Testing for Lack of Fit in Blocked and Split-Plot Response Surface Designs

Peter Goos
Universiteit Antwerpen, Belgium
Erasmus Universiteit, Rotterdam, the Netherlands

Steven G. Gilmour
University of Southampton, UK

June 21, 2012

Abstract

Textbooks on response surface methodology emphasize the importance of lack-of-fit tests when fitting response surface models, and stress that, to be able to test for lack of fit, designed experiments should have replication and allow for pure-error estimation. In this paper, we show how to obtain pure-error estimates and how to carry out a lack-of-fit test when the experiment is not completely randomized, but a blocked experiment, a split-plot experiment, or any other multi-stratum experiment. Our approach to calculating pure-error estimates is based on residual maximum likelihood (REML) estimation of the variance components in a full treatment model. It generalizes the one suggested by Vining et al. (2005) in the sense that it works for a broader set of designs and for replicates other than center point replicates. Our lack-of-fit test also generalizes the test proposed by Khuri (1992) for data from blocked experiments because it exploits replicates other than center point replicates and works for split-plot and other multi-stratum designs as well. We provide analytical expressions for the test statistic and the corresponding degrees of freedom, and demonstrate how to perform the lack-of-fit test in the SAS procedure MIXED. We re-analyze several published data sets and discover a few instances in which the usual response surface model exhibits significant lack of fit.

Keywords: Kenward-Roger degrees of freedom, multi-stratum design, replication, residual maximum likelihood (REML), split-split-plot design, treatment model.

1 Introduction

When analysing data using empirical response surface models, it is often desirable to allow detection of failures of assumptions. In particular, an analysis which allows separation of lack of fit from pure error is useful. When experiments are completely randomized, this is easily accomplished since the pure-error estimate is obtained from replicate points, and is implemented in several packages for analysing experimental data - see Box and Draper (2007) for a full explanation.

In blocked response surface designs, when the block effects are taken as fixed, more care is needed with the definition of pure error, but the most reasonable, discussed in detail by Gilmour and Trinca (2000), is that it is the expectation of the residual mean square from the block-treatment model. In this model, each combination of factor levels used in the experiment is taken to be a discrete treatment. It is sometimes desirable to treat block effects as random and similar models are used with split-plot and other multi-stratum structures. The purpose of this paper is to show how a test for lack of fit can be conducted in response surface models with these structures.

Khuri (1992) tested for lack of fit with random block effects, but based his pure-error estimates only on replicated center points, although we will see that extending the definition of Gilmour and Trinca (2000) allows more precise pure-error estimation. Vining et al. (2005) and Vining and Kowalski (2008) recommended a simple analysis based on estimation of each variance component using the sample variance obtained from replicate points. However, this method is only applicable to particular types of design and only uses replicate points within whole plots and completely replicated whole plots to obtain pure-error estimates. Gilmour and Trinca (2000) showed, in the context of blocked response surface designs, that this is a stronger definition of pure error than is used in completely randomized designs, which requires only the use of the full treatment model. Parker et al. (2007) noted that the pure-error estimates could also be used to test for lack of fit, but did not give detailed explanation of how this could be done.

Almimi et al. (2009) pointed out the importance of developing procedures for checking the adequacy of fit of split-plot models. To do so, they proposed the use of two different coefficients of determination or R^2 values, one for the whole-plot stratum in the analysis and one for the subplot stratum. Similarly, they suggest PRESS values for both the whole-plot and subplot strata. However, they do not present a formal lack-of-fit test for models estimated from split-plot experimental data. In this paper, we show how to carry out formal lack-of-fit tests for random block models and split-plot models. We apply the test to several data sets described in the literature and use a simulated data set to show that the test can be extended for use in a split-split-plot design.

2 Model

In any experimental design, blocking factors arise from restrictions to the randomization, so that particular sets of treatments must appear together in blocks. Unless each block consists of the same set of treatments, some information for comparing treatments is confounded with block effects. If the order of the blocks is randomized, we can use random block effects in the model to recover this inter-block information. Split-plot designs have treatments defined by combinations of the levels of several factors applied in two strata, i.e. some factors have main effects completely confounded with the effects of blocks. Hence, blocked reponse surface designs with random block effects and split-plot response surface designs have exactly the same structure and model, the only difference being that in the former no main effects are completely confounded with block effects.

We assume that the model is

$$\mathbf{Y} = \mathbf{X}_t\boldsymbol{\tau} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y}$ is a random variable of which the response vector $\mathbf{y}$ is assumed to be a realization, $\mathbf{X}_t$ is the full treatment design matrix, having (i,t)th element equal to 1 if treatment t appears in run i and 0 otherwise, $\boldsymbol{\tau}$ is the corresponding vector of treatment means, $\boldsymbol{\delta}$ is a vector of random block or whole-plot effects, $\mathbf{Z}$ is the design matrix for these random effects and $\boldsymbol{\epsilon}$ is the vector of random experimental unit errors. We further assume that $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_1^2\mathbf{I})$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_0^2\mathbf{I})$ and that $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are independent. We refer to model (1) as the full treatment model.

In a typical response surface experiment, we want to further interpret the treatment effects, for example by assuming that

$$\mathbf{X}_t\boldsymbol{\tau} = \mathbf{X}\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{X}$ is the model matrix for a polynomial regression model and $\boldsymbol{\beta}$ is the vector of parameters of this model. Obviously, adopting the polynomial regression model is a much stronger assumption than that of model (1), which allows any pattern of treatment effects. Therefore, we would like to be able to test for lack of fit of this polynomial model.

