DEPARTMENT OF APPLIED ECONOMICS

International
ment Informatics • M
omics and Strategy • Marke
anisation Studies • Production
erations
lethods
ncial Sc
nagement
Marketing
ction and Operations Manag
titative Methods • Accounti
Actuarial Science •
Managemen

RESEARCH REPORT

PRACTICAL INFERENCE FROM INDUSTRIAL SPLIT-PLOT DESIGNS

PETER GOOS • IVAN LANGHANS • MARTINA VANDEBROEK

OR 0407

Practical inference from industrial split-plot designs

Peter Goos
Katholieke Universiteit Leuven

Ivan Langhans
CQ Consultancy

Martina Vandebroek
Katholieke Universiteit Leuven

Abstract

Many industrial response surface experiments are deliberately not conducted in a completely randomized fashion. This is because some of the factors investigated in the experiment are hard to change. The resulting experimental design then is of the split-plot type and the observations in the experiment are in many cases correlated. A proper analysis of the experimental data therefore is a mixed model analysis involving generalized least squares estimation. Many people, however, analyze the data as if the experiment was completely randomized, and estimate the model using ordinary least squares. The purpose of the present paper is to quantify the differences in conclusions reached from the two methods of analysis and to provide the reader with guidance for analyzing split-plot experiments in practice. The problem of choosing the number of degrees of freedom for significance tests in the mixed model analysis is discussed as well.

Keywords: generalized least squares, ordinary least squares, Satterthwaite's method, method of Kenward and Roger, residual method, containment method.

1 Introduction

Split-plot experiments are often used in industry because some of the experimental variables are costly or hard to change. The split-plot nature of the experiment comes from the fact that the levels of the hard-to-change variables are not reset independently for several sequences of runs. Simpson, Kowalski and Landman (2003), for example, describe a wind tunnel experiment to investigate how the total drag and the down force of a race car depend on the front ride height, the rear ride height, the yaw angle and the covering of the car's grille. To change the front and rear ride heights was extremely cumbersome. Therefore, the levels of these factors were held constant for five different runs in the experiment. The levels of the other variables, yaw angle and grille cover, were reset in each run. The final design used for the experiment comprised 45 runs. Conducting this split-plot experiment reduced the test time by 20% over a completely randomized experiment with only 20 observations, in which the front and rear ride heights would have been reset independently for every run. Bisgaard (2000) observed that split-plotting is common and much more frequently used than the literature on design of experiments in engineering would suggest. He states that split-plot designs play a key role in the industrial application of factorial experiments.

The design of industrial split-plot experiments has received a considerable amount of attention in the recent literature on experimental design. First, minimum aberration two-level fractional factorial split-plot designs have been derived by Huang, Chen and Voelkel (1998) and Bingham and Sitter (1999, 2001). D-optimal split-plot designs for response surface experiments have been computed by Goos and Vandebroek (2001, 2003, 2004). Tailor-made split-plot designs can also be constructed using the algorithm presented in Trinca and Gilmour (2001). Kowalski, Cornell and Vining (2002) proposed standard designs for split-plot experiments with both mixture variables and process variables. Vining, Kowalski and Montgomery (2003) show how response surface designs can be modified to fit in a split-plot structure. They also discuss the conditions for ordinary and generalized least squares estimation to be equivalent. Earlier papers on split-plot designs in industry were written by Cornell (1988), who points out that experiments involving mixture and process variables are often conducted in a split-plot fashion, and Box and Jones (1992), who discuss the split-plot analysis of robust product experiments. The split-plot design and analysis of prototype experiments is discussed in Bisgaard and Steinberg (1997). Anbari and Lucas (1994) showed that split-plot designs are sometimes statistically more efficient for making predictions than completely randomized experiments.

Ganju and Lucas (1997) point out that the split-plot nature of experiments is often hidden because the runs of an experiment may be conducted in a random order but without resetting one or more of the hard-to-change experimental variables. They point out that analyzing such experiments as if it were completely randomized experiments may lead to erroneous conclusions. The topic of these so-called randomized-not-reset (RNR) experiments is revisited in Ganju and Lucas (1999, 2004), who argue that the run order of experiments should be reported as well as the presence of hard-to-change factors in order to allow a proper mixed model analysis of the data. The statistical efficiency of RNR experiments has been further investigated by Ju and Lucas (2002) and Webb, Lucas and Borkowski (2004). Further references concerning split-plot designs can be found in Myers, Montgomery, Vining, Borror and Kowalski (2004). A summary of some of the recent work on split-plot designs can be found in Goos (2002).

Despite all interest in running and designing split-plot experiments, practitioners still use ordinary least squares (OLS) to analyze the experimental results, thereby ignoring the split-plot nature of the experiment and the fact that the observations are most likely correlated. A major reason for this is that performing a generalized least squares (GLS) analysis is not supported in popular packages used by industrial statisticians. It is well-known that OLS and GLS will lead to different results when significant factor effects need to be determined. The purpose of this paper is to investigate the differences between the two approaches in detail. We will do so from the viewpoint of an industrial statistician. Therefore, we use small experimental designs and focus our attention on two key aspects of statistical decision making. Firstly, we will investigate how the probability of finding significant effects depends on the inference method used. For that purpose, we have performed an extensive simulation study involving the OLS approach and four different GLS approaches. The four GLS approaches, all of which are available in SAS 8.02, differ in the way the degrees of freedom and the test statistics are computed. Secondly, we will investigate if different answers are obtained from OLS and GLS estimation in terms of the optimum settings for the experimental variables in a response surface experiment.

In the next section, we provide a motivating example, illustrating the use of split-plot experiments in practice and the fact that OLS and GLS might lead to different conclusions. In Section 3, the statistical model is described and OLS and GLS estimation are discussed. In Section 4, the focus is on the detection of significant effects. In the beginning of that section, the four different approaches for determining the denominator degrees of freedom in significance tests are introduced succinctly. Finally, in Section 5, the focus is on the determination of the optimum settings of a process. A short discussion containing practical recommendations for the analysis and the design of split-plot response surface experiments concludes the paper.

2 Motivating example

One typical example of an industrial experiment that is conducted in a split-plot fashion is reported by Webb et al. (2004). They describe a Box-Behnken design conducted at a computer components manufacturing company for improving the performance of a wrapper machine for packaging products in air-tight bags. The factors investigated in the experiment were the spacing of the seal crimper (w), the speed at which the machine was run (s_1) and the temperature of the crimper (s_2). Being aware that the spacing of the crimper was hard to reset, the experimenters reset its level only four times in order to save time and costs. The design and the data for this experiment are displayed in Table 1.

If the data are analyzed ignoring the split-plot nature of the experiment, for example using the SAS procedure GLM (the code needed is given in Appendix A), the main effects of the factors speed and temperature are significant at the 5% level. If, on

Table 1: Data for the wrapper machine example.

wp	spacing	speed	temp	response
1	0	+1	−1	5.005
	0	+1	+1	9.170
	0	0	0	9.235
2	+1	0	+1	8.450
	+1	0	−1	5.110
	+1	−1	0	9.155
	+1	+1	0	5.010
3	−1	+1	0	5.800
	−1	−1	0	10.885
	−1	0	−1	5.940
	−1	0	+1	9.110
4	0	0	0	8.090
	0	−1	−1	9.100
	0	−1	+1	10.150
	0	0	0	8.195

the contrary, the correlation structure, resulting from the fact that the spacing of the crimper was held constant for several successive runs, is taken into account using the SAS procedure MIXED (the code is given in Appendix A as well), the two-way interaction between speed and temperature is significantly different from zero as well. If the insignificant terms are removed from the model, the model obtained using the GLM procedure is

$$\mathrm{E}(y) = 8.51 - 1.79w + 1.47s_1,$$

where w and s_1 represent the speed and the temperature respectively. Using the MIXED procedure, the model obtained is given by

$$\mathrm{E}(y) = 8.74 - 2.14w + 1.47s_1 + 0.78ws_1,$$

no matter what degrees of freedom option is selected (see Section 4.1 for more details about degrees of freedom options). In this example, both models suggest that a low level for the factor speed and a high level for the factor temperature should be selected if a higher response is desired. However, this example demonstrates that ignoring the fact that the experiment was not completely randomized may lead to a different model and thus to a different decision regarding the settings of the experimental variables.

3 Model and analysis

In this section, we describe the statistical model which is commonly used for split-plot experiments and derive the corresponding correlation structure of the data they produce. We also describe how the correlation structure can be taken into account when estimating the split-plot model as well as how split-plot experiments are often analyzed in practice.

3.1 The split-plot model

In a split-plot experiment, there are two types of experimental variables. The m_w hard-to-change factors are called whole plot variables and denoted by $w_1, w_2, \ldots, w_{m_w}$ or simply by $\mathbf{w}$. The remaining $m_s = m - m_w$ variables are the sub-plot variables $s_1, s_2, \ldots, s_{m_s}$ or $\mathbf{s}$. In the wrapper machine experiment described in the previous section, there is one whole plot factor, namely the spacing of the seal crimper, and there are two sub-plot factors: the machine speed and the temperature of the crimper.

A split-plot design consists of sets of runs of which the levels of the hard-to-change variables $\mathbf{w}_i$ are held constant. With these sets, which are henceforth referred to as whole plots, a whole plot error is associated to indicate these runs are correlated. For the wrapper machine experiment, the whole plots are indicated in Table 1. They correspond to the independent resettings of the seal crimper spacing. Next, the combinations of $\mathbf{s}_{ij}$ are randomized within the whole plot, generating the residual or sub-plot error variance. In the machine wrapper experiment, there are three whole plots of size four and one whole plot of size three. We denote the number of runs in the ith whole plot by k_i and the total number of runs in the experiment by $n = \sum_{i=1}^{b} k_i$, where b is the number of whole plots or, equivalently, the number of independent resettings of the hard-to-change variables.

