Accepted: 21 December 2017

DOI: 10.1111/jac.12267

REVIEW ARTICLE

WILEY Journal of Agronomy and Crop Science

A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels

H. P. Piepho[1] | R. N. Edmondson[2]

[1]Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany

[2]Rana House, Wellesbourne, UK

Correspondence
H. P. Piepho, Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany.
Email: piepho@uni-hohenheim.de

Abstract

Agronomic experiments are often complex and difficult to interpret, and the proper use of appropriate statistical methodology is essential for an efficient and reliable analysis. In this paper, the basics of the statistical analysis of designed experiments are discussed using real examples from agricultural field trials. Factorial designs allow for the study of two or more treatment factors in the same experiment, and here we discuss the analysis of factorial designs for both qualitative and quantitative level treatment factors. Where treatment factors have quantitative levels, models of treatment effects are essential for efficient analysis and in this paper we discuss the use of polynomials for empirical quantitative modelling of treatment effects. The example analyses cover experiments with a single quantitative level factor, experiments with mixtures of quantitative and qualitative level factors, polynomial regression designs with two quantitative level factors, split-plot designs with quantitative level factors and repeated-measures designs with correlated data and a quantitative treatment response over time. Modern mixed model computer software for routine analysis of experimental data is now readily available, and we demonstrate the use of two alternative software packages, the SAS package and the R language. The main purpose of the paper is to exemplify standard statistical methodology for routine analysis of designed experiments in agricultural research, but in our discussion we also provide some references for the study of more advanced methodology.

KEYWORDS

factorial analysis, linear mixed models, polynomial regression, R, repeated-measures analysis, response surface models, SAS, split-plot analysis

1 | INTRODUCTION

Agronomic experiments are often difficult and complex both to plan and to execute and may depend on many factors that can affect both efficiency and reliability. They are usually expensive and may sometimes take many years to accomplish. In addition, agronomic experiments are typically subject to high background variability due to effects such as fertility trends in fields or initial weight differences between animals or year-to-year variability in weather as well as to the high natural variability of biological and agronomic systems. For these and other reasons, the proper statistical design and analysis and the effective control of background variability by blocking or by the use of covariates together with sufficient replication of treatments are vital for the precision and accuracy of agronomic experiments.

This paper exemplifies the use of modern statistical methods for the analysis of designed experiments assuming various block and treatment structures including designs with repeated measurements on the same experimental units. A good statistical analysis addresses

the aims of the experiment and produces accurate and efficient information that provides a simple and parsimonious description of the results while still taking proper account of the error structure of the design. A properly chosen analysis will provide not only a full summary of the observed data but will also give additional insight into the effects of the treatments on the system under study.

The focus of this paper is on designed experiments with treatment structures that can be modelled by polynomial regression methods. All the experiments considered are comparative, meaning that the focus is exclusively on comparisons between treatments in the same experiment (Bailey, 2008; Giesbrecht & Gumpertz, 2004; Hinkelmann & Kempthorne, 1994). Polynomial regression is a standard linear model methodology that fits naturally within the standard analysis of variance. We will show how modern mixed model computer software can be used for the routine analysis of blocked experiments based on a range of block and treatment structures, including designs with repeated observations on the same experimental units.

The paper does not explicitly discuss the design of experiments, but it will be assumed that the example data sets have been collected from efficient and well-designed experiments. Good references on the design of agronomic experiments include Mead and Curnow (1983), John and Williams (1995), Dean and Voss (1999), Kuehl (2000), Quinn and Keough (2002), Bailey, 2008; Mead, Gilmour, and Mead (2012) and Welham, Gezan, Clark, and Mead (2015).

The paper is organized into sections, and in the first section, we provide some theoretical background on important concepts for the analysis of designed experiments based on linear statistical models. In the subsequent sections, we consider five worked examples illustrating various aspects and general principles for the analysis of factorial experiments. Example 1 shows factorial interactions; Example 2 demonstrates lack-of-fit testing for regression models; Example 1 re-visited illustrates how regression curves for one factor can be fitted at different levels of a second factor; Example 3 introduces polynomial modelling for two or more quantitative level factors; Example 4 uses regression methods for the analysis of repeated-measures data over time; and Example 5 deals with the transformation of variables in polynomial regression.