3 Estimation and Testing

3.1 Estimated Standard Errors of Fixed Effects

In response surface studies, the main interest is usually in estimating the fixed effects $\boldsymbol{\beta}$ in the polynomial regression model (2). However, to test for lack of fit, we must fit the

full treatment model (1). This is usually done using the empirical GLS estimator

$$\hat{\boldsymbol{\tau}} = (\mathbf{X}_t'\hat{\mathbf{V}}^{-1}\mathbf{X}_t)^{-1}\mathbf{X}_t'\hat{\mathbf{V}}^{-1}\mathbf{Y}, \tag{3}$$

where

$$\hat{\mathbf{V}} = \hat{\sigma}_1^2\mathbf{Z}'\mathbf{Z} + \hat{\sigma}_0^2\mathbf{I},$$

and $\hat{\sigma}_1^2$ and $\hat{\sigma}_0^2$ are the estimators of the variance components obtained from residual maximum likelihood (REML) (McCulloch et al., 2008) applied to the full treatment model. The variance matrix of these estimators is usually estimated by

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\tau}}) = \hat{\boldsymbol{\Psi}} = (\mathbf{X}_t'\hat{\mathbf{V}}^{-1}\mathbf{X}_t)^{-1}. \tag{4}$$

The corresponding estimated standard errors of the fixed effects' estimates are known to be negatively biased. A correction, which usually gives much less biased estimated standard errors, was suggested by Kenward and Roger (1997).

Simple orthogonal block structures are those made up of crossed and nested blocking factors, in which each block contains equal numbers of units (Nelder (1965); see also Gilmour and Trinca (2006)), irrespective of the treatment structure or model. In simple orthogonal block structures, the approximate variance matrix for fixed effects, with the Kenward-Roger correction, is

$$\widehat{\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})} = \hat{\boldsymbol{\Psi}}_A = \hat{\boldsymbol{\Psi}} + 2\hat{\boldsymbol{\Lambda}}, \tag{5}$$

where $\hat{\boldsymbol{\Lambda}}$ is obtained by plugging the REML estimators of the variance components into

$$\boldsymbol{\Lambda} = \boldsymbol{\Psi}\sum_{i=0}^{1}\sum_{j=0}^{1}\left\{u_{ij}\left(\mathbf{Q}_{ij} - \mathbf{P}_i\boldsymbol{\Psi}\mathbf{P}_j\right)\right\}\boldsymbol{\Psi},$$

$\boldsymbol{\Psi} = (\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t)^{-1}$, $u_{ij} = Cov(\hat{\sigma}_i^2, \hat{\sigma}_j^2)$, $i, j \in \{0, 1\}$, $\hat{\sigma}_i^2$ is the estimator of σ_i^2,

$$\mathbf{P}_i = \mathbf{X}_t'\frac{\partial\mathbf{V}^{-1}}{\partial\sigma_i^2}\mathbf{X}_t$$

and

$$\mathbf{Q}_{ij} = \mathbf{X}_t'\frac{\partial\mathbf{V}^{-1}}{\partial\sigma_i^2}\mathbf{V}\frac{\partial\mathbf{V}^{-1}}{\partial\sigma_j^2}\mathbf{X}_t.$$

Hence

$$\mathbf{V}^{-1} = \frac{1}{\sigma_0^2}\mathbf{I} - \frac{\sigma_1^2}{\sigma_0^4 + k\sigma_1^2\sigma_0^2}\mathbf{Z}\mathbf{Z}'. \tag{6}$$

By direct differentiation, we obtain

$$\frac{\partial \mathbf{V}^{-1}}{\partial \sigma_1^2} = -\frac{1}{\left(\sigma_0^2 + k\sigma_1^2\right)^2}\mathbf{Z}\mathbf{Z}'$$

and

$$\frac{\partial \mathbf{V}^{-1}}{\partial \sigma_0^2} = \frac{1}{\sigma_0^4}\left\{\frac{\sigma_1^2\left(2\sigma_0^2 + k\sigma_1^2\right)}{\left(\sigma_0^2 + k\sigma_1^2\right)^2}\mathbf{Z}\mathbf{Z}' - \mathbf{I}\right\}.$$

We use the asymptotic sampling variance of the REML estimators of variance components, given by McCulloch et al. (2008) for example,

$$u_{00} = V\left(\hat{\sigma}_0^2\right) = 2tr(\mathbf{Z}'\mathbf{C}\mathbf{Z}\mathbf{Z}'\mathbf{C}\mathbf{Z})/c,$$

$$u_{11} = V\left(\hat{\sigma}_1^2\right) = 2tr(\mathbf{C}\mathbf{C})/c$$

and

$$u_{01} = u_{10} = Cov\left(\hat{\sigma}_0^2, \hat{\sigma}_1^2\right) = -2tr(\mathbf{Z}'\mathbf{C}\mathbf{C}\mathbf{Z})/c,$$

where

$$c = tr(\mathbf{C}\mathbf{C})tr(\mathbf{Z}'\mathbf{C}\mathbf{Z}\mathbf{Z}'\mathbf{C}\mathbf{Z}) - \left\{tr(\mathbf{Z}'\mathbf{C}\mathbf{C}\mathbf{Z})\right\}^2$$

and

$$\mathbf{C} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}_t\left(\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t\right)^{-1}\mathbf{X}_t'\mathbf{V}^{-1}.$$

3.2 Testing for Lack of Fit

In completely randomized response surface designs, it is common practice to carry out a hypothesis test to check for lack of fit of the second order model (Box and Draper, 2007); it is also straightforward to do in blocked response surface designs with fixed block effects (Gilmour and Trinca, 2000). The extension to random block effects and split-plot designs is not trivial. One possibility would be to perform a likelihood ratio test to compare the polynomial model with the full treatment model. Although it should have good asymptotic properties, such a test suffers from problems in realistic sized experiments. Instead, we recommend using the approximate F-test proposed by Kenward and Roger (1997), which uses their adjusted estimated variance-covariance matrix in Wald-type test statistics.