For a polynomial model, the jth observation ($j = 1, 2, \ldots, k_i$) within the ith whole plot ($i = 1, 2, \ldots, b$) of a split-plot experiment can be written as

$$y_{ij} = \mathbf{f}'(\mathbf{w}_i, \mathbf{s}_{ij})\boldsymbol{\beta} + \gamma_i + \varepsilon_{ij}, \tag{1}$$

where $\mathbf{f}'(\mathbf{w}_i, \mathbf{s}_{ij})$ represents the polynomial expansion of the whole plot variables and the sub-plot variables, the $p \times 1$ vector $\boldsymbol{\beta}$ contains the p model parameters, γ_i is the random effect of the ith whole plot or the ith whole plot error, and ε_{ij} is the sub-plot error. Note that the polynomial $\mathbf{f}'(\mathbf{w}_i, \mathbf{s}_{ij})\boldsymbol{\beta}$ contains terms involving whole plot factors only, terms involving sub-plot factors only, and terms involving interactions between both types of factors. We will refer to the corresponding model parameters as whole plot coefficients, sub-plot coefficients and whole plot by sub-plot interaction coefficients. In matrix notation, the model corresponding to a split-plot design is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ represents the design matrix containing the settings of both the whole plot variables $\mathbf{w}$ and the sub-plot variables $\mathbf{s}$. The matrix $\mathbf{Z}$ is a $n \times b$ matrix of zeroes and ones assigning the n observations to the b whole plots: the (i, j)th entry of $\mathbf{Z}$ is equal to one if the jth observation belongs to the ith whole plot, and zero otherwise. The random effects of the b whole plots are contained within the b-dimensional vector $\boldsymbol{\gamma}$, and the random errors are contained within the n-dimensional vector $\boldsymbol{\varepsilon}$. It is assumed that

$$\begin{aligned} \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_n \text{ and } \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}_n, \\ \mathrm{E}(\boldsymbol{\gamma}) = \mathbf{0}_b \text{ and } \mathrm{Cov}(\boldsymbol{\gamma}) = \sigma_\gamma^2 \mathbf{I}_b, \\ \mathrm{Cov}(\boldsymbol{\gamma}, \boldsymbol{\varepsilon}) = \mathbf{0}_{b \times n}. \end{aligned}$$

Under these assumptions, the variance-covariance matrix of the observations $\mathrm{Var}(\mathbf{y})$ can be written as

$$\mathbf{V} = \sigma_\varepsilon^2 \mathbf{I}_n + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}'.$$

Suppose the entries of $\mathbf{y}$ are arranged per whole plot, then

$$\mathbf{V} = \mathrm{diag}(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_b),$$

where

$$\begin{aligned} \mathbf{V}_i &= \sigma_\varepsilon^2 \mathbf{I}_{k_i} + \sigma_\gamma^2 \mathbf{1}_{k_i} \mathbf{1}_{k_i}', \\ &= \sigma_\varepsilon^2 (\mathbf{I}_{k_i \times k_i} + \eta \mathbf{1}_{k_i} \mathbf{1}_{k_i}'), \end{aligned}$$

and $\eta = \sigma_\gamma^2 / \sigma_\varepsilon^2$ is a measure for the extent to which observations within the same whole plot are correlated. Since both σ_ε^2 and σ_γ^2 are positive numbers, η is also positive. The larger η, the more the observations within a whole plot are correlated. According to Bisgaard and Steinberg (1997), the whole plot error is often larger than the sub-plot error, so that $\eta > 1$. In practice, the variance components and thus also η have to be estimated. Letsinger, Myers and Lentner (1996), for instance, obtain $\hat{\eta} = \hat{\sigma}_\gamma^2 / \hat{\sigma}_\varepsilon^2 = 2433/2332 = 1.04$ for a chemical split-plot experiment. The wrapper machine example introduced in Section 2 yields $\hat{\eta} = 1.0801/0.1562 = 6.91$ and Goos (2002) obtains $\hat{\eta} = 1.5876/1.9470 = 0.82$ for the example in Kowalski et al. (2002).

3.2 Proper analysis of a split-plot experiment

When the random error terms as well as the whole plot effects are normally distributed, the maximum likelihood estimate of the unknown model parameter $\boldsymbol{\beta}$ in (1) is the GLS estimate instead of the OLS estimate. As a result, the unknown model parameters $\boldsymbol{\beta}$ should be estimated by

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{2}$$

and the variance-covariance matrix of the estimators is given by

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{3}$$

Usually, however, the variances σ_γ^2 and σ_ε^2 are not known and therefore, (2) and (3) cannot be used directly. Instead the variance components σ_γ^2 and σ_ε^2 are estimated, and the estimates $\hat{\sigma}_\gamma^2$ and $\hat{\sigma}_\varepsilon^2$ are substituted in the GLS estimator (2), yielding the so-called feasible GLS or FGLS estimator

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y},$$

where

$$\hat{\mathbf{V}} = \hat{\sigma}_\varepsilon^2\mathbf{I}_n + \hat{\sigma}_\gamma^2\mathbf{Z}\mathbf{Z}'.$$

The variance-covariance matrix (3) then becomes

$$\text{var}(\hat{\boldsymbol{\beta}}_{\text{FGLS}}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}, \tag{4}$$

which is used in GLS inference procedures even though it is biased downward in small samples if utilized as a measure of precision for $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$ (see, e.g. Kenward and Roger (1997)). However, Letsinger et al. (1996) as well as Goos and Vandebroek (2001) showed that (4) provides a reasonable approximation to the true finite sample variance-covariance matrix. Variance component estimates are thoroughly described in Letsinger et al. (1996). They recommend restricted maximum likelihood (REML) estimation, which is the default estimation method in the SAS procedure MIXED, of the variance components σ_γ^2 and σ_ε^2 because this method performs well for various values of η and because it is also a good estimation option when smaller designs and near full second order models are used.

3.3 Improper analysis of a split-plot experiment

Researchers are sometimes unaware of the split-plot nature of the experiment and therefore ignore the fact that the observations may be correlated. In other cases, researchers who know that the experiment they performed is of the split-plot type simply ignore this because the software that is available to them is unable to compute the FGLS estimates. The unknown model parameters are then estimated using OLS:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{5}$$

The correct variance-covariance matrix of this estimator is

$$\text{var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{6}$$

but, in practice,

$$\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{7}$$

where $\hat{\sigma}^2$ is the estimated residual error variance, is used for statistical inference. In the sequel of this paper, we assume that the variance-covariance matrix (7) is used if a split-plot design is analyzed using OLS. In Appendix B, it is shown that the expected value of the estimator for σ^2 in this approach equals

$$\text{E}\Big(\frac{\mathbf{e}'\mathbf{e}}{n-p}\Big) = \sigma_\varepsilon^2 + \sigma_\gamma^2\frac{n - \text{trace}\{\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\}}{n-p}, \tag{8}$$

Table 2: Comparison of the variances of the parameter estimators obtained from the split-plot Box-Behnken design in Table 1.

		Split-plot design		
			corr.	incorr.
Effect	CRD	GLS	OLS	OLS
Intercept	0.4121	0.5956	0.6521	0.2681
w	0.1545	0.5600	0.5600	0.1005
s_1	0.1545	0.0274	0.1545	0.1005
s_2	0.1545	0.0195	0.0195	0.1005
w^2	0.3348	1.1233	1.1374	0.2178
s_1^2	0.3348	0.0432	0.0573	0.2178
s_2^2	0.3348	0.0432	0.0573	0.2178
ws_1	0.3091	0.0391	0.0391	0.2011
ws_2	0.3091	0.0391	0.0391	0.2011
s_1s_2	0.3091	0.0391	0.0391	0.2011

where $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}_{\text{OLS}}$ is the vector of residuals obtained from an OLS regression. For the example in Section 2, this yields 0.8043, which lies between σ_ε^2 and the total variance $\sigma_\varepsilon^2 + \sigma_\gamma^2$.

The risks of improperly analyzing a split-plot experiment are pointed out in Kempthorne (1952), Nelson (1985), Box and Jones (1992), Davison (1995) and Ganju and Lucas (1997). By using a split-plot design, a loss of precision in the estimation of whole plot coefficients is incurred, while the opposite is true for the sub-plot coefficients and the whole plot by sub-plot interaction coefficients. This is illustrated in Table 2, where the parameter estimator variances for a completely randomized Box-Behnken design (CRD) with 15 observations and the split-plot Box-Behnken design in Table 1 are compared. It was assumed that $\sigma_\varepsilon^2 = 0.1562$ and $\sigma_\gamma^2 = 1.0801$ (these are the REML estimates for the variance components for the data in Table 1). It was also assumed that the model was estimated using OLS for the CRD. For the split-plot design, both OLS and GLS estimation were used. The variances for GLS estimation are given by the diagonal elements of the estimated version of (4). For OLS estimation, the correct variances are given by the diagonal elements of the estimated version of (6). These variances are in the column labelled "correct OLS". The variances in the column labelled "incorrect OLS" are computed using (7) and (8). The table illustrates that the whole plot coefficients, i.e. the coefficients of w and w^2, are estimated less efficiently from a split-plot design, no matter what estimation method is used. This can be seen by comparing the variances in the columns labelled "GLS" and "correct OLS" with the other columns. The opposite is illustrated for the other effects. The loss of precision in the whole plot coefficients is intuitive since the whole plot factor effects are confounded with the whole plot errors. The sub-plot coefficients and the whole plot by sub-plot interaction coefficients are estimated more precisely because they aren't confounded with the whole plot errors. In the next section, we investigate how serious this problem is in small response surface split-plot experiments. For that purpose, we distinguish between several methods for determining the denominator degrees of freedom for the statistical tests.