In presenting these examples, we aim to show that the proper use of statistical methodology can both enhance and simplify the interpretation of complex agronomic experiments. We use SAS (SAS Institute Inc., 1999) and the R environment (R Core Team, 2016) throughout, and we provide sample programme code for all the examples discussed. The R code and the data for the examples have been implemented in the R package 'agriTutorial' (Edmondson & Piepho, 2018). The methodology is standard and widely available, and any other modern reputable statistical software package should provide access to the methodology discussed in this paper. While the methods discussed are often adequate for a full and reliable analysis of a designed experiment, there are numerous generalizations and extensions that can sometimes provide additional insight into an analysis. In the discussion, we give a brief mention of some further regression methodologies for designed experiments.

2 | STATISTICAL MODELS—SOME BACKGROUND

2.1 | Treatment structures

The design and analysis of experiments depends critically on assumptions about the measurement structures of the underlying treatment factors, and an understanding of these structures is essential for a correct analysis of experimental data. We briefly outline below three of the major measurement categories which frequently occur in agricultural experiments.

2.1.1 | Qualitative level factors

Treatment factors are qualitative (or nominal) level factors if they have no particular ordering of factor levels. For example, a variety factor in a variety trial is a qualitative level factor because the factor levels are varieties which usually have no particular ordering on a quantitative scale. The most basic analysis for a qualitative level factor is by tabulation of the summary means of the individual factor levels. There is no special relationship between the individual factor levels, so there is no special ordering of the levels and no quantitative relationship between the factor levels. Assuming the experiment is adequately replicated, the variances of the estimated mean differences can be calculated and the factor levels compared either descriptively by a pairwise standard error of a difference (*SED*) or analytically by a suitable multiple comparisons test (Bretz, Hothorn, & Westfall, 2011; Hsu, 1996).

2.1.2 | Quantitative level factors

Treatment factors are quantitative (or metric) level factors if they have a strict ordering of the factor levels with measurable quantitative differences between the levels. For example, a factor with levels equal to different amounts or proportions or rates of the same fertilizer would be a quantitative level factor. Regression models use the distance information on the scale of the quantitative predictor variables and are essential for the capture of quantitative information on the levels of a quantitative level factor. By comparison, a simple qualitative factorial analysis would ignore any information on the distance and order relationship between factor levels.

A common model for a quantitative level factor with not too many levels (usually not more than say 7 or 8) is a low-order polynomial regression model fitted to the quantitatively spaced treatment levels. The fitted regression model should explain the observed data as parsimoniously as possible and where a polynomial regression analysis is used should normally include only polynomial terms up to the highest degree term that is statistically significant in the fitted model. A meaningful polynomial requires at least three quantitative factor levels; therefore, a treatment factor with only two levels is normally regarded as qualitative irrespective of the actual treatment levels.

2.1.3 | Ordinal level factors

Treatment factors are ordinal level factors if they have a strict ordering of the factor levels but do not have measurable quantitative differences between the factor levels. For example, susceptibility to lodging of cereals can be scored on an arbitrary integer scale ranging from 1 for no lodging to N for complete lodging with intermediate scores representing intermediate levels of lodging. Score data are usually regarded as ordinal because although the scores can be ranked in order of importance, there is usually no underlying quantitative scale that measures the separation of the successive ordinal levels. A full analysis for an ordinal levels factor can be complex, but often it can be reasonable to assume a suitable spacing (usually equal spacing) for the ordinal factor levels (Agresti, 2010) and then a polynomial model can be fitted for that assumed spacing. More general methods are available and are discussed by a number of authors including Agresti (2010) and Gertheiss (2014) but will not be discussed further here.