We rewrite the full treatment model by separating the polynomial model parameters,

$$\mathbf{X}_t\boldsymbol{\tau} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_l\mathbf{L}'\boldsymbol{\tau} = \begin{bmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_l\end{bmatrix}\boldsymbol{\tau}^*,$$

where $\boldsymbol{\tau}^{*\prime} = [\boldsymbol{\beta}' \;\; \boldsymbol{\tau}'\mathbf{L}]$ and $\mathbf{L}$ has dimensions $l \times t$. The additional terms $\mathbf{L}'\boldsymbol{\tau}$ represent higher order polynomial terms, though it is not essential for them to be parameterized in this way.

The lack-of-fit test should test the null hypothesis $H_0 : \; \mathbf{L}'\boldsymbol{\tau} = \mathbf{0}$ against the alternative $H_1 : \; \mathbf{L}'\boldsymbol{\tau} \neq \mathbf{0}$. The natural test statistic is

$$F = \frac{1}{l}\hat{\boldsymbol{\tau}}'\mathbf{L}\left(\mathbf{L}'\hat{\boldsymbol{\Psi}}_A\mathbf{L}\right)^{-1}\mathbf{L}'\hat{\boldsymbol{\tau}},$$

but this does not reduce to the standard F-test when the latter is appropriate, e.g. in orthogonal designs. Kenward and Roger derived an approximation which does and, in our case, this gives the test statistic $F^* = \lambda F$, where

$$\lambda = \frac{m}{E^*(m-2)},$$

$$m = 4 + \frac{l+2}{l\rho - 1},$$

$$\rho = \frac{V^*}{2E^{*2}},$$

$$E^* = \frac{l}{l - A_2},$$

$$V^* = \frac{2(1 + c_1 B)}{l(1 - c_2 B)^2(1 - c_3 B)},$$

$$c_1 = \frac{g}{3l + 2(1-g)},$$

$$c_2 = \frac{l - g}{3l + 2(1-g)},$$

$$c_3 = \frac{l + 2 - g}{3l + 2(1-g)},$$

$$g = \frac{(l+1)A_1 - (l+4)A_2}{(l+2)A_2},$$

$$A_1 = \sum_{i=0}^{1}\sum_{j=0}^{1} u_{ij}\mathrm{tr}(\boldsymbol{\Theta\Psi P}_i\boldsymbol{\Psi})\mathrm{tr}(\boldsymbol{\Theta\Psi P}_j\boldsymbol{\Psi}),$$

$$A_2 = \sum_{i=0}^{1}\sum_{j=0}^{1} u_{ij}\mathrm{tr}(\boldsymbol{\Theta\Psi P}_i\boldsymbol{\Psi\Theta\Psi P}_j\boldsymbol{\Psi})$$

and

$$\boldsymbol{\Theta} = \mathbf{L}(\mathbf{L}'\boldsymbol{\Psi}\mathbf{L})^{-1}\mathbf{L}'.$$

Under H_0, F^* has approximately an F distribution with l and m degrees of freedom. This test for lack of fit is easily programmed, so that checking for lack of fit in a randomized block or split-plot response surface design becomes almost as simple as in a completely randomized response surface design. Also, the lack-of-fit test can be readily performed in the SAS procedure MIXED. We refer to the Appendix for two example SAS programs.

4 Examples

4.1 Pastry dough experiment

Gilmour and Ringrose (1999) and Gilmour and Trinca (2000) describe a pastry dough experiment carried out in the Department of Food Science and Technology at the University of Reading. The factors investigated in the experiment were the feed flow rate (x_1), the initial moisture content (x_2) and the screw speed (x_3) of a mixing process for pastry dough. The goal of the experiment was to acquire an understanding of how the various properties of a dough depend on the settings of the three factors and to develop an overall control scheme for the process. The experiment involved seven days, on each of which four runs were performed. The design for the experiment was obtained using the blocking algorithm of Trinca and Gilmour (2000). It is displayed in Table 1, along with five of the responses: a longitudinal expansion index (y_1) and a cross-sectional expansion index (y_2), and three variables representing the color of the pastry using the CIE 1976 (L^* a^* b^*) color system (Commission internationale de l'éclairage, 1986), the lightness (y_3), the redness (y_4) and the yellowness (y_5). The redness response has not previously been analyzed in the literature. The design involves 15 distinct factor level combinations, labeled 1–15 in Table 1. As a result, the full treatment model required for the pure-error estimates of the variance components and the lack-of-fit test involves 15 parameters.

For each of the five responses, we tested the second-order response surface model for lack of fit. The pure-error estimates of the block error variance σ_1^2 and the residual error variance σ_0^2, along with the denominator degrees of freedom, F test statistic and p-value for the lack-of-fit test, are given in Table 2. For the purpose of comparison, we have also shown the variance component estimates obtained by using the response surface model in the table. Note that the numerator degrees of freedom for the lack-of-fit tests equal 5 for each of the five responses. This is because the design involves 15 different factor level combinations or treatments and the response surface model has 10 unknown parameters.