4 Finding significant effects

As described above, it is well-known that whole plot effects are estimated less precisely from a split-plot design than from a completely randomized design. The opposite is true for the effects of the sub-plot variables and for the interactions between the whole plot and the sub-plot variables. Therefore, analyzing the data from a split-plot experiment using OLS, that is by ignoring the fact that the experiment wasn't properly randomized, makes that whole plot effects are often erroneously considered as statistically significant. On the contrary, sub-plot effects and whole plot by sub-plot interaction effects are considered insignificant too often. Despite the fact that this is known, we did not find any reference where the seriousness of the problem has been quantified in the context of small response surface experiments typically used in industry.

4.1 Degrees of freedom

Using the SAS procedure MIXED, five methods to determine the denominator degrees of freedom for the tests on the individual model parameters can be used. Two of them, i.e. the RESIDUAL method and the BETWEEN-WITHIN option, are equivalent for the type of experimental design examined in the present paper.

4.1.1 Containment method

The CONTAIN option invokes the containment method, which, for the experimental designs investigated in this paper, takes $n - \text{rank}\,[\,\mathbf{X}\ \mathbf{Z}\,]$ as the degrees of freedom. Since $p < \text{rank}\,[\,\mathbf{X}\ \mathbf{Z}\,] < p + b$ for the designs under investigation, the denominator degrees of freedom lie between $\max\{1, n - (p + b)\}$ and $n - p$. This choice of degrees of freedom matches the tests performed for balanced split-plot designs and should be adequate for moderately unbalanced designs.

4.1.2 Residual method

The RESIDUAL option performs all tests using the residual degrees of freedom, $n - p$, as the denominator degrees of freedom. For a split-plot design, this method produces the same number of degrees of freedom as the BETWITHIN option, which, in general, divides the residual degrees of freedom into between-subject and within-subject portions. Therefore, we do not treat this option explicitly. Note also that the same degrees of freedom are used if the split-plot nature of the experiment is ignored.

4.1.3 Satterthwaite's method

If the SATTERTH option is chosen, the degrees of freedom for a t-statistic

$$t = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{c}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{c}}}, \tag{9}$$

where $\mathbf{c}$ is a p-dimensional vector defining an estimable linear combination of $\boldsymbol{\beta}$, are computed as

$$\nu = \frac{2\{\mathbf{c}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{c}\}^2}{\mathbf{g}'\mathbf{A}\mathbf{g}},$$

where $\mathbf{g}$ is the gradient of $\mathbf{c}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{c}$ with respect to the unknown model parameters $\boldsymbol{\beta}$, σ_δ^2 and σ_ε^2, evaluated at $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_\delta^2$ and $\hat{\sigma}_\varepsilon^2$, and $\mathbf{A}$ is the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_\delta^2$ and $\hat{\sigma}_\varepsilon^2$. For testing the statistical significance of the ith element of $\boldsymbol{\beta}$, $\mathbf{c}$ is a vector having a one as its ith element and zeroes elsewhere.

4.1.4 Method of Kenward & Roger

The method of Kenward and Roger (1997) involves inflating the estimated variance-covariance matrix (3) of parameter estimators to take into account the fact that the variance components σ_δ^2 and σ_ε^2, and thus $\mathbf{V}$, need to be estimated. Satterthwaite degrees of freedom are then computed based on the adjusted variance-covariance matrix. For the experimental design investigated in the present paper, the inflation of (3) didn't noticeably affect the degrees of freedom but it did affect the denominator of the t-statistic (9) and thus the statistical test and the corresponding conclusion.

4.2 Simulation results

We have performed a number of simulation studies to investigate how severe the differences between the analyses with OLS and GLS are. We have considered a 2^4 factorial design, a three-variable central composite design (CCD) and a D-optimal design in two variables on a constrained design region. First, we consider the situation in which none of the experimental variables has an effect on the response. The purpose of this is to investigate how many type I errors are made if OLS or GLS inference procedures are used. Next, we investigate how sensitive each of the inference methods is when it comes to detecting nonzero effects.

In all our simulations, the random whole plot effects γ_i and the random errors ε_{ij} were drawn from independent normal distributions with zero means and variances σ_γ^2 and σ_ε^2 respectively. We performed computations using three values for $\sigma^2 = \sigma_\gamma^2 + \sigma_\varepsilon^2$ (5, 20 and 45) and using four different variance ratios $\eta = \sigma_\gamma^2/\sigma_\varepsilon^2$ (1, 2, 4 and 8).

Table 3: 2^4 design arranged in four whole plots of size four.

wp	w	s_1	s_2	s_3
1	-1	-1	-1	-1
	-1	-1	$+1$	$+1$
	-1	$+1$	$+1$	-1
	-1	$+1$	-1	$+1$
2	-1	-1	$+1$	-1
	-1	-1	-1	$+1$
	-1	$+1$	-1	-1
	-1	$+1$	$+1$	$+1$
3	$+1$	-1	-1	-1
	$+1$	-1	$+1$	$+1$
	$+1$	$+1$	$+1$	-1
	$+1$	$+1$	-1	$+1$
4	$+1$	-1	$+1$	-1
	$+1$	-1	-1	$+1$
	$+1$	$+1$	-1	-1
	$+1$	$+1$	$+1$	$+1$

4.2.1 Type I errors

2^4 *factorial design*

Consider a 2^4 factorial experiment conducted in a split-plot fashion because one of the factors, w, is hard to change, whereas the three remaining factors, s_1, s_2 and s_3, are easy to change. Using the defining relation I=W=S$_1$S$_2$S$_3$, the 16 runs of the experiment were split in four whole plots of equal sizes. The design obtained this way is shown in Table 3. It is a D-optimal arrangement of the runs of the factorial design in four whole plots of size four if the interest is in estimating a model containing the linear and the two-factor interaction effects (see Goos and Vandebroek (2003)). This is because the levels of the sub-plot variables add up to zero in every whole plot and because this is also true for the two-factor interactions of the four variables. The same estimates for $\boldsymbol{\beta}$ will therefore be obtained from OLS and FGLS. The standard errors used in both analyses will however be different, as well as the degrees of freedom used in the four FGLS approaches.

The model used to simulate data was the zero model

$$\mathrm{E}(y) = 50 + 0w + 0s_1 + 0s_2 + 0s_3 + 0ws_1 + 0ws_2 + 0ws_3 + 0s_1s_2 + 0s_1s_3 + 0s_2s_3. \quad (10)$$

The purpose of using this model is to check whether the different estimation methods used lead to type I error rates of 5% if a 5% significance level is utilized. Computational results for $\sigma^2 = 20$ and $\eta = 1$ and 8 are given in Table 4. It turns out

Table 4: Percentages of type I errors for several types of effects obtained from the design in Table 3 with $\sigma^2 = 20$.

	$\eta = 1$				$\eta = 8$			
Method	w	s_1	ws_1	s_1s_2	w	s_1	ws_1	s_1s_2
Containment	5.9%	4.8%	4.7%	4.4%	7.0%	3.8%	3.9%	4.1%
Residual	10.5%	8.7%	8.7%	7.0%	10.8%	7.0%	7.5%	7.7%
Satterthwaite	6.2%	5.6%	5.2%	4.8%	4.2%	3.8%	3.9%	4.2%
Kenward & Roger	6.2%	5.6%	5.2%	4.8%	4.2%	3.8%	3.9%	4.2%
OLS	14.8%	0.2%	1.4%	1.7%	21.8%	0.1%	0.0%	0.2%

that the methods of Satterthwaite and Kenward & Roger, which produce identical degrees of freedom and standard errors for all terms in this example, yield error rates that are close to 5% for all settings of σ^2 and η. The containment method is a good alternative for the sub-plot and whole plot by sub-plot interaction effects. The residual method is the worst GLS inference procedure as it leads to about 7-9% of rejections of the null hypothesis for the pure sub-plot effects and the whole plot by sub-plot interaction effects. For the whole plot effect, it leads to 10-12% of false positives.

Analyzing the experiment ignoring the split-plot structure leads to type I error rates for the whole plot effect ranging from 12.6% (for $\sigma^2 = 5$ and $\eta = 1$) to 21.8% (for $\sigma^2 = 20$ and $\eta = 8$). For all the other terms, the type I error rate produced by this analysis method is substantially lower than 5% and it approaches zero if the variance ratio η increases.

We have also investigated the effect of duplicating the factorial design. Firstly, we have considered the situation in which the duplicated design was run in four whole plots of size eight. For this design, a selection of the results is displayed in Table 5. The containment method yields type I error rates of over 14% for the whole plot term. For the other terms, the containment method gives almost identical results compared to the methods of Satterthwaite and Kenward & Roger, which perform extremely well for all model terms. The large difference between the containment method and the methods of Satterthwaite and Kenward & Roger can be explained by the fact that the former method uses 15 degrees of freedom for the test on the whole plot coefficient, whereas the latter methods use on average 3.995 degrees of freedom when $\eta = 1$ and 2.304 when $\eta = 8$. Using OLS, error rates of more than 40% were obtained for the whole plot term. For the sub-plot terms, the error rates are close to 2% when η is small and virtually zero when η is large.

When the duplicated factorial design is run using eight whole plots of size four, the differences in type I error rates between the containment method and the methods

Table 5: Percentages of type I errors for several types of effects obtained by running a duplicated 2^4 design in four whole plots of size eight ($\sigma^2 = 20$).