2.2 | Block and plot structures

Plots are the experimental units of a field experiment, and by association, the term is often used to denote the experimental units of other kinds of experiments. Plots are normally grouped into homogeneous blocks as a good block design will help ensure improved precision of comparison between treatments. Where blocks contain less than a complete set of treatments, they are called incomplete blocks and may be nested within other, larger, blocks. Treatments must be randomized within each set of blocks subject to the constraint that the design allocation of treatments to blocks remains unchanged. Randomization helps justify the usual assumptions of an analysis of variance and helps ensure that the treatment comparisons are unbiased after fitting an appropriate block and treatment model (Bailey, 2008). The plots and the blocks of a design both need to be properly modelled by plot and block effects appropriate to the design of the experiment.

2.2.1 | Block effects

Treatment factors are of direct interest and are usually modelled as fixed effects, but nuisance factors such as block effects can be modelled either as fixed or as random effects. Fixed block effects models are robust against distributional assumptions and are useful for eliminating block differences in designs with large complete or near-complete blocks (Dixon, 2017). In general, however, incomplete block designs will confound substantial treatment information between blocks and then an analysis with recovery of inter-block information will be necessary. Fixed blocks are not useful for recovery of inter-block treatment information, but an assumption of random blocks will allow treatment information to be recovered by combining inter- and intra-block information using the reciprocals of the estimates of inter- and intra-block variances as weights (Mead et al., 2012; Piepho, Williams, & Ogutu, 2013). If a random blocks model is fitted using a mixed model package, recovery of information is provided for automatically without the need for any further data manipulation. This is of particular importance for designs with relatively small incomplete blocks where substantial information may be contained in the inter-block analysis. Modern mixed model software has greatly facilitated the analysis of designs that include random effects and has made the analysis of designs with mixtures of random effects and fixed effects both feasible and practicable.

2.2.2 | Plot effects

Standard analysis of simple randomized block designs assumes plots with a simple uncorrelated constant variance error structure, but more complex designs can have more complex error structures. One important class of design is the split-plot design where different treatment factors can have differently sized plots and hence different error structures. For example, an experiment with several different varieties of a crop and several different fertilizer treatments could use main plots for the fertilizer treatments and sub-plots within main plots for the individual varieties. Another important class of design is the repeated-measures design where the same experimental plot or unit is measured repeatedly over time. For example, a growth curve experiment on animal diets might involve repeated measurements over time on each individual animal, which would probably mean that repeated measurements on the same animal would be correlated with measurements close together in time on the same animal more highly correlated than measurements further apart in time. Many other error structures are possible, but split-plot designs and repeated-measures designs are of special importance in practical research and the analysis of these two important classes of design will be exemplified in the following sections.

3 | EXAMPLE 1—FACTORIAL INTERACTIONS

Fully crossed factorial designs examine all combinations of the levels of a set of factors and can generate tables of factorial treatment means that are sometimes complex and difficult to interpret. Factorial analysis simplifies the interpretation of factorial designs by finding smaller, marginal, tables that give a simplified summary of the factor effects. A marginal table contains a subset of the factorial treatments averaged across all other factors in the design. For example, in a factorial design with two factors A and B there is a full table of factorial treatment means for A $\times$ B and a table of marginal A-means averaged across the levels of B and a table of marginal B-means averaged across the levels of A. If a factorial analysis of variance shows no evidence of interaction between factors A and B, then the interpretation of the factorial treatment effects can be based on the separate main effects (marginal means) tables for A and B without loss of treatment information. If, on the other hand, there is a significant interaction between factors A and B, then the factorial table for A $\times$ B contains non-additive factorial treatment

information and cannot be marginalized without a significant loss of treatment information.

Factorial analysis generalizes in a straightforward way for any number of factors, but, for simplicity, our paper will focus mainly on just two factors. The following example will illustrate some of the key features of a factorial analysis of variance.

Example 1: Gomez and Gomez (1984, p. 143) report an experiment with three management practices (minimum, optimum and intensive), five different amounts of nitrogen (N) fertilizer (0, 50, 80, 110 and 140 kg/ha) and three varieties (V1, V2 and V3). The experiment involved variety and management as qualitative treatment factors and nitrogen fertilizer as a quantitative treatment factor. Overall, there were 45 treatment combinations (3 management practices × 3 varieties × 5 nitrogen levels) and each treatment combination was replicated three times.