Response y_4 is the only one for which there is a significant lack of fit, giving a p-value of 0.0345. By adding either of the linear-by-quadratic interaction terms $x_1x_2^2$ or $x_1x_3^2$ to the full quadratic model, we can get rid of the significant lack of fit. In that case,

Table 1: Design and response data for the pastry dough experiment

Block	Treatment	x_1	x_2	x_3	y_1	y_2	y_3	y_4	y_5
1	1	−1	−1	−1	15.0	6.14	77.89	0.20	11.46
1	15	0	0	0	13.0	4.97	77.31	0.12	11.93
1	15	0	0	0	11.7	5.41	77.91	0.13	11.63
1	8	1	1	1	14.8	4.83	78.10	0.09	11.32
2	4	−1	1	1	11.2	4.25	76.93	0.26	12.17
2	15	0	0	0	12.2	3.86	77.51	0.16	11.85
2	15	0	0	0	11.6	4.34	77.38	0.05	11.64
2	5	1	−1	−1	14.1	4.93	77.96	0.02	11.28
3	2	−1	−1	1	15.9	6.26	78.68	−0.05	10.76
3	3	−1	1	−1	10.8	3.92	77.74	−0.02	14.41
3	5	1	−1	−1	15.6	4.92	76.90	0.11	12.27
3	8	1	1	1	15.8	5.48	77.24	−0.04	12.13
4	9	−1	0	0	11.2	4.36	76.99	0.31	13.33
4	13	0	0	−1	12.7	4.12	76.72	−0.15	14.19
4	12	0	1	0	11.4	4.24	76.34	−0.05	13.84
4	6	1	−1	1	18.6	6.11	78.07	0.20	10.55
5	3	−1	1	−1	10.1	4.35	76.79	0.24	14.22
5	11	0	−1	0	13.0	5.02	76.75	0.07	12.35
5	14	0	0	1	11.1	4.32	77.64	0.10	12.54
5	10	1	0	0	11.7	4.18	76.70	0.10	13.50
6	1	−1	−1	−1	14.6	5.85	77.00	−0.08	12.92
6	4	−1	1	1	12.8	4.89	76.73	0.00	13.91
6	6	1	−1	1	17.6	6.67	78.38	0.14	11.66
6	7	1	1	−1	15.4	4.80	77.19	−0.04	14.48
7	2	−1	−1	1	15.0	6.38	77.74	−0.02	12.20
7	15	0	0	0	10.7	4.21	76.97	0.00	14.94
7	15	0	0	0	9.6	4.29	76.97	0.02	14.61
7	7	1	1	−1	10.9	4.30	77.19	0.08	14.78

the F test statistic and the p-value equal 2.74 and 0.1076, respectively. Both linear-by-quadratic interaction effects lead to the same p-value for the lack-of-fit test because they are completely aliased with each other in this design. The next smallest p-value is for y_1. Unlike for the other responses, the two sets of variance component estimates for responses y_1 and y_4 are somewhat different. In both cases, σ_0^2 seems to be overestimated in the response surface model, presumably due to contamination by higher-order effects. For the other responses, the variance component estimates are not much different in the two models, which strongly suggests that higher-order terms are not needed in the model

Table 2: Results for the lack-of-fit test and variance component estimates for the data from the pastry dough experiment

	Lack of fit			Pure error		Response surface	
Response	df	F value	p-value	σ_1^2	σ_0^2	σ_1^2	σ_0^2
y_1	10.5	2.38	0.1101	2.0596	1.3752	1.8100	2.3246
y_2	9.92	0.67	0.6557	0.1970	0.2977	0.1799	0.2904
y_3	9.09	0.51	0.7626	0.1178	0.1258	0.1408	0.1003
y_4	7.03	4.63	0.0345	0.0124	0.0033	0.0012	0.0107
y_5	8.18	1.71	0.2360	0.9782	0.0721	0.9703	0.0970

for these responses.

4.2 Galvanized steel experiment

Khuri (1992) analyzes data from an experiment in which the impact of two factors, temperature (x_1) and curing time (x_2), on the shear strength y of the bonding of galvanized steel bars was investigated. The two factors had three levels each, 375, 400 and 450°F for temperature and 30, 35 and 40 seconds for curing time. These levels were coded as -1, 0 and 2 for the first factor, and -1, 0 and 1 for the second factor. The design involved nine different treatment combinations, and included twelve blocks. Eight of these blocks had nine runs, one for each treatment combination. Two of the four remaining blocks had twelve runs, and the remaining two blocks had eleven runs. The difference in size of the blocks was entirely due to replications of the center run in the latter four blocks. The design and the data are shown in Table 3.

For these data, Khuri (1992) reports the results for a test for lack of fit. His test is based on the replicated observations within the blocks, i.e. on the replicated center points only. With his test involving 85 degrees of freedom for lack of fit and ten degrees of freedom for pure error, Khuri obtains a p-value of 0.225. Our test for lack of fit differs from Khuri's because it is based on the full treatment model instead of on the response surface model and because it also exploits the replication of treatment combinations other than the center run. Our test uses three degrees of freedom for lack of fit and 98.9 degrees of freedom for pure error, and results in a test statistic of 3.10 and a p-value of 0.0301. Hence, unlike Khuri's, our test suggests a significant lack of fit. The lack of fit can be accounted for by adding a linear-by-quadratic interaction term, $x_1x_2^2$, to the response surface model. The test statistic for the corresponding lack-of-fit test (involving two degrees of freedom for lack of fit and 99.1 for pure error) drops to 2.72, giving a p-value of 0.0708.

Table 3: Design and response data for the galvanized steel experiment

Treat-	Factor		Block											
ment	x_1	x_2	1	2	3	4	5	6	7	8	9	10	11	12
1	−1	−1	1226	1075	1172	1213	1282	1142	1281	1305	1091	1281	1305	1207
2	0	−1	1898	1790	1804	1961	1940	1699	1833	1774	1588	1992	2011	1742
3	2	−1	2142	1843	2061	2184	2095	1935	2116	2133	1913	2213	2192	1995
4	−1	0	1472	1121	1506	1606	1572	1608	1502	1580	1343	1691	1584	1486
5	0	0	2010	2175	2279	2450	2291	2374	2417	2393	2205	2142	2052	2339
5	0	0	1882			2355					2268		2032	
5	0	0	1915			2420					2103		2190	
5	0	0	2106			2240								
6	2	0	2352	2274	2168	2298	2147	2413	2430	2440	2093	2208	2201	2216
7	−1	1	1491	1691	1707	1882	1741	1846	1645	1688	1582	1692	1744	1751
8	0	1	2078	2513	2392	2531	2366	2392	2392	2413	2392	2488	2392	2390
9	2	1	2531	2588	2617	2609	2431	2408	2517	2604	2477	2601	2588	2572

In this example, the pure-error estimates of σ_1^2 and σ_0^2 are 3630.80 and 11813, respectively, whereas the estimates obtained from the second-order response surface model are 3521.59 and 12238, respectively. Hence, using the response surface model leads to an overestimation of σ_0^2, when compared to the full treatment model.