	$\eta = 1$				$\eta = 8$			
Method	w	s_1	ws_1	s_1s_2	w	s_1	ws_1	s_1s_2
Containment	15.8%	5.2%	4.5%	4.3%	14.4%	5.6%	5.1%	5.4%
Residual	16.1%	5.3%	4.5%	4.5%	14.6%	5.8%	5.2%	5.6%
Satterthwaite	6.0%	5.2%	4.5%	4.3%	4.9%	5.6%	5.1%	5.4%
Kenward & Roger	6.0%	5.2%	4.5%	4.3%	4.9%	5.6%	5.1%	5.4%
OLS	40.3%	2.1%	1.8%	1.6%	53.0%	0.2%	0.2%	0.2%

Table 6: Percentages of type I errors for several types of effects obtained by running a duplicated 2^4 design in eight whole plots of size four ($\sigma^2 = 20$).

	$\eta = 1$				$\eta = 8$			
Method	w	s_1	ws_1	s_1s_2	w	s_1	ws_1	s_1s_2
Containment	7.9%	5.3%	5.6%	5.9%	7.2%	4.3%	4.8%	5.4%
Residual	8.5%	5.5%	5.8%	6.3%	8.1%	4.8%	5.0%	5.6%
Satterthwaite	5.5%	5.3%	5.6%	5.9%	4.9%	4.3%	4.8%	5.4%
Kenward & Roger	5.5%	5.3%	5.6%	5.9%	4.9%	4.3%	4.8%	5.4%
OLS	19.2%	0.7%	1.4%	1.4%	27.9%	0.0%	0.0%	0.0%

of Satterthwaite and Kenward & Roger are substantially smaller, even though the difference between the degrees of freedom used by the methods is still large: 15 for the containment method versus 6 to 6.3 for the other two methods. Note that doubling the number of whole plots leads to a substantial reduction in the type I error rate for the whole plot coefficient if the OLS analysis is applied. It ranges from 19.2% to 27.9% which is less than the 40% obtained by using four whole plots of size eight, but still substantially greater than 5%. Computational results for this case are displayed in Table 6.

The Satterthwaite method and the method of Kenward & Roger perform well for the duplicated factorial designs with four and with eight whole plots. The other methods for determining the degrees of freedom perform well for the sub-plot coefficients and the whole plot by sub-plot interactions. For the whole plot coefficient however, they perform poorly if the duplicated factorial design option with four whole plots is chosen.

Central composite design

Consider a 16-run CCD conducted in a split-plot fashion because one of the factors,

Table 7: 16-run CCD arranged in eight, six and four whole plots.

8 whole plots				6 whole plots				4 whole plots			
	w	s_1	s_2		w	s_1	s_2		w	s_1	s_2
1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1
	-1	1	1		-1	1	1		-1	1	1
2	-1	1	-1		-1	1	-1		-1	1	-1
	-1	-1	1		-1	-1	1		-1	-1	1
3	1	-1	-1	2	1	-1	-1		-1	0	0
	1	1	1		1	1	1	2	1	-1	-1
4	1	1	-1		1	1	-1		1	1	1
	1	-1	1		1	-1	1		1	1	-1
5	-1	0	0	3	-1	0	0		1	-1	1
6	1	0	0	4	1	0	0		1	0	0
7	0	-1	0	5	0	-1	0	3	0	-1	0
	0	0	0		0	0	0		0	0	0
	0	1	0		0	1	0		0	1	0
8	0	0	-1	6	0	0	-1	4	0	0	-1
	0	0	0		0	0	0		0	0	0
	0	0	1		0	0	1		0	0	1

w, is hard to change, whereas the two remaining factors, s_1 and s_2, are easy to change. We have used three different arrangements of the CCD in our simulation study. These arrangements have eight, six and four whole plots and are displayed in Table 7. The levels of s_1 and s_2 add up to zero within each whole plot. This is true for the two-factor interactions as well except for s_1s_2 in the design with eight whole plots. Therefore, the effect of s_1s_2 is partially confounded with the whole plot errors so that the results for the corresponding term will resemble those for a whole plot effect.

Again, we used a model with zero coefficients to investigate to what extent the 5% type I error rate was approximated:

$$\mathrm{E}(y) = 50 + 0w + 0s_1 + 0s_2 + 0w^2 + 0s_1^2 + 0s_2^2 + 0ws_1 + 0ws_2 + 0s_1s_2. \qquad (11)$$

Like for the factorial design, the methods attributed to Satterthwaite and to Kenward & Roger produce nearly identical results. The only difference between both methods lies in the estimates of the standard errors for the linear whole plot terms and for the quadratic sub-plot effects that are larger (about 2-3%) for the Kenward & Roger method. Overall, the method of Kenward & Roger leads to an acceptance rate closest to 5%. For the arrangement with eight whole plots, the acceptance rate lies between 4% and 6% (between 4% and 7.4%) for the linear (quadratic) whole plot term. For the arrangement with four whole plots, the acceptance rate lies be-

Table 8: Percentages of type I errors for the 3 arrangements of the CCD displayed in Table 7 ($\sigma^2 = 20$ and $\eta = 1$).

b	Method	w	w^2	s_1	s_1^2	ws_1	s_1s_2
4	Containment	17.2%	11.8%	5.3%	8.5%	4.7%	5.6%
	Residual	18.6%	13.7%	5.7%	9.8%	5.7%	6.5%
	Satterthwaite	17.2%	11.4%	5.3%	9.0%	5.2%	6.2%
	Kenward & Roger	17.2%	11.4%	5.3%	7.0%	5.2%	6.2%
	OLS	32.7%	22.5%	3.0%	12.6%	3.0%	4.6%
6	Containment	5.9%	4.6%	4.7%	5.7%	5.7%	5.3%
	Residual	9.6%	9.7%	10.1%	11.9%	9.1%	9.9%
	Satterthwaite	6.7%	7.6%	6.3%	9.9%	6.1%	6.1%
	Kenward & Roger	6.7%	7.6%	6.3%	7.4%	6.1%	6.1%
	OLS	19.4%	14.0%	2.7%	11.3%	2.2%	2.6%
8	Containment	1.0%	1.1%	5.8%	4.7%	4.7%	0.7%
	Residual	8.1%	7.8%	13.3%	14.6%	12.6%	7.8%
	Satterthwaite	5.9%	6.0%	6.3%	11.2%	6.2%	5.7%
	Kenward & Roger	5.8%	6.0%	6.3%	8.3%	6.2%	5.7%
	OLS	12.3%	12.2%	1.8%	8.1%	1.8%	12.9%

tween 16.9% and 20.4% (between 11.7% and 16.1%) for the linear (quadratic) whole plot term. These acceptance rates are quit large indicating that the design with four whole plots has too few whole plots to allow a good correction for the standard errors by the Kenward & Roger approach. Table 8 shows the results obtained for $\sigma^2 = 20$ and $\eta = 1$. For the arrangements with four and six whole plots, the containment method performs as well as the approach of Kenward & Roger. For the design option with eight whole plots, it leads to only 1% of type I errors for the whole plot coefficients. The residual method performs worse than the other three GLS inference procedures. For the sub-plot coefficients and the whole plot by sub-plot interactions, the Satterthwaite and Kenward & Rogers approaches perform well too.

Analyzing the design using OLS leads to large type I error rates for the whole plot coefficients, especially if the number of whole plots is small. If the arrangement with four whole plots is used, the type I error rate for the linear whole plot term ranges from 30.8% when $\sigma^2 = 5$ and $\eta = 1$ to 61.6% when $\sigma^2 = 45$ and $\eta = 8$. For the quadratic whole plot term, the error rate lies between 21% and 49.5%. These large error rates can also be seen in Table 8. It can also be seen that they decrease with the number of whole plots. Also, the error rates for the quadratic sub-plot effects are large. They range from 12.6% when η is small to 30.2% when η is large and $b = 4$. For the other sub-plot coefficients, the type I error rate is around 3% when the CCD with four whole plots is run and 1% or less when eight whole plots are used.

Table 9: Design with four whole plots of size three on a constrained design region.

wp	w	s
1	-0.5000	-0.1667
	-0.5000	0.5000
	-0.5000	0.5000
2	-0.1667	0.0000
	-0.1667	0.5000
	-0.1667	-0.5000
3	0.1667	0.5000
	0.1667	-0.5000
	0.1667	0.0000
4	0.5000	-0.5000
	0.5000	0.1667
	0.5000	-0.5000

D-optimal design for a constrained design region

We have also performed simulations with a two-factor design involving four whole plots of three observations on a constrained design region ($-1/2 \leq x_1, x_2 \leq 1/2$, $-2/3 \leq x_1 + x_2 \leq 2/3$). One of the factors, w, was a hard-to-change factor. The experimental design used to simulate data from the model

$$\mathrm{E}(y) = 50 + 0w + 0s + 0w^2 + 0s^2 + 0ws \tag{12}$$

is displayed in Table 9. It was obtained using the algorithm of Goos and Vandebroek (2003).

For all model terms, the methods of Satterthwaite and Kenward & Roger yield the best approximation to the 5% significance level. For the linear (quadratic) whole plot term, the type I error rate using these methods ranges from 8.3% for small values of η to 12.7% for large values of η (from 8.7% to 12.7%). For the sub-plot terms and the whole plot by sub-plot interaction, between 4.7% and 6.6% of the coefficients are judged significantly different from zero. For the tests on the sub-plot coefficients, the containment method and even the residual method are close competitors. For the whole plot coefficients, they perform considerably worse than the methods of Satterthwaite and Kenward & Roger. The OLS analysis for the design in Table 9 shows a picture similar to that for the CCD with four whole plots.