The experiment was laid out as a split-split-plot design, with fertilizer as the main-plot factor with the five rates randomly assigned to five main plots in each of three complete replicate blocks, management practice as the sub-plot (or split-plot) factor with the three management practices randomly assigned to three sub-plots within each main plot and variety as the sub-sub-plot (or split-split-plot) factor with the three varieties randomly assigned to individual sub-sub-plots within each sub-plot. A classical analysis of variance for this example is given in Table 1 where all treatment factors are regarded as qualitative. This analysis displays sources of variation corresponding to the different types of experimental unit to which the treatment factor levels were randomly applied, i.e., main plots (for nitrogen), split plots (for management) and split-split plots (for variety). Each different type of experimental unit (or "stratum") requires a separate error term in the model.

Treatment means for the variety-by-fertilizer combinations, averaged over management practices, are reported in Table 2. In this example, averaging across management practices is justified because there is no statistical evidence of any interaction of management practices with variety or nitrogen (Table 1). The table shows the individual nitrogen-by-variety treatment means and also reports marginal means for varieties averaged over the levels of nitrogen and marginal means for nitrogen averaged over the levels of varieties. Comparison between the marginal means of these two factors is meaningful, however, only if the effects of these two factors are independent of each other, and, in Table 2, this would imply that (i) the expected differences between fertilizers are the same for each variety and are equal to the expected differences between the marginal fertilizer means and (ii) the expected differences between varieties are the same for each fertilizer and are equal to the expected differences between the variety marginal means.

In fact, the analysis of variance shows a significant two-factor interaction between the variety and the fertilizer effects (Table 1), indicating non-additivity between the effects of these two factors. For example, the marginal means of varieties V1 and V2 have a difference of 1.27 t/ha, whereas at a nitrogen level of 140 kg/ha, the difference of the two varieties is 2.21 t/ha. When there is a significant interaction between factors, those factors are no longer independent and their effects cannot be explained by using a simple marginal model. Instead, the factor combinations must be assessed individually or must be fitted by a more general model.

The analysis of variance (ANOVA) of a factorial experiment is often followed by a pairwise comparison of treatment means. However, where treatment factors have quantitative factor levels, such as varying amounts of the same type of fertilizer, a simple comparison between treatment means will not take proper account of the treatment structure of the design. In Table 2, for example, it is difficult to see the variety-specific pattern of yield response to increased amounts of nitrogen simply from an inspection of the treatment means. In this situation, the set of pairwise comparisons between treatment means is not fully informative and instead a regression model of the treatment response against the actual factor levels will usually give a more informative analysis.

Figure 1 shows a regression model of yield against amount of applied nitrogen fertilizer for the three varieties. The interaction

TABLE 1 Classical analysis of variance for rice experiment with three varieties of rice, five fertilizer treatments and three management practices (Example 1). All factors are treated as qualitative in this analysis, including nitrogen

Source of variation	Degrees of freedom	Sum of squares	Mean square	*F*-value	*p*-value
Blocks (replicates)	2	0.73	0.366		
Nitrogen	4	61.64	15.410	27.70	<.0001
Main-plot error	8	4.45	0.556		
Management	2	42.94	21.468	81.996	<.0001
Management × nitrogen	8	1.10	0.138	0.527	.8226
Split-plot error	20	5.24	0.262		
Variety	2	206.01	103.007	207.867	<.0001
Variety × management	4	3.85	0.963	1.943	.1149
Variety × nitrogen	8	14.14	1.768	3.568	.0019
Variety × management × nitrogen	16	3.70	0.231	0.467	.9538
Split-split-plot error	60	29.73	0.496		

TABLE 2 Variety-by-fertilizer means (yield; t/ha) and standard errors of a difference (*SED*) for rice experiment with three varieties of rice and five fertilizer treatments (Example 1)

	Amount of nitrogen (kg N/ha)					
Variety	0	50	80	110	140	Average[a]
V1	4.51	4.76	5.83	5.44	5.08	5.12
V2	5.16	6.02	6.59	6.92	7.29	6.40
V3	6.48	7.88	8.56	8.44	9.34	8.14
Average[b]	5.38	6.22	6.99	6.93	7.24	

SED for variety-by-fertilizer means at the same level of fertilizer = 0.300; *SED* for variety-by-fertilizer means with the same variety = 0.318; Both *SED*s were obtained using residual maximum likelihood (REML) (see the continuation of Example 1).