4.3 Strength of ceramic pipes

The experiment on ceramic pipes reported by Vining et al. (2005) has 12 whole plots, each with four runs, and three complete blocks consisting of replicated center points. The experimental factors were zone-1 temperature (x_1), zone-2 temperature (x_2), amount of binder (x_3) and grinding speed (x_4). The former two factors were whole-plot factors, while the latter two were subplot factors. The design for the ceramic pipe experiment was based on a four-factor central composite design. Hence, it involves 25 distinct factor level combinations or treatments. The design and the response data are shown in Table 4. The response, y, was the strength of a ceramic pipe.

When fitting a second-order response surface model to the ceramic pipe data, there is no evidence of lack of fit. The F test statistic equals 1.13, while the numerator and denominator degrees of freedom amount to 10 and 6.96, respectively. This results in a p-value of 0.4499. The pure-error estimates of the variance components σ_1^2 and σ_0^2 are 0.5263 and 0.0936, while those obtained from fitting the response surface model equal 1.4176 and 0.0756, respectively. The numerator degrees of freedom equal 10 because there are 25 treatments in the design and 15 parameters in the second-order response surface model.

Table 4: Design and response data for the ceramic pipe experiment

Block	Treatment	x_1	x_2	x_3	x_4	y	Block	Treatment	x_1	x_2	x_3	x_4	y
1	1	−1	−1	−1	−1	80.40	7	10	0	−1	0	0	80.07
1	2	−1	−1	−1	1	89.91	7	10	0	−1	0	0	80.79
1	3	−1	−1	1	−1	71.88	7	10	0	−1	0	0	80.20
1	4	−1	−1	1	1	76.87	7	10	0	−1	0	0	79.95
2	17	1	−1	−1	−1	87.48	8	16	0	1	0	0	68.98
2	18	1	−1	−1	1	90.84	8	16	0	1	0	0	68.64
2	19	1	−1	1	−1	84.49	8	16	0	1	0	0	69.24
2	20	1	−1	1	1	83.61	8	16	0	1	0	0	69.20
3	6	−1	1	−1	−1	62.99	9	11	0	0	−1	0	78.56
3	7	−1	1	−1	1	79.91	9	12	0	0	0	−1	74.59
3	8	−1	1	1	−1	49.95	9	14	0	0	0	1	82.52
3	9	−1	1	1	1	63.23	9	15	0	0	1	0	68.63
4	22	1	1	−1	−1	73.06	10	13	0	0	0	0	74.86
4	23	1	1	−1	1	84.45	10	13	0	0	0	0	74.22
4	24	1	1	1	−1	66.13	10	13	0	0	0	0	74.06
4	25	1	1	1	1	73.29	10	13	0	0	0	0	74.82
5	5	−1	0	0	0	71.87	11	13	0	0	0	0	73.60
5	5	−1	0	0	0	71.53	11	13	0	0	0	0	73.59
5	5	−1	0	0	0	72.08	11	13	0	0	0	0	73.34
5	5	−1	0	0	0	71.58	11	13	0	0	0	0	73.76
6	21	1	0	0	0	82.34	12	13	0	0	0	0	75.52
6	21	1	0	0	0	82.20	12	13	0	0	0	0	74.74
6	21	1	0	0	0	81.85	12	13	0	0	0	0	75.00
6	21	1	0	0	0	81.85	12	13	0	0	0	0	74.90

4.4 Wind tunnel experiment

Simpson et al. (2004) report the results from a wind tunnel experiment, involving four different responses: coefficient of lift at the front of the car (y_1), coefficient of lift at the rear of the car (y_2), drag (y_3) and lift over drag ratio (y_4). The design for the experiment, which is shown in Table 5 along with the response data, had nine whole plots of five runs. Four experimental variables were studied: front ride height (x_1), rear ride height (x_2), yaw angle (x_3) and grille coverage (x_4). The first two of these are whole-plot factors, whereas the others are sub-plot factors. The design involved 25 distinct factor level combinations or treatments, 20 of which were duplicated. A special feature of the design was that only one of the quadratic whole-plot effects and only one of the quadratic sub-plot effects could be estimated. Hence, we estimated a model including main effects, two-factor interaction effects and two of the four quadratic effects. The results of the lack-of-fit tests for the four responses are given in Table 6.

For two of the responses in the wind tunnel experiment, y_2 and y_4, there is significant lack of fit. For these responses, there are substantial differences between the pure-error estimates of the variance components and the estimates obtained from the response surface model. Note that, for the wind tunnel experiment, the denominator degrees of freedom are 16 for each of the responses. This is due to the orthogonality of the subplot design to the whole plots.