Summary

Over all models, designs, variances and variance ratios, the method of Kenward & Roger and Satterthwaite's method yield type I errors closest to 5%. The containment method performs almost equally good for the sub-plot terms. For the whole plot terms, the containment method leads to large numbers of errors when the number of observations is large and the number of whole plots is small. Ignoring the split-plot nature of the experiment leads to extremely large type I error rates for the whole plot coefficients.

4.2.2 Detecting significant effects

If the interest is in finding effects that are significantly and substantially different from zero, the picture obtained is totally different. As expected, methods that lead to large numbers of type I errors are the most sensitive ones to detect significant effects, whereas approaches that lead to too few type I errors often fail to detect active factors. We illustrate this by means of the factorial design, the CCD and the design for the constrained design region.

2^4 *factorial design*

The model used to simulate data is given by

$$\mathrm{E}(y) = 50 + 4w + 4s_1 + 2s_2 + 6s_3 + 2ws_1 - 2ws_2 + 1ws_3 + 2s_1s_2 - 2s_1s_3 + 1s_2s_3. \quad (13)$$

First, we discuss the simulation results in Table 10 for the single factorial design displayed in Table 3. When it comes to detecting the whole plot effect, ignoring the split-plot nature of the experiment is superior to all GLS approaches. Among the GLS approaches, the residual method was by far the best, whereas the methods of Satterthwaite and Kenward & Roger performed poorest. For $\sigma^2 = 20$, the difference in detection rate between the OLS analysis and the GLS approach using the residual method was around 16% when $\eta = 1$ and around 20% when $\eta = 8$. The difference between OLS analysis and the methods of Satterthwaite and Kenward & Roger amounted to 40% and 43% for these parameters. For detecting the sub-plot effects, the residual method is slightly better than the other GLS inference procedures and between 22% and 28% better than the OLS approach. For the model coefficients and settings for σ^2 and η used in the study, the differences in hit rates for the sub-plot effects and the whole plot by sub-plot interaction effects between the residual method and OLS were never smaller than 20% and reached peaks of 70% for small effects. These large differences were obtained for large values of η and signify that the residual method detected the effect in most instances, whereas the OLS approach failed to do so in a large number of simulations. For example, the effects of the interaction terms ws_1 or ws_2 (with model coefficients of 2 and -2) were detected in 80% of the simulations when $\sigma^2 = 45$ and $\eta = 8$, whereas this was only 10% when the OLS analysis was adopted.

Doubling the number of observations in the four whole plots of the experiment doesn't change the relative performances of the inference procedures for detecting the whole plot effect: ignoring the split-plot nature of the design still produces the

Table 10: Frequencies of detecting whole plot and sub-plot effects for the design in Table 3 ($\sigma^2 = 20$).

	$\eta = 1$			$\eta = 8$		
D.f. method	w	s_1	s_2s_3	w	s_1	s_2s_3
Containment	36.5%	92.8%	14.6%	33.8%	100.0%	45.7%
Residual	49.1%	97.8%	22.7%	44.6%	100.0%	62.0%
Satterthwaite	25.4%	93.9%	15.8%	21.7%	100.0%	46.3%
Kenward & Roger	25.4%	93.9%	15.8%	21.7%	100.0%	46.3%
OLS	65.5%	75.6%	5.6%	64.9%	72.4%	7.5%

Table 11: Frequencies of detecting whole plot and sub-plot effects obtained by running a duplicated 2^4 design in four whole plots of size eight ($\sigma^2 = 20$).

	$\eta = 1$			$\eta = 8$		
D.f. method	w	s_1	s_2s_3	w	s_1	s_2s_3
Containment	66.5%	100.0%	40.1%	52.1%	100.0%	96.1%
Residual	66.8%	100.0%	40.7%	52.2%	100.0%	96.2%
Satterthwaite	29.7%	100.0%	40.1%	20.3%	100.0%	96.1%
Kenward & Roger	29.7%	100.0%	40.1%	20.3%	100.0%	96.1%
OLS	92.4%	99.8%	25.0%	90.0%	100.0%	31.5%

best results, whereas Satterthwaite's method and that of Kenward & Roger perform poorly. They are beaten by the residual and the containment methods, which are still worse than the OLS approach. The effect of doubling the number of observations is that the number of statistically significant whole plot effects detected increases substantially for the containment, the residual and the OLS methods. This is shown in Table 11. As to detecting effects other than whole plot effects, all GLS approaches produce almost identical results. Like for the term involving s_2s_3 in Table 11, they outperform the OLS analysis by up to 65% when the sub-plot effects are small and $\eta = 8$.

Doubling the number of whole plots instead of the number of observations within each whole plot leads to a smaller difference between the GLS approaches and the OLS analysis. As it can be seen from Table 12, the residual method remains the best GLS method for detecting the whole plot effect. It is still considerably less effective than using OLS as it finds 10% to 21% less significant effects. For the other effects, all GLS approaches yield similar results, with a slight edge for the residual method. They all perform much better than OLS in terms of detecting the smaller sub-plot effects in model (13).

Table 12: Frequencies of detecting whole plot and sub-plot effects obtained by running a duplicated 2^4 design in eight whole plots of size four ($\sigma^2 = 20$).

	$\eta = 1$			$\eta = 8$		
D.f. method	w	s_1	s_2s_3	w	s_1	s_2s_3
Containment	84.5%	100.0%	40.9%	68.7%	100.0%	94.8%
Residual	85.8%	100.0%	42.6%	70.7%	100.0%	95.2%
Satterthwaite	78.5%	100.0%	40.9%	58.8%	100.0%	94.8%
Kenward & Roger	78.5%	100.0%	40.9%	58.8%	100.0%	94.8%
OLS	96.1%	100.0%	17.7%	92.4%	100.0%	9.8%

Central composite design

For the CCD, the model used to simulate data is given by

$$\mathrm{E}(y) = 50 + 4w + 4s_1 - 4s_2 + 2w^2 + 2s_1^2 - 2s_2^2 + 3ws_1 + 3ws_2 - 3s_1s_2 \qquad (14)$$

As to the linear and the quadratic whole plot effects, the results using the CCD confirm the results from factorial designs. For the linear term, the differences between the OLS approach and the residual method are as large as 20% to 30% when four whole plots are used. Satterthwaite's and Kenward & Roger's methods have even more difficulties detecting nonzero effects. Increasing the number of whole plots has no effect on the relative performances of the inference procedures. An excerpt from the computational results is given in Table 13.

As to the sub-plot terms, a distinction has to be madė between the quadratic terms, on the one hand, and the linear effects and whole plot by sub-plot interaction effects, on the other hand. As it is shown in Table 13, the residual method is overall the most effective to detect significant linear effects and whole plot by sub-plot interaction effects. The differences with the other GLS approaches depend heavily on the exact number of whole plots and on σ^2 and η. The OLS analysis performs poorly compared to the GLS method, especially when η is large. As to the detection of quadratic sub-plot effects, the residual method is for some instances, for example for a CCD with four whole plots and $\eta = 1$, outperformed by the OLS approach. While the performance of the residual method is good for all the settings under study, the OLS method is less robust and may perform poorly. For example, it finds about 20% less significant quadratic sub-plot effects when $\eta = 8$ the CCD is run using six whole plots.

D-optimal design for a constrained design region

For the D-optimal design for a constrained design region, the model used to simulate

Table 13: Frequency of detecting nonzero effects for the three arrangements of the CCD displayed in Table 7 ($\sigma^2 = 20$ and $\eta = 1$).

b	D.f. method	w	w^2	s_1	s_1^2	ws_1	s_1s_2
4	Containment	48.7%	16.6%	90.1%	16.9%	61.1%	61.1%
	Residual	50.5%	18.1%	92.2%	17.8%	64.8%	65.6%
	Satterthwaite	38.7%	15.5%	91.3%	15.9%	62.7%	63.7%
	Kenward & Roger	38.7%	15.5%	91.3%	13.3%	62.7%	63.7%
	OLS	74.9%	27.2%	87.0%	20.3%	54.9%	57.2%
6	Containment	32.5%	6.4%	78.1%	10.4%	46.2%	46.3%
	Residual	45.3%	12.5%	91.6%	17.1%	65.5%	65.9%
	Satterthwaite	27.7%	9.0%	84.7%	14.2%	54.5%	55.3%
	Kenward & Roger	27.7%	9.0%	84.7%	11.1%	54.5%	55.3%
	OLS	68.5%	18.0%	79.4%	17.6%	45.6%	45.0%
8	Containment	18.6%	2.1%	57.2%	6.9%	29.8%	8.8%
	Residual	53.2%	10.4%	90.9%	21.1%	65.9%	30.4%
	Satterthwaite	40.1%	8.1%	78.2%	16.7%	48.0%	21.8%
	Kenward & Roger	39.6%	8.1%	78.2%	13.3%	48.0%	21.8%
	OLS	68.0%	15.4%	75.3%	14.3%	40.0%	42.7%

data is given by

$$\mathrm{E}(y) = 50 - 6w + 6s + 3w^2 - 3s^2 + 4ws. \tag{15}$$

This model lead to conclusions similar to those drawn from the three arrangements of the CCD, the only difference being that the residual method is clearly better than the OLS analysis when the detection of the quadratic sub-plot effect is concerned.

Summary

For detecting whole plot effects, analyzing the data ignoring the correlation caused by the split-plot design of the experiment leads to the best results. The GLS approach combined with the residual method for determining the degrees of freedom is second best, but it is substantially less powerful. For detecting significant sub-plot effects and whole plot by sub-plot interaction effects, the GLS analysis utilizing the residual degrees of freedom is better than the other GLS approaches. Remarkably, the GLS approaches have trouble detecting the quadratic sub-plot effects. For some cases, they are even beaten by the OLS analysis. This is surprising since the standard errors produced by the GLS analysis are smaller than those obtained from an OLS analysis for these terms.