[a]Marginal means for fertilizers.

[b]Marginal means for varieties.

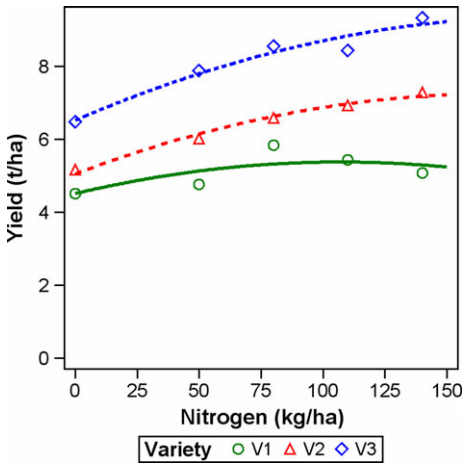

FIGURE 1 Quadratic models for relation between yield and amount of nitrogen (N) for three rice varieties (Example 1). Plotted points are observed variety-by-fertilizer means. The models have a common quadratic term for all varieties but separate linear terms (see Table 6 and further explanation in the continuation of Example 1) [Colour figure can be viewed at wileyonlinelibrary.com]

pattern which emerges from this analysis indicates that variety V1 showed little or no response to increased amounts of nitrogen, whereas the other two varieties showed a positive response. This indicates that the effect of fertilizer was dependent on the variety, which accounts for the observed interaction between these two treatment factors. A proper statistical analysis with suitable tests of fit guided the choice of the fitted polynomial model trend curves, and this analysis will be fully explained in the continuation of Example 1 later in this paper.

In this example, treatment means and marginal means are equal to arithmetic means, but it should be emphasized that more generally model-based mean estimates (adjusted means or least-square means) are preferable. With balanced data as in the present example, arithmetic means and model-based means coincide, but with unbalanced data they will be different. Model-based means along with their appropriate standard errors and confidence limits are produced automatically when using a mixed model package, such as the emmeans package in R (Lenth, 2018) or the lsmeans statement in SAS.

4 | MIXED MODEL ANALYSIS OF EXAMPLE 1 BY REML

A classical analysis of variance is based on sums of squares as shown in Table 1, but a more modern approach is to fit a linear mixed model which estimates variance components by the residual maximum likelihood (REML) method (Kenward & Roger, 1997). Using this approach, the sub-sub-plot (split-split-plot) error variance estimate for Example 1 equals 0.4039, and the main-plot error variance estimate is 0.0169, whereas the sub-plot (split-plot) error variance estimate is zero, meaning that this effect is effectively removed from the model. Generally with mixed models, determination of the denominator degrees of freedom for Wald-type *F*- and *t*-statistics becomes an issue, and we here use the method proposed by Kenward and Roger (1997, 2009). Also, REML generally does not provide a simple sums of squares analysis which means that classical *F*-statistics based on ratios of sums of squares are unavailable. Instead, REML packages usually provide Wald-type *F*-statistics which are equivalent to the classical *F*-statistics for simple classical designs but not otherwise, e.g., when there are missing values or the covariance structure is complex. For details the

TABLE 3 Sequential Wald-type *F* tests for factorial model fitted to the field experiment with rice (Example 1). All factors are treated as qualitative in this analysis, including nitrogen

Source	Numerator degrees of freedom	Denominator degrees of freedom[a]	*F*-value[a]	*p*-value
Blocks	2	8	0.66	.5439
Nitrogen	4	8	27.70	<.0001
Management	2	80	49.11	<.0001
Management × nitrogen	8	80	0.32	.9581
Variety	2	80	235.65	<.0001
Variety × management	4	80	2.20	.0760
Variety × nitrogen	8	80	4.04	.0004
Variety × management × nitrogen	16	80	0.53	.9241

[a]Obtained using the Kenward and Roger (1997) method. Sub-plot error variance component estimated to be zero.

reader is referred to pertinent textbooks such as Hocking (1985), Schabenberger and Pierce (2002) and West, Welch, and Gałecki (2014).