To remove the lack of fit for the y_2 and y_4 responses, more than just a few higher-order interactions have to be added to the model. For y_2, for instance, adding all three-factor interactions to the model does not suffice. Adding all three-factor interactions and all linear-by-quadratic interactions, however, does remove the lack of fit. The F statistic and the p-value of the corresponding lack-of-fit test equal 1.87 and 0.1654, respectively. For the y_4 response, one model that does not exhibit significant lack of fit is a model including all three-factor interactions and the four-factor interaction. The corresponding F statistic and p-value are 2.38 and 0.0719, respectively. In any case, it should be clear that, for two of the responses in the wind tunnel experiment, no simple response surface model exists that fits the data well. This may be due to the rounding used for the responses, the extremely small estimates for the variance components, and a few outlying observations. Especially for the y_2 response, the estimates for σ_1 and σ_0 are of the same order of magnitude as the rounding error. A transformation of the y_2 and y_4 responses did not result in simpler solutions to avoid lack of fit.

Table 5: Design and response data for the wind tunnel experiment

Block	Treatment	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
1	4	1	−1	−1	−1	−0.079	−0.219	0.416	0.715
1	9	1	−1	−1	1	−0.130	−0.227	0.409	0.875
1	14	1	−1	0	0	−0.097	−0.219	0.401	0.788
1	19	1	−1	1	−1	−0.069	−0.210	0.398	0.700
1	24	1	−1	1	1	−0.121	−0.199	0.385	0.830
2	2	−1	1	−1	−1	−0.120	−0.281	0.419	0.955
2	7	−1	1	−1	1	−0.168	−0.290	0.410	1.118
2	12	−1	1	0	0	−0.127	−0.276	0.400	1.005
2	17	−1	1	1	−1	−0.097	−0.238	0.393	0.852
2	22	−1	1	1	1	−0.151	−0.259	0.386	1.061
3	5	1	1	−1	−1	−0.112	−0.249	0.435	0.831
3	10	1	1	−1	1	−0.168	−0.259	0.428	0.996
3	15	1	1	0	0	−0.139	−0.252	0.421	0.926
3	20	1	1	1	−1	−0.105	−0.229	0.414	0.807
3	25	1	1	1	1	−0.157	−0.228	0.405	0.952
4	2	−1	1	−1	−1	−0.123	−0.279	0.420	0.958
4	7	−1	1	−1	1	−0.173	−0.289	0.412	1.123
4	12	−1	1	0	0	−0.138	−0.270	0.404	1.012
4	17	−1	1	1	−1	−0.104	−0.240	0.394	0.872
4	22	−1	1	1	1	−0.155	−0.261	0.387	1.074
5	4	1	−1	−1	−1	−0.081	−0.221	0.418	0.721
5	9	1	−1	−1	1	−0.128	−0.226	0.408	0.867
5	14	1	−1	0	0	−0.098	−0.219	0.400	0.793
5	19	1	−1	1	−1	−0.070	−0.212	0.399	0.708
5	24	1	−1	1	1	−0.118	−0.198	0.383	0.825
6	5	1	1	−1	−1	−0.118	−0.249	0.436	0.843
6	10	1	1	−1	1	−0.168	−0.255	0.426	0.994
6	15	1	1	0	0	−0.138	−0.246	0.419	0.918
6	20	1	1	1	−1	−0.107	−0.227	0.412	0.810
6	25	1	1	1	1	−0.160	−0.225	0.403	0.956
7	3	0	0	−1	−1	−0.111	−0.248	0.420	0.853
7	8	0	0	−1	1	−0.158	−0.252	0.409	1.004
7	13	0	0	0	0	−0.128	−0.238	0.401	0.912
7	18	0	0	1	−1	−0.093	−0.217	0.394	0.785
7	23	0	0	1	1	−0.149	−0.211	0.384	0.939
8	1	−1	−1	−1	−1	−0.096	−0.250	0.402	0.861
8	6	−1	−1	−1	1	−0.150	−0.257	0.394	1.033
8	11	−1	−1	0	0	−0.108	−0.231	0.382	0.887
8	16	−1	−1	1	−1	−0.082	−0.221	0.380	0.797
8	21	−1	−1	1	1	−0.133	−0.220	0.369	0.959
9	1	−1	−1	−1	−1	−0.097	−0.249	0.400	0.863
9	6	−1	−1	−1	1	−0.154	−0.257	0.391	1.051
9	11	−1	−1	0	0	−0.118	−0.239	0.383	0.932
9	16	−1	−1	1	−1	−0.091	−0.226	0.382	0.830
9	21	−1	−1	1	1	−0.132	−0.217	0.365	0.955

Table 6: Results for the lack-of-fit test and variance component estimates for the data from the wind tunnel experiment.

	Lack of fit			Pure error		Response surface	
Response	df	F value	p-value	σ_1^2	σ_0^2	σ_1^2	σ_0^2
y_1	16	1.87	0.1213	0.65×10^{-5}	0.57×10^{-5}	0.61×10^{-5}	0.78×10^{-5}
y_2	16	8.37	<.0001	0.70×10^{-6}	0.49×10^{-5}	0	0.19×10^{-4}
y_3	16	1.98	0.1001	0.51×10^{-6}	0.16×10^{-5}	0.38×10^{-6}	0.23×10^{-5}
y_4	16	3.60	0.0094	0.42×10^{-4}	0.72×10^{-4}	0.26×10^{-4}	0.15×10^{-3}

