GLIMMIX 54% ($t=10$) and 66% ($t=30$) of all runs did not converge with the default settings; similar behaviour was found for analysis by gm. Despite these gross differences in the effective number of runs, the bias for S.E.D. was comparable with both procedures. With the other combinations of simulation and analysis model the effective number of runs and the calculated bias was more similar between both procedures; in a few cases, however, differences in bias estimates ranged up to 11·8% ($t=30$, simulation by baseline model and analysis by sm). The reason was that with analysis according to sm there was no automatic transition to the baseline model for analysis as with MIXED. In almost all instances where the bias estimates differed by more than 1%, the bias was larger with GLIMMIX. Regarding the control of the nominal Type I error rate with *F*- and *t*-tests, similar results were obtained with GLIMMIX as those presented in Table 4 for $t=10$. With $t=30$ the control was slightly poorer with GLIMMIX than with MIXED. The comparison of KR1 and KR2 within GLIMMIX shows slight advantages of KR2 in *t*-tests for $t=10$, but much less favourable results in the *F*-test (instead of 13 combinations of simulation and analysis model in case of KR1, only for one combination was there a satisfactory control of the Type I error rate). For $t=30$, results were about the same with both procedures for the *t*-test, whereas for the *F*-test the KR2 method was again inferior.

The occasional differences in convergence behaviour between GLIMMIX and MIXED mean that particularly in the case of stronger correlations, different models were sometimes selected as best models by GLIMMIX and by MIXED. Regarding the Type I error control by KR1, both procedures yielded quite similar results for the best-fitting model. Only in the exceptional cases described above (simulation by the Gaussian model, $t=10$ with $A^{+}=10$ and $t=30$ with $A^{+}=7$) did use of GLIMMIX lead a conservative empirical Type I error rate.

The comparison of KR1 and KR2 in GLIMMIX did not suggest a clear advantage for KR2. This result is somewhat unexpected, because Kenward & Roger (2009) pointed out that the KR2 method is to be preferred over the KR and KR1 method for nonlinear covariance models, which includes spatial covariance models. Further simulations comparing the KR1 and KR2 methods are certainly needed to gain a broader picture. Based on the current results, it is recommended to use the KR1 method in MIXED.

CONCLUSIONS

Conclusions for the example

In the introductory sugar beet example the power model PM was the best-fitting model among those converging; the criteria (i)–(iv) did not hold for the fitted PM model. In contrast to the simulations, the 'true' model is not known here. It is therefore important to make sure that the analysis strategy is such that the best-fitting model controls the nominal Type I error rate. This could be demonstrated for the *t*-test in case of $t=30$ when using the KR1 method, while for $t=10$ the tests tended to be on the liberal side. In the example there were $t=18$ treatments. Hu *et al.* (2006) showed for the SW and KR methods that even $t=20$ leads to a notable improvement compared to $t=10$. It may therefore be assumed that in the example, the control of Type I error rate is also satisfactory for the *t*-test so that results based on the KR1 method are reliable. Often, pairwise *t*-tests are preceded by a global *F*-test. But the *F*-test revealed relatively poor control of Type I error when based on the best-fitting model, leaning either to the liberal or the conservative side. It is therefore advisable to perform such *F*-tests using the baseline model (B-RCB) and switch to the best-fitting spatial model only for the pairwise *t*-tests.

General remarks

The current results show that, independently of the model for spatial correlation and the strength of correlation, a randomized complete block design and analysis according to B-RCB provide control of the nominal Type I error for the *t*-test (for 10 and 30 treatments) and for the *F*-test (for 30 treatments). In case of strong correlation and 10 treatments, the *F*-test tends to be conservative. Similar results were found for incomplete block designs and analysis by B-IB, the KR method (=KR1=KR2 in this case) yielding slightly better results than the SW method. When spatial simulation and analysis model do not coincide, then departures from the nominal Type I error rates may be more marked depending on the method of adjustment, which are larger for the *F*-test than for the *t*-test. For this reason, the choice of model selection criterion as well as the choice of candidate models is of paramount importance. In contrast, it makes little difference whether spatial correlation is modelled across the whole experimental field or only for plots within the same block (exceptions: SM and sm). When analysing by SM or sm, it is recommended to check whether a