Table 3 shows Wald-type *F* tests for Example 1, and comparison with Table 1 shows that the *F* tests for variety and management effects are not identical to those in Table 1 because of the zero estimate for the sub-plot error variance. Effectively, the mean squares for sub-plot and sub-sub-plot error in Table 1 are pooled as are corresponding sets of error degrees of freedom, thus changing the error term for the treatment effects involving variety and management. Some REML packages have an option to lift the non-negativity constraint on the variance component estimates (e.g., the NOBOUND option in SAS). If this is used and the data are balanced, the REML-based *F* test and traditional ANOVA coincide (Hocking, 1985; Chapters 9 and 10; Piepho & Spilke, 1999), and this is also found for this example, with the sub-plot component of error variance estimate equalling −0.07791. Using this option may be seen as a conservative approach because the sub-plot error term is forced to remain in the model. With balanced data, the resulting *F* tests are guaranteed to be exact, whereas with the constraint in place the tests are approximate. Also, the use of negative variance components is in line with a randomization-based paradigm to linear modelling for designed experiments, as opposed to a model-based paradigm (Nelder, 1954). We should caution the reader, however, that with unbalanced data (e.g., due to missing values) negative variance components are bound to cause convergence problems when a variance estimate becomes negative. Negative variance components should therefore be allowed only with balanced data (Spilke, Hu, & Piepho, 2005). The probability of a negative variance estimate due to chance variation alone can be quite large in small data sets (Verdooren, 1982).

In field experiments, it is often reasonable to assume that yield or productivity in neighbouring plots will be positively correlated and that the correlation between neighbours will decrease with increasing distance. Assuming this kind of correlation structure, nested plot error variances will normally decrease with successive levels of nesting due to the nested plots being located, on average, closer together. Sometimes, however, nested variances may increase either due to random chance or due to some other cause and then a conventional nested analysis may give rise to negative variance component estimates. By definition, a variance must be positive. Variance structures with negative estimates for a variance component are perhaps best interpreted as implying negative correlations among adjacent experimental units, which could be due to competition effects. Negative variance component estimates are not expected to occur in well-designed field experiments with sufficient guard space between plots to avoid competition, so there is a strong argument for setting negative variance estimates equal to zero (Nelder, 1954) or imposing a non-negativity constraint as in the default setting with most REML packages. Some packages do not even permit lifting the non-negativity constraint, e.g., the R packages "nlme" and "lme4." In this paper, we focus entirely on analyses that impose a non-negativity constraint.

Table 1 shows that the ratio of the sub-sub-plot error mean square to the sub-plot error mean square for the example analysis was 0.496/0.262 = 1.893 with 60 and 20 degrees of freedom. Assuming an *F*-distribution for the variance ratios, such a value or larger would occur by chance with a probability of $p = .057$. While we accept that this variance ratio is quite large, we prefer, for the purposes of our illustrative example, to fit an analysis with negative variances set to zero as we believe this will be the normal analysis for this type of design. However, we recognize that some further analysis of this example could be valuable and although we do not have space to explore the example further here, we have provided some additional analysis in the Supporting information. We also provide some additional analysis of the assumed blocks-by-treatments model for this example in Section 3 of Example 1 in the agriTutorial package.

5 | EXAMPLE 2—LACK OF FIT AND MARGINALITY FOR A SINGLE QUANTITATIVE TREATMENT FACTOR

Designed experiments with replication allow for tests of the lack of fit of a candidate regression model (Dean & Voss, 1999). To illustrate this important step in the regression analysis of a replicated experiment, we now consider an experiment with a single quantitative level treatment factor.