4.5 Split-split-plot example

The lack-of-fit test we propose can also be applied to data from multi-stratum experiments other than split-plot experiments. For lack of published data sets from industrial split-split-plot response surface experiments, we simulated data for a D-optimal 48-run split-split-plot design obtained using the algorithm of Jones and Goos (2009), as implemented in the JMP 10 software, and assuming a full quadratic model in four factors, x_1–x_4. The design involves eight whole plots, and each whole plot has two subplots of three runs. The first factor, x_1, is the whole-plot factor, and the second, x_2, is the subplot factor. The design, which is shown in Table 7 along with the response data, has 41 distinct factor level combinations or treatments. The treatments 2, 9, 10, 14, 28, 37 and 41 are duplicated. The seven duplicate treatments can be used to obtain pure-error estimates for the variance components corresponding to the whole plots, the subplots and the runs in the split-split-plot model

$$\mathbf{Y} = \mathbf{X}_t\boldsymbol{\tau} + \mathbf{Z}_2\boldsymbol{\delta} + \mathbf{Z}_1\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{7}$$

where $\mathbf{Y}$ is a random variable of which the response vector $\mathbf{y}$ is assumed to be a realization, $\mathbf{X}_t$ is the full treatment design matrix, $\boldsymbol{\tau}$ is the corresponding vector of treatment means, $\boldsymbol{\delta}$ is a vector of random whole-plot errors, $\mathbf{Z}_2$ is the design matrix for these random effects, $\boldsymbol{\gamma}$ is a vector of random whole-plot errors, $\mathbf{Z}_1$ is the design matrix for these random effects, and $\boldsymbol{\epsilon}$ is the vector of random experimental unit errors. We further assume that $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_2^2\mathbf{I})$, $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_1^2\mathbf{I})$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_0^2\mathbf{I})$, and that all random effects are independent. We call model (7) the full treatment model.

We simulated responses for the D-optimal design assuming the values -9, -1, 3 and -5 for the linear main effects, 7, -1, 6 and -8 for the quadratic effects, and 5, -7, -1, -4, -3, -8 for the two-factor interaction effects. We also assumed that two three-factor interactions (one involving x_1, x_2 and x_3 with coefficient -6.5, and one involving x_1, x_2 and x_4 with coefficient -5) were active, and that the variance components σ_2^2, σ_1^2

Table 7: Design and response data for the split-split-plot example

WP	SP	Treatment	x_1	x_2	x_3	x_4	y	WP	SP	Treatment	x_1	x_2	x_3	x_4	y
1	1	9	−1	1	−1	−1	62.36	5	9	9	−1	1	−1	−1	20.03
1	1	12	−1	1	1	−1	99.70	5	9	10	−1	1	−1	1	17.59
1	1	14	−1	1	1	1	86.91	5	9	13	−1	1	1	0	56.40
1	2	6	−1	0	−1	0	83.28	5	10	1	−1	−1	−1	−1	34.54
1	2	7	−1	0	0	1	77.68	5	10	2	−1	−1	−1	1	44.23
1	2	8	−1	0	1	−1	97.63	5	10	4	−1	−1	1	0	56.33
2	3	27	1	−1	−1	−1	51.44	6	11	35	1	1	−1	−1	33.80
2	3	28	1	−1	−1	1	65.75	6	11	37	1	1	−1	1	33.82
2	3	30	1	−1	1	−1	85.28	6	11	40	1	1	1	0	18.50
2	4	37	1	1	−1	1	83.41	6	12	32	1	0	−1	−1	18.30
2	4	38	1	1	0	−1	78.43	6	12	33	1	0	0	1	3.45
2	4	41	1	1	1	1	32.19	6	12	34	1	0	1	0	23.55
3	5	2	−1	−1	−1	1	93.18	7	13	28	1	−1	−1	1	18.88
3	5	3	−1	−1	0	−1	85.02	7	13	29	1	−1	0	−1	11.14
3	5	5	−1	−1	1	1	94.08	7	13	31	1	−1	1	1	−3.51
3	6	10	−1	1	−1	1	60.83	7	14	36	1	1	−1	0	40.33
3	6	11	−1	1	0	−1	75.64	7	14	39	1	1	1	−1	35.54
3	6	14	−1	1	1	1	86.61	7	14	41	1	1	1	1	−15.48
4	7	24	0	1	−1	0	80.57	8	15	16	0	−1	−1	0	19.65
4	7	25	0	1	0	1	61.15	8	15	18	0	−1	0	1	7.58
4	7	26	0	1	1	−1	92.65	8	15	19	0	−1	1	−1	32.10
4	8	15	0	−1	−1	−1	64.96	8	16	21	0	0	−1	−1	10.54
4	8	17	0	−1	0	0	76.17	8	16	22	0	0	0	0	19.65
4	8	20	0	−1	1	1	68.28	8	16	23	0	0	1	−1	35.02

and σ_0^2 were 9, 4 and 1, respectively. When estimating a second-order response surface model, the lack-of-fit test has 26 numerator degrees of freedom (41 treatments minus 15 parameters in the response surface model) and one denominator degree of freedom. The F test statistic is 260.36, which results in a p-value of 0.0489. Thus, despite the single denominator degree of freedom, there is an indication of some lack of fit. This is in line with the model assumed to simulate the data. The pure-error estimates of σ_2^2, σ_1^2 and σ_0^2 are 2.7925, 4.1887 and 0.2314, respectively. Each of these are completely different from the estimates obtained from the second-order response surface model: 0, 0 and 57.1494. As in the pastry dough experiment, the galvanized steel experiment and the wind tunnel experiment, a symptom of the lack of fit is the large estimate for σ_0^2 obtained from the response surface model, relative to the pure-error estimate obtained from the full treatment model. Adding the two active three-factor interactions to the model leads to an F statistic of 5.30 and a p-value of 0.3321 for the lack-of-fit test. Adding one of the two active three-factor interactions leads to p-values of 0.0749 and 0.0610, each of which suggests that adding one three-factor interaction effect to the model is not enough.