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

nugget effect is needed. If the best-fitting model is selected by AICC and LRT and the KR1 method is used, the nominal Type I error rate is well controlled most of the time for the *t*-test in case of 30 treatments, while for 10 treatments both the *t*- and *F*-tests were on the liberal side. The power analyses show that the gain from spatial modelling can be substantial compared to a baseline model with complete block effects in the case of strong spatial correlation, and, to somewhat lesser degree, also for a baseline model with incomplete block effects. When correlations between neighbouring plots were below 0·3, neither incomplete block effects nor spatial models afforded any gain compared to analysis based on a complete block model. In the interpretation of results for incomplete blocks, it should be kept in mind that no incomplete block effects were simulated. The results show that incomplete block designs are capable of improving power compared to complete blocks when there is spatial correlation among plots.

In addition to the analyses considered in the present paper for designs with incomplete blocks, one may, of course, also combine random effects for incomplete blocks with a spatial error model. Selected additional simulations were performed (detailed results not shown) to compare the combined model with analyses based on B-IB, a purely spatial model and B-RCB. With stronger correlation, the combined model was selected as best in up to 10% of the cases based on AICC, although no incomplete block effects were simulated.

Given that the true model is among the candidate models, the current results suggest that in large trials with a larger number of treatments, a *t*-test using the KR1 method provides valid results for spatial models. The uncertainty associated with model selection can be alleviated somewhat by employing a more comprehensive class of spatial models, thus increasing the chances that the true underlying model is close to at least one of the candidate models. For example, the Matérn model, which has an additional shape parameter, provides additional flexibility, and it encompasses the exponential and Gaussian models as special cases (Matérn 1986; Haskard *et al.* 2007). The convergence behaviour of this model may be problematic, however, when no suitable starting parameter values are supplied.

The current simulations indicate that in trials with a smaller number of treatments there is a trend for tests to be on the liberal side when a spatial model is assumed in analysis. Based on the randomization theory, and independently of the number of treatments, designs with incomplete blocks and analysis by the baseline model are a viable alternative that minimizes the computational demand and yields valid tests in conjunction with the KR (=KR1=KR2 in this case) method, but at the cost of some loss in power, particularly in case of a large number of treatments and strong spatial correlation.

SUPPLEMENTARY MATERIAL

The supplementary material for this article can be found at http://journals.cambridge.org/AGS

We thank the colleagues of the Division of Agronomy and Crop Science (Humboldt-Universität zu Berlin, Faculty of Life Sciences, Albrecht Daniel Thaer-Institute of Agricultural and Horticultural Sciences, Germany) for providing the sugar beet data.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second international Symposium on Information Theory* (Eds B. N. Petrov & F. Caski), pp. 267–281. Budapest, Hungary: Akademiai Kiado. Reprinted 1992 in *Breakthroughs in Statistics I. Foundations and Basic Theory* (Eds S. Kotz & N. L. Johnson), pp. 610–624. New York: Springer.

Besag, J. & Kempton, R. (1986). Statistical analysis of field experiments using neighbouring plots. *Biometrics* **42**, 231–251.

Brownie, C., Bowman, D. T. & Burton, J. W. (1993). Estimating spatial variation in analysis of data from yield trials: a comparison of methods. *Agronomy Journal* **85**, 1244–1253.

Burnham, K. P. & Anderson, D. R. (1998). *Model Selection and Inference*. New York: Springer.

Fai, A. H. T. & Cornelius, P. L. (1996). Approximate *F*-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation* **54**, 363–378.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Giesbrecht, F. G. & Burns, J. C. (1985). Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics* **41**, 477–486.