5 Optimum settings

In order to investigate whether analyzing a split-plot experiment using GLS allows a researcher to determine the optimum settings of a process better than by using OLS, we have performed a simulation study using the three arrangements (in four, six and eight whole plots) of the 16-point CCD for three experimental variables displayed in Table 7. We have also investigated the same absolute and relative magnitudes of the error components σ_w^2 and σ_s^2 as in Section 4. For every situation, we have analyzed the data using OLS and GLS. In this section, the different methods for determining the degrees of freedom are of no importance because only the parameter estimates matter. As to the GLS estimation, we have distinguished between the case where the true variance components are known and the case where the variance components are unknown and need to be estimated. As indicated in Section 3.2, we have labelled the two cases GLS and FGLS respectively. The two GLS approaches take into account the fact that the experiment was conducted in a split-plot format, whereas the OLS approach ignores this. It is expected that, because the GLS estimates have smaller variances than their OLS counterparts (see Table 2 for an illustration and Appendix C for a proof), the model obtained by applying GLS will be closer to the unknown true model. Therefore, the optimum settings computed from this model should be closer to the true optimum and therefore lead to a better value for the response than does OLS.

One of the models we used to generate responses was

$$\begin{aligned} \mathrm{E}(Y) &= 50 + 8w + 3s_1 + 0s_2 - 7w^2 - 3s_1^2 + 0s_2^2 - 4ws_1 + 0ws_2 + 0s_1s_2, \\ &= \mathbf{f}'(\mathbf{x})\boldsymbol{\beta}, \end{aligned} \tag{16}$$

where $\mathbf{f}'(\mathbf{x})$ is the vector containing the expansions of the experimental variables and $\boldsymbol{\beta}$ is the p-dimensional vector containing the unknown model parameters. Under this model, the variable s_2 is inactive. The maximum value of 52.3382 is achieved at $w = 0.5294$ and $s_1 = 0.1471$, no matter what value of s_2 is used. We denote the polynomial expansion corresponding to the optimal settings of w, s_1 and s_2 by $\mathbf{x}_{\text{opt}}$.

Using the split-plot CCDs and model (16), we have generated data under different scenarios. We denote the estimated model coefficients by the p-dimensional vectors $\mathbf{b}_{\text{FGLS}}$, $\mathbf{b}_{\text{GLS}}$ and $\mathbf{b}_{\text{OLS}}$. Using these estimates, we have computed the settings for $\mathbf{x} = [\ w\ s_1\ s_2\ ]'$ that produced the largest response based on the estimated models, i.e. we have computed

$$\max_{\mathbf{x}\in[-1,+1]^3} \mathbf{f}'(\mathbf{x})\mathbf{b}_{\text{FGLS}},$$

$$\max_{\mathbf{x}\in[-1,+1]^3} \mathbf{f}'(\mathbf{x})\mathbf{b}_{\text{GLS}}$$

and

$$\max_{\mathbf{x}\in[-1,+1]^3} \mathbf{f}'(\mathbf{x})\mathbf{b}_{\text{OLS}}$$

for every simulation run. We denote the vectors corresponding to the optima found in the ith simulation by $\hat{\mathbf{x}}_{i,\text{FGLS}}$, $\hat{\mathbf{x}}_{i,\text{GLS}}$ and $\hat{\mathbf{x}}_{i,\text{OLS}}$ respectively. In order to investigate the quality of these solutions, we have compared the values

$$\Delta_{i,\text{FGLS}} = \mathbf{f}'(\mathbf{x}_{\text{opt}})\boldsymbol{\beta} - \mathbf{f}'(\hat{\mathbf{x}}_{i,\text{FGLS}})\boldsymbol{\beta},$$

$$\Delta_{i,\text{GLS}} = \mathbf{f}'(\mathbf{x}_{\text{opt}})\boldsymbol{\beta} - \mathbf{f}'(\hat{\mathbf{x}}_{i,\text{GLS}})\boldsymbol{\beta}$$

and

$$\Delta_{i,\text{OLS}} = \mathbf{f}'(\mathbf{x}_{\text{opt}})\boldsymbol{\beta} - \mathbf{f}'(\hat{\mathbf{x}}_{i,\text{OLS}})\boldsymbol{\beta}.$$

These values, which are all positive because $\mathbf{f}'(\mathbf{x}_{\text{opt}})\boldsymbol{\beta}$ is the highest possible value of the true response, indicate how far off the solution is from the one that is optimal. As explained above, it is expected that $\Delta_{i,\text{GLS}} < \Delta_{i,\text{FGLS}} < \Delta_{i,\text{OLS}}$. The results reported are based on 10,000 simulations and were obtained by averaging $\Delta_{i,\text{FGLS}}$, $\Delta_{i,\text{GLS}}$ and $\Delta_{i,\text{OLS}}$ over all simulations. The averages are denoted by Δ_{FGLS}, Δ_{GLS} and Δ_{OLS} respectively. In order to have an idea of the relative importance of the deviations Δ_{GLS}, Δ_{FGLS} and Δ_{OLS}, it is interesting to point out that the simulated responses ranged from 21 to 57 on average.

5.1 A single central composite design

First, consider the situation in which a single CCD design is used to estimate the full quadratic model in the three experimental variables. The three arrangements displayed in Table 7 were used. The deviations displayed in Table 14 show that the relative performance of the GLS and FGLS estimation compared to the OLS estimation is better for smaller numbers of whole plots w and for larger variance ratios η. It is striking that the improvement realized by utilizing the more complicated GLS and FGLS estimation methods is quite small. This is even more so if the deviations Δ_{FGLS}, Δ_{GLS} and Δ_{OLS} are compared to the overall variance $\sigma^2 = \sigma_w^2 + \sigma_s^2$.

We have also investigated whether a different arrangement of the 16 runs of the CCD in whole plots produces different results. The designs we have used for that purpose are displayed in Table 15. For these designs, the levels of the sub-plot factors s_1 and s_2 no longer add up to zero which makes the design is no longer orthogonal with respect to s_1 and s_2 and that the OLS and GLS estimates for the model parameters more different. It was therefore expected that the OLS and (F)GLS estimates of $\boldsymbol{\beta}$ would be more different than for the designs in Table 14. As a result, larger differences between the OLS and (F)GLS results were probable. The simulation results for the designs in Table 15 are given in Table 16. The differences between Δ_{OLS} and Δ_{FGLS} are larger, so that using GLS is more beneficial here. There are also larger difference between Δ_{GLS} and Δ_{FGLS}. This means that part of the benefit of using GLS is lost here because the variance components have to be estimated. It is interesting to note as well that, for many combinations of w, η and σ^2, the values of Δ_{OLS} and Δ_{GLS} are larger than their counterparts in Table 14. This means that

Table 14: Simulation results for a 16-run CCD with four, six and eight whole plots for various η- and σ^2-values.

		$\sigma^2 = 5$			$\sigma^2 = 20$			$\sigma^2 = 45$		
w	η	Δ_{OLS}	Δ_{GLS}	Δ_{FGLS}	Δ_{OLS}	Δ_{GLS}	Δ_{FGLS}	Δ_{OLS}	Δ_{GLS}	Δ_{FGLS}
4	1	0.5962	0.5356	0.5769	1.3416	1.2962	1.3154	1.8972	1.8916	1.8908
	2	0.5991	0.4843	0.5038	1.3160	1.2201	1.2361	1.9105	1.8641	1.8757
	4	0.5955	0.4269	0.4323	1.2890	1.1115	1.1251	1.8964	1.8098	1.8177
	8	0.5954	0.3837	0.3838	1.2708	0.9980	0.9986	1.8818	1.7235	1.7142
6	1	0.6097	0.5442	0.5815	1.3156	1.2562	1.2886	1.8095	1.7625	1.7679
	2	0.6125	0.4860	0.5233	1.2807	1.1675	1.2068	1.7565	1.6638	1.6798
	4	0.6076	0.4211	0.4463	1.2441	1.0660	1.0910	1.7015	1.5406	1.5623
	8	0.6006	0.3621	0.3690	1.2132	0.9471	0.9517	1.6824	1.4183	1.4330
8	1	0.6249	0.5769	0.6158	1.3499	1.3214	1.3439	1.7748	1.7713	1.7677
	2	0.6412	0.5478	0.5870	1.3337	1.2737	1.3006	1.7447	1.7260	1.7316
	4	0.6467	0.5108	0.5383	1.3135	1.2217	1.2432	1.7043	1.6766	1.6727
	8	0.6490	0.4704	0.4788	1.2972	1.1760	1.1816	1.6689	1.6262	1.6237

it is good to arrange the runs of the experiment in whole plots in such a way that the average level of the sub-plot factors within the whole plots is equal.

We have also investigated whether performing a GLS analysis when the 16-run CCD has been conducted in a random run order, i.e. when the CCD has been conducted as a RNR experiment. We have generated 10,000 different random run orders. On average, the number of whole plots obtained in this way was 11.63. The minimum and maximum number of whole plots equalled 6 and 16 respectively. If the case of 16 whole plots, the number of whole plots equals the number of observations and the resulting design is a completely randomized design. In that case, the whole plot error variance and the sub-plot error variance cannot be estimated separately. The simulation results are shown in Table 17. The large numbers of whole plots make that the differences between a GLS and an OLS analysis are not as large as in Table 14. The values of Δ_{OLS} and Δ_{GLS} for $\sigma^2 = 20$ and $\sigma^2 = 45$ are large compared to those in Table 14. This is an indication that performing a properly designed split-plot experiment is likely to produce better results than a RNR experiment.