5 Discussion

Testing for lack of fit is a routine part of response surface methodology, when the runs can be completely randomized. It should similarly be done routinely in blocked, split-plot and other multi-stratum response surface experiments. We have shown that this testing is fairly straightforward, given an appropriate linear mixed models program. We recommend that this test should always be done before interpreting a fitted polynomial response surface model. If no evidence of lack of fit is found, then we can interpret the response surface model output with more confidence. If lack of fit is found then further investigation is required. Sometimes a single third-order term can explain the lack of fit, in which case we can modify our model accordingly; at other times, the lack of fit might be caused by an outlier, or indicate the need for a transformation of the response. In other cases the lack of fit, though statistically significant, might have little impact on the interpretation of the data and can be effectively ignored. The most difficult cases are those like the wind tunnel data, where it is very difficult to see a clear pattern indicated by the lack of fit. In such cases, we should procede to interpetation with caution.

At present, construction methods for efficient split-plot and other multi-stratum designs that allow for pure-error estimation and lack-of-fit testing are still lacking. An interesting avenue for research would be to extend the approach of Gilmour and Trinca (2012) for completely randomized designs and designs with fixed block effects to split-plot and other multi-stratum designs, and to experiments involving random block effects.

Appendix. SAS Programs

Galvanized steel example

```
data steel;
input block treat x1 x2 y;
datalines;
1        1      -1    -1    1226
1        2       0    -1    1898
...
12       8       0     1    2390
12       9       2     1    2572
;
* response surface model;
proc mixed;
```

```
class block;
model y = x1 x2 x1*x2 x1*x1 x2*x2 / ddfm=kr solution;
run;
* lack-of-fit test second-order model;
proc mixed;
class block treat;
model y = x1 x2 x1*x2 x1*x1 x2*x2 treat/ ddfm=kr solution;
random block / solution;
run;
* lack-of-fit test second-order model + linear-by-quadratic interactions;
proc mixed;
class block treat;
model y = x1 x2 x1*x2 x1*x1 x2*x2 x1*x2*x2 treat/ ddfm=kr solution;
random block / solution;
run;
```

Split-split-plot example

```
data splitsplitplot;
input run wp sp x1-x4 y treat;
datalines;
1     1       9     -1     1    -1    -1     62.36
1     1      12     -1     1     1    -1     99.70
1     1      14     -1     1     1     1     86.91
1     2       6     -1     0    -1     0     83.28
...
8    15      19      0    -1     1    -1     82.10
8    16      21      0     0    -1    -1     60.54
8    16      22      0     0     0     0     69.65
8    16      23      0     0     1    -1     85.02
;
* response surface model;
proc mixed data = splitsplitplot;
class wp sp;
model y = x1|x2|x3|x4@2 x1*x1 x2*x2 x3*x3 x4*x4 / solution ddfm = kr;
random wp sp;
run;
* lack-of-fit test second-order model;
proc mixed data = splitsplitplot;
class wp sp treat;
```

```
model y = x1|x2|x3|x4@2 x1*x1 x2*x2 x3*x3 x4*x4 treat/ ddfm = kr solution;
random wp sp;
run;
* lack-of-fit test second-order model + 2 three-factor interactions;
proc mixed data = splitsplitplot;
class wp sp treat;
model y = x1|x2|x3|x4@2 x1*x1 x2*x2 x3*x3 x4*x4
          x1*x2*x3 x1*x3*x4 treat/ ddfm = kr solution;
random wp sp;
run;
```

References

Almimi, A. A., M. Kulahci, and D. C. Montgomery (2009). Checking the adequacy of fit of models from split-plot designs. *Journal of Quality Technology 41*, 272–284.

Box, G. E. P. and N. R. Draper (2007). *Response Surfaces, Mixtures and Ridge Analyses* (2nd ed.). New York: Wiley.

Commission internationale de l'éclairage (1986). *Colorimetry, CIE publication 15.2.* (2nd ed.). Vienna: Bureau Central CIE.

Gilmour, S. G. and T. J. Ringrose (1999). Controlling processes in food technology by simplifying the canonical form of fitted response surfaces. *Applied Statistics 48*, 91–101.

Gilmour, S. G. and L. A. Trinca (2000). Some practical advice on polynomial regression analysis from blocked response surface designs. *Communications in Statistics: Theory and Methods 29*, 2157–2180.

Gilmour, S. G. and L. A. Trinca (2006). Response surface experiments on processes with high variation. In A. I. Khuri (Ed.), *Response Surface Methodology and Related Topics*, pp. 19–46. Singapore: World Scientific.

Gilmour, S. G. and L. A. Trinca (2012). Optimum design of experiments for statistical inference (with discussion). *Applied Statistics 61*, 345–401.

Jones, B. and P. Goos (2009). D-optimal design of split-split-plot experiments. *Biometrika 96*, 67–82.

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

Khuri, A. I. (1992). Response surface models with random block effects. *Technometrics 34*, 26–37.

McCulloch, C. E., S. R. Searle, and J. M. Neuhaus (2008). *Generalized, Linear, and Mixed Models* (2nd ed.). New York: Wiley.

Nelder, J. A. (1965). The analysis of randomized experiments with orthogonal block structure. i. block structure and the null analysis of variance. *Proceedings of the Royal Society of London, Series A 283*, 147–162.

Parker, P., S. M. Kowalski, and G. G. Vining (2007). Unbalanced and minimal point equivalent estimation second-order split-plot designs. *Journal of Quality Technology 39*, 376–388.

Simpson, J. R., S. M. Kowalski, and D. Landman (2004). Experimentation with randomization restrictions: Targeting practical implementation. *Quality and Reliability Engineering International 20*, 481–495.

Trinca, L. A. and S. G. Gilmour (2000). An algorithm for arranging response surface designs in small blocks. *Computational Statistics and Data Analysis 33*, 25–43.

Vining, G. G. and S. M. Kowalski (2008). Exact inference for response surface designs within a split-plot structure. *Journal of Quality Technology 40*, 394–406.

Vining, G. G., S. M. Kowalski, and D. C. Montgomery (2005). Response surface designs within a split-plot structure. *Journal of Quality Technology 37*, 115–128.