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

Gomez, E. V., Schaalje, G. B. & Fellingham, G. W. (2005). Performance of the Kenward–Roger method when the covariance structure is selected using AIC and BIC.

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press

Communications in Statistics – Simulation and Computation **34**, 377–392.

Grondona, M. O. & Cressie, N. (1991). Using spatial considerations in the analysis of experiments. *Technometrics* **33**, 381–392.

Harville, D. A. & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association* **87**, 724–731.

Haskard, K. A., Cullis, B. R. & Verbyla, A. P. (2007). Anisotropic Matérn correlation and spatial prediction using REML. *Journal of Agricultural, Biological, and Environmental Statistics* **12**, 147–160.

Hu, X., Spilke, J. & Richter, C. (2006). The influence of spatial covariance on the Type I error and the power for different evaluation models. *Biometrical Letters* **43**, 19–37.

Hurvich, C. M. & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Irvine, K. M., Gitelman, A. I. & Hoeting, J. A. (2007). Spatial designs and properties of spatial correlation: effects on covariance estimation. *Journal of Agricultural, Biological and Environmental Statistics* **12**, 450–469.

Kenward, M. G. & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997.

Kenward, M. G. & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis* **53**, 2583–2595.

Matérn, B. (1986). *Spatial Variation*, 2nd edn. Lecture Notes in Statistics no. 36. New York: Springer.

Müller, B. U., Kleinknecht, K., Möhring, J. & Piepho, H.-P. (2010). Comparison of spatial models for sugar beet and barley trials. *Crop Science* **50**, 794–802.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13–27.

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

Piepho, H.-P., Büchse, A. & Richter, C. (2004). A mixed modelling approach for randomized experiments with repeated measures. *Journal of Agronomy and Crop Science* **190**, 230–247.

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.

Pilarczyk, W. (2009). The extent and prevailing shape of spatial relationship in Polish variety testing trials on wheat. *Plant Breeding* **128**, 411–415.

Richter, C. & Kroschewski, B. (2012). Geostatistical models in agricultural field experiments: investigations based on uniformity trials. *Agronomy Journal* **104**, 91–105.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika* **6**, 309–316.

Schaalje, G. B., McBride, J. B. & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics* **7**, 512–524.

Schabenberger, O. & Pierce, F. J. (2002). *Contemporary Statistical Models for the Plant and Soil Sciences*. Boca Raton: CRC Press.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Spilke, J., Piepho, H.-P. & Hu, X. (2005). A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological, and Environmental Statistics* **10**, 374–389.

Spilke, J., Richter, C. & Piepho, H.-P. (2010). Model selection and its consequences for different split-plot designs with spatial covariance and trend. *Plant Breeding* **129**, 590–598.

Stroup, W. W. (2002). Power analysis based on spatial effects mixed-models: a tool for comparing design and analysis strategies in the presence of spatial variability. *Journal of Agricultural, Biological and Environmental Statistics* **7**, 491–511.

Whitaker, D., Williams, E. R. & John, J. A. (2009). *CycDesigN 4.0: A Package for the Computer Generation of Experimental Designs*. Naseby, New Zealand: CycSoftware Ltd.

Wu, T. & Dutilleul, P. (1999). Validity and efficiency of neighbor analyses in comparison with classical complete and incomplete block analyses of field experiments. *Agronomy Journal* **91**, 721–731.

Wu, T., Mather, D. E. & Dutilleul, P. (1998). Application of geostatistical and neighbor analyses to data from plant breeding trials. *Crop Science* **38**, 1545–1553.

Yates, F. (1939). The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments. *Biometrika* **30**, 440–466.

Zhang, H. & Zimmerman, D. L. (2005). Toward reconciling two asymptotic frameworks in spatial statistics. *Biometrika* **92**, 921–936.

Zimmerman, D. L. & Harville, D. A. (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics* **47**, 223–239.

https://doi.org/10.1017/S0021859614000823 Published online by Cambridge University Press