Finally, we have simulated the situation in which the CCD was run and analyzed as a completely randomized experiment. For $\sigma^2 = 5$, 20 and 45, this gave deviations Δ of 0.5818, 1.3873 and 1.8543 respectively. Comparing these values to the deviations in the Tables 14, 16 and 17 shows that using a split-plot design and analyzing it by GLS or FGLS leads to smaller deviations than running a completely randomized design. This is especially true for large values of the variance ratio η. Even when the split-plot design is improperly analyzed using OLS, the deviations are not that different from those of a completely randomized design.

Table 15: 16-run CCD arranged in eight and six whole plots.

8 whole plots				6 whole plots			
	w	s_1	s_2		w	s_1	s_2
1	-1	-1	-1	1	-1	-1	-1
	-1	0	0		-1	1	1
2	-1	1	-1		-1	1	-1
	-1	1	1	2	1	-1	-1
3	1	-1	-1		1	1	-1
	1	0	0		1	-1	1
4	1	1	-1	3	-1	0	0
	1	-1	1		-1	-1	1
5	-1	-1	1	4	1	0	0
6	1	1	1		1	1	1
7	0	-1	0	5	0	-1	0
	0	0	0		0	0	0
	0	1	0		0	1	0
8	0	0	-1	6	0	0	-1
	0	0	0		0	0	0
	0	0	1		0	0	1

Table 16: Simulation results for the 16-run CCDs with six and eight whole plots displayed in Table 15 for various η- and σ^2-values.

		$\sigma^2 = 5$			$\sigma^2 = 20$			$\sigma^2 = 45$		
w	η	Δ_{OLS}	Δ_{GLS}	Δ_{FGLS}	Δ_{OLS}	Δ_{GLS}	Δ_{FGLS}	Δ_{OLS}	Δ_{GLS}	Δ_{FGLS}
6	1	0.6319	0.5418	0.5875	1.3837	1.3022	1.3437	1.8590	1.8018	1.8343
	2	0.6397	0.4712	0.5053	1.3729	1.2068	1.2390	1.8540	1.7177	1.7538
	4	0.6413	0.3984	0.4134	1.3617	1.0771	1.0932	1.8360	1.6008	1.6264
	8	0.6428	0.3418	0.3394	1.3440	0.9271	0.9188	1.8189	1.4538	1.4548
8	1	0.6383	0.5620	0.6441	1.3629	1.3119	1.3853	1.8384	1.8016	1.8994
	2	0.6535	0.5087	0.5845	1.3608	1.2384	1.3293	1.8371	1.7503	1.8507
	4	0.6640	0.4424	0.4995	1.3544	1.1390	1.2253	1.8541	1.6677	1.7798
	8	0.6702	0.3763	0.4135	1.3525	1.0191	1.0763	1.8433	1.5444	1.6261

Table 17: Simulation results for 10,000 random run orders of the 16-run CCD for various η- and σ^2-values.

	$\sigma^2 = 5$		$\sigma^2 = 20$		$\sigma^2 = 45$	
η	Δ_{OLS}	Δ_{FGLS}	Δ_{OLS}	Δ_{FGLS}	Δ_{OLS}	Δ_{FGLS}
1	0.5617	0.5683	1.3440	1.3380	1.8424	1.8184
2	0.5606	0.5481	1.3502	1.3140	1.8356	1.8204
4	0.5572	0.5201	1.3437	1.2909	1.8326	1.7717
8	0.5538	0.4873	1.3154	1.2145	1.8534	1.6940

5.2 Duplicating the central composite design

We have also performed simulations with a duplicated central composite design. Firstly, we have doubled the number of observations in each whole plot. As a result, the designs considered still possess four, six and eight whole plots, but the total number of observations has doubled. Such a design may only be slightly more cumbersome than the original non-replicated central composite design because the number of changes of the whole plot factor level remains small. Secondly, we have duplicated the entire central composite design, leading to designs with 8, 12 and 16 whole plots. These design options involve more changes in the whole plot factor levels compared to the original central composite designs. However, they may still be cheaper or easier to run than a completely randomized 16-run central composite design.

5.2.1 Doubling the whole plot sizes

Doubling the number of observations in every whole plot of the central composite design of course leads to smaller deviations from the optimum. It is particularly beneficial if GLS or FGLS estimation are used, $\eta = 1$ and σ^2 is large. If OLS estimation is utilized, the improvements are smaller and even decrease with η. Remarkably, the split-plot designs investigated often produce larger deviations than the duplicated completely randomized design, even if GLS or FGLS are used for the analysis of the split-plot designs. This is in contrast with the non-replicated CCDs.

5.2.2 Doubling the number of whole plots

The benefits of doubling the number of whole plots and leaving the whole plot sizes unchanged are larger. The beneficial effect is larger for the OLS results than for those based on GLS or FGLS. The GLS and FGLS deviations are smaller than the deviations obtained for the completely randomized design, indicating that running a split-plot design does not harm in terms of the optimum response achieved.

Table 18: Frequency with which a zero estimate is obtained for the whole plot error.

w	η	$\sigma^2 = 5$	$\sigma^2 = 20$	$\sigma^2 = 45$
4	1	0.2372	0.2369	0.2414
	2	0.2059	0.205	0.2019
	4	0.1584	0.1632	0.1719
	8	0.1293	0.1231	0.1294
6	1	0.1711	0.1766	0.1718
	2	0.1222	0.1237	0.1281
	4	0.0818	0.0868	0.0856
	8	0.0505	0.0484	0.0522
8	1	0.1398	0.1404	0.1389
	2	0.0935	0.0965	0.0987
	4	0.0594	0.0558	0.0589
	8	0.0394	0.0351	0.0361

6 Variance component estimation

A possible explanation of the small differences between Δ_{FGLS} and Δ_{OLS} is that the whole plot error variance σ_w^2 is often estimated to be zero. In that case, FGLS estimation reduces to OLS estimation. The frequencies of this happening for the non-replicated CCDs in Table 7 are displayed in Table 18. For the design option with four whole plots, an estimate of zero for σ_w^2 is obtained for almost one fourth of the simulated data sets generated with $\eta = 1$. The number of zeroes obtained diminishes with increasing η. The design options with six or eight whole plots, and thus with more degrees of freedom for estimating σ_w^2, produce substantially less zeroes. Despite the large numbers of zero estimates obtained, σ_w^2 is, on average, overestimated in 30 of the 36 situations investigated. This is shown in Table 19 which displays the average estimated error components. The average estimates can be compared to the true values used in the simulations, which can be obtained as $\sigma_\varepsilon^2 = \sigma^2/(1+\eta)$ and $\sigma_\gamma^2 = \eta\sigma^2/(1+\eta)$. In light of this overestimation, it is not surprising that the sub-plot error variance σ_s^2 is underestimated in 27 of the 36 cases. As a consequence of this, the variance ratio η is overestimated in 30 of the 36 instances. The magnitude of the overestimation of η decreases with the number of whole plots and with the variance ratio η. For $\eta = 8$, the variance ratio is even underestimated if six or eight whole plots are used. Notice that exactly the same pattern is obtained for the three σ^2-values investigated. These simulation results thus show that the FGLS approach leads to an overestimation of η and this makes that the optima obtained from this analysis method are closer to those obtained from a GLS analysis than to those obtained by applying OLS estimation which uses a zero η-value.

The conclusion of these simulation results is of course that the estimation of the

Table 19: Average estimates for the whole plot error σ_w^2 and the sub-plot error σ_s^2.

		$\sigma^2 = 5$			$\sigma^2 = 20$			$\sigma^2 = 45$		
w	η	σ_w^2	σ_s^2	$\hat{\eta}$	σ_w^2	σ_s^2	$\hat{\eta}$	σ_w^2	σ_s^2	$\hat{\eta}$
4	1	3.6209*	2.3566	1.5365*	14.4835*	9.4272	1.5364*	32.5785*	21.2100	1.5360*
	2	4.0102*	1.5883	2.5248*	16.0354*	6.3527	2.5242*	36.0819*	14.2947	2.5241*
	4	4.3752*	0.9642	4.5376*	17.4970*	3.8568	4.5367*	39.3793*	8.6781	4.5378*
	8	4.6566*	0.5415	8.5994*	18.6262*	2.1661	8.5990*	41.9038*	4.8736	8.5981*
6	1	3.2524*	2.2363	1.4544*	13.1266*	8.9267	1.4705*	29.4110*	20.1039	1.4629*
	2	3.6815*	1.5548	2.3678*	14.7888*	6.2090	2.3818*	33.3562*	13.9560	2.3901*
	4	4.0826*	0.9759	4.1834*	16.3408*	3.9003	4.1896*	36.8619*	8.7616	4.2072*
	8	4.3779	0.5697*	7.6846	17.5349	2.2737*	7.7121	39.4568	5.1165*	7.7117
8	1	3.0732*	2.2266	1.3802*	12.2837*	8.9060	1.3793*	27.6264*	20.0380	1.3787*
	2	3.5986*	1.5747	2.2853*	14.3933*	6.2977	2.2855*	32.3803*	14.1713	2.2849*
	4	4.0800*	1.0004*	4.0784*	16.3196*	4.0023*	4.0776	36.7120*	9.0026*	4.0779*
	8	4.4135	0.5928*	7.4452	17.6525	2.3720*	7.4420	39.7185	5.3341*	7.4461

* Overestimated.

variance components is highly inefficient. Even large whole plot error variances are often not detected, so that OLS inevitably has to be used. In other cases, the whole plot error variance is by far overestimated.

7 Discussion and recommendations

In this paper, an extensive simulation study was described to investigate the impact of several analysis approaches to small response surface split-plot experiments like factorial experiments, central composite experiments and D-optimal experiments on constrained design regions. The simulation results showed that a proper generalized least squares analysis of the experimental results produced by a split-plot design does not allow researchers to discover many whole plot effects. Even large whole plot effects are identified as insignificant, so that the power of the split-plot design used in industry is small for detecting whole plot effects with generalized least squares. If ordinary least squares is used, linear whole plot effects are detected much more often and the number of nonzero quadratic whole plot effects identified is substantially larger than when generalized least squares is utilized. For detecting sub-plot effects, analyzing the data using generalized least squares and using the residual degrees of freedom as the denominator degrees of freedom yields the best results. Using more advanced methods for determining the denominator degrees of freedom for the significance tests, like for instance the methods of Satterthwaite and Kenward & Roger, leads to a smaller number of identified active effects.

If the interest is in identifying nonzero effects, it is tempting to rely on ordinary least squares for detecting active whole plot effects and to use generalized least squares for detecting other effects. A more correct approach would however be to use a different level of significance for whole plot effects. As the simulation results show that, overall, the correct p-value is approximated best by using a generalized least squares analysis combined with the Kenward and Roger (1997) degrees of freedom method, we recommend using this approach to compute p-values. Using a level of significance of 20% for the whole plot effects instead of the usual 5% would, for the factorial design in Table 3 and $\eta = 1$, lead to a detection of the whole plot effect in 66.8% of the instances compared to 25.4% reported in Table 10. For $\eta = 8$, the likelihood of detecting the whole plot effect would then be 54.9%. In order to attain 95%, a level of significance of more than 60% would have to be used. A simple alternative might be to assume that any whole plot effect is significant and carry it over to the next stage of experimentation. For the sub-plot effects, a GLS estimation approach combined with Kenward and Roger's (1997) approach to determine the denominator degrees of freedom for the significance tests is recommended for inference purposes. This is because this method leads to the best results in terms of approximating the 5% type I error rate. For quadratic sub-plot effects, increasing the level of significance might be useful as not many such effects would be detected otherwise.

Another result of our study is that there is not a huge difference between the optimum settings determined from a model estimated by ordinary least squares and one estimated by generalized least squares. This is especially the case when the observations of the split-plot experiment are correlated only to a small extent. It is also interesting to point out that running a split-plot design in many cases leads to smaller deviations from the optimal response than running a completely randomized designs. This will even be more so if split-plot designs with a larger number of observations are compared to completely randomized designs. Such a comparison is not at all unfair because split-plot designs are much easier and cheaper to run than fully randomized experiments, even if they possess more observations.

A final conclusion from the simulation studies is that a proper assignment of the experimental runs to the whole plots leads to better results in terms of finding the optimum settings for a process. Arranging the runs in the whole plots such that the average levels of the regressors is the same in every whole plot works well.

Acknowledgement

The research that was described in this paper was carried out while the first author was a postdoctoral researcher of the Fund for Scientific Research Flanders (Belgium) and while the second author was working on a research project funded by the same institution. The authors would like to thank Steven Gilmour and Geert Verbeke for their comments on earlier drafts of the paper.

Appendix A

The SAS commands needed to analyze the wrapper machine data are:

```
data wrapper;
input wp spacing speed temp y;
datalines;
1  0  1  -1  5.005
...
4 0  0  0  8.195
;
proc glm;
model y = spacing|speed|temp@2
          spacing*spacing speed*speed temp*temp / solution;
proc mixed;
class wp;
model y = spacing|speed|temp@2
          spacing*spacing speed*speed temp*temp / ddfm=kr solution;
random wp;
run;
```

The options for determining the degrees of freedom in the MIXED procedure are CONTAIN, BETWITHIN, RESIDUAL, SATTERTH and KR (or KENWARDROGER).

Appendix B

The usual estimator for the residual error variance in OLS regression is $\mathbf{e}'\mathbf{e}/(n-p)$. If OLS is used for the estimation, the residual vector for a split-plot experiment is

$$\begin{aligned}
\mathbf{e} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}} \\
&= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \\
&= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}), \\
&= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}), \\
&= \mathbf{M}(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}),
\end{aligned}$$

where $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. As a result,

$$\begin{aligned}
\mathrm{E}(\mathbf{e}'\mathbf{e}) &= \mathrm{E}\{(\boldsymbol{\varepsilon}' + \boldsymbol{\gamma}'\mathbf{Z}')\mathbf{M}(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})\}, \\
&= \mathrm{E}\{(\boldsymbol{\varepsilon}' + \boldsymbol{\gamma}'\mathbf{Z}')\mathbf{M}(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})\}, \\
&= \mathrm{E}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) + \mathrm{E}(\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}), \\
&= \sigma_\varepsilon^2(n-p) + \sigma_\gamma^2 \mathrm{trace}(\mathbf{Z}'\mathbf{M}\mathbf{Z}), \\
&= \sigma_\varepsilon^2(n-p) + \sigma_\gamma^2[n - \mathrm{trace}\{\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\}].
\end{aligned}$$

Dividing this expression by $n-p$ yields (8).

Appendix C

The variance-covariance matrix of the GLS estimator (4) equals

$$\begin{aligned}
&(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&\quad = \sigma_\varepsilon^2(\mathbf{X}'(\mathbf{I}_n - \eta\mathbf{Z}(\mathbf{I}_b + \eta\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{X})^{-1}, \\
&\quad = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X} - \eta\mathbf{X}'\mathbf{Z}(\mathbf{I}_b + \eta\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}, \\
&\quad = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma_\gamma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{I}_b + \eta\mathbf{Z}'\mathbf{Z} - \eta\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \\
&\quad = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma_\gamma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{I}_b + \eta\mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \\
&\quad = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma_\gamma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&\qquad\quad - \sigma_\gamma^2\eta(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{I}_b + \mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Z}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},
\end{aligned} \tag{17}$$

whereas the true variance-covariance matrix of the OLS estimator, given by (6), is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma_\gamma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

References

Anbari, F. T. and Lucas, J. M. (1994). Super-efficient designs: How to run your experiment for higher efficiency and lower cost, *ASQC Technical Conference Transactions*, pp. 852–863.

Bingham, D. R. and Sitter, R. R. (2001). Design issues in fractional factorial split-plot designs, *Journal of Quality Technology* **33**: 2–15.

Bingham, D. and Sitter, R. R. (1999). Minimum-aberration two-level fractional factorial split-plot designs, *Technometrics* **41**: 62–70.

Bisgaard, S. (2000). The design and analysis of $2^{k-p} \times 2^{q-r}$ split plot experiments, *Journal of Quality Technology* **32**: 39–56.

Bisgaard, S. and Steinberg, D. M. (1997). The design and analysis of $2^{k-p} \times s$ prototype experiments, *Technometrics* **39**: 52–62.

Box, G. E. P. and Jones, S. P. (1992). Split-plot designs for robust product experimentation, *Journal of Applied Statistics* **19**: 3–26.

Cornell, J. A. (1988). Analyzing data from mixture experiments containing process variables: A split-plot approach, *Journal of Quality Technology* **20**: 2–23.

Davison, J. J. (1995). *Response Surface Designs and Analysis for Bi-Randomiza-tion Error Structures*, Ph.D. thesis, Virginia Polytechnic Institute and State University.

Ganju, J. and Lucas, J. M. (1997). Bias in test statistics when restrictions in randomization are caused by factors, *Communications in Statistics: Theory and Methods* **26**: 47–63.

Ganju, J. and Lucas, J. M. (1999). Detecting randomization restrictions caused by factors, *Journal of Statistical Planning and Inference* **81**: 129–140.

Ganju, J. and Lucas, J. M. (2004). Randomized and random run order experiments. To appear.

Goos, P. (2002). *The Optimal Design of Blocked and Split-plot Experiments*, New York: Springer.

Goos, P. and Vandebroek, M. (2001). Optimal split-plot designs, *Journal of Quality Technology* **33**: 436–450.

Goos, P. and Vandebroek, M. (2003). D-optimal split-plot designs with given numbers and sizes of whole plots, *Technometrics* **45**: 235–245.

Goos, P. and Vandebroek, M. (2004). Outperforming completely randomized designs. To appear.

Huang, P., Chen, D. and Voelkel, J. (1998). Minimum-aberration two-level split-plot designs, *Technometrics* **40**: 314–326.

Ju, H. L. and Lucas, J. M. (2002). L^{k} factorial experiments with hard-to-change and easy-to-change factors, *Journal of Quality Technology* **34**: 411–421.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*, New York: Wiley.

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

Kowalski, S. M., Cornell, J. A. and Vining, G. G. (2002). Split-plot designs and estimation methods for mixture experiments with process variables, *Technometrics* **44**: 72–79.

Letsinger, J. D., Myers, R. H. and Lentner, M. (1996). Response surface methods for bi-randomization structures, *Journal of Quality Technology* **28**: 381–397.

Myers, R. H., Montgomery, D. C., Vining, G. G., Borror, C. M. and Kowalski, S. M. (2004). Response surface methodology: A retrospective and current literature survey. To appear.

Nelson, L. S. (1985). What do low F ratios tell you?, *Journal of Quality Technology* **17**: 237–238.

Simpson, J., Kowalski, S. and Landman, D. (2003). Experimentation with randomization restrictions: Targeting practical implementation, Paper presented at the 47th Annual Fall Technical Conference, El Paso.

Trinca, L. A. and Gilmour, S. G. (2001). Multi-stratum response surface designs, *Technometrics* **43**: 25–33.

Vining, G. G., Kowalski, S. M. and Montgomery, D. C. (2003). Response surface designs within a split-plot structure. Unpublished manuscript.

Webb, D., Lucas, J. M. and Borkowski, J. J. (2004). Factorial experiments when factor levels are not necessarily reset. To appear.