Example 2: Petersen (1994, p. 125) describes an experiment conducted to assess the effects of five different quantities of N-fertilizer (0, 35, 70, 105 and 140 kg N/ha) on root dry matter yield of sugar beets (t/ha) with three complete replications laid out in three randomized complete blocks. One objective of this experiment was to determine the amount of fertilizer maximizing yield. The data are plotted in Figure 2, together with a straight line relationship fitted through the data points. Inspection of the scatter plot is the most important initial step of regression analysis as it helps in detecting any anomalies in the data and in identifying a suitable regression model. The scatter of the replicated observations at each treatment level shows some evidence that the variance increased with the mean although with only three

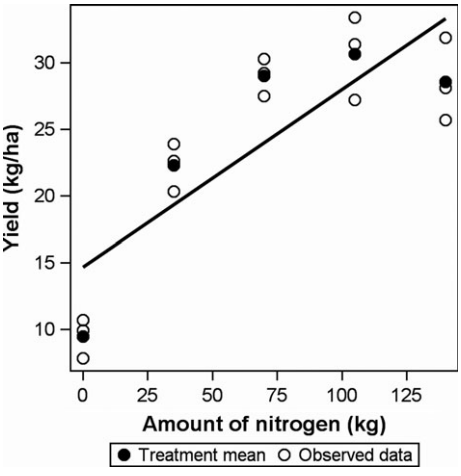

FIGURE 2 Fitted linear regression for yield of sugar beet (t/ha) against five amounts of N-fertilizer (0, 35, 70, 105 and 140 kg N/ha) (Example 2)

observations per treatment the evidence is rather weak. It is clear, however, from a visual inspection of the observed data (Figure 2) that the straight line relationship is not a good fit to the data and in the remainder of this section, we will discuss the problem of fitting a more general polynomial to the observed data.

The appropriate degree of a polynomial can be selected by sequentially adding significant polynomial terms of increasing degree starting with the linear term (Nelder, 2000) until the lack of fit becomes non-significant. Polynomials higher in degree than cubic (third degree) are not normally used, however, because a polynomial of high degree usually indicates either an over-fitted or a miss-specified model. Also note that the maximum possible degree of the polynomial fitted for a factor is given by the total degrees of freedom for that factor, which is one less than the number of factor levels (Dean & Voss, 1999).

The model for the linear relationship in this example is,

$$y_{ij} = \mu + b_j + \beta_1 x_i + e_{ij}, \quad (1)$$

where y_{ij} is the yield of the j-th replicate of the i-th fertilizer treatment, μ is a general intercept, b_j is the effect of the j-th block, β_1 is the linear slope, x_i is the amount of nitrogen applied with the i-th fertilizer treatment, and e_{ij} is a residual error term assumed to follow a normal distribution with zero mean and variance σ^2.

If a linear model is not valid for the observed data, there will be a bias, or lack of fit, in the prediction for each treatment level. This lack of fit corresponds to the vertical displacement of the observed treatment means from the fitted line at the observed nitrogen levels in Figure 2, and we can denote this lack-of-fit as δ_i for the i-th treatment. Adding this lack of fit as an effect, an invalid regression model can be adjusted locally at each observed treatment level x_i. In the case of a linear regression, this locally adjusted model becomes,

$$y_{ij} = \mu + b_j + \beta_1 x_i + \delta_i + e_{ij}. \quad (2)$$

.

The validity of linear model (1) can be assessed by testing the lack-of-fit effect δ_i for significance. Under the null hypothesis H_0: $\delta_i = 0$ for all i, the lack of fit vanishes and linear regression model (1) holds. Conversely, if the lack of fit is significant, it may be concluded that candidate model (1) does not fit. The analysis of variance for model (2), based on sequential sums of squares (Type I SS in SAS), is shown in Table 4. Note that the degrees of freedom for the polynomial term and the lack-of-fit effect add up to the total fertilizer degrees of freedom, which equal the number of fertilizer treatments minus one. The lack of fit is significant (p = .0002), confirming the visual impression from Figure 2, so we add the quadratic term $\beta_2 x_i^2$:

$$y_{ij} = \mu + b_j + \beta_1 x_i + \beta_2 x_i^2 + \delta_i + e_{ij}. \quad (3)$$

The lack of fit for the quadratic model is non-significant (p = .8039; Table 5), showing that the quadratic model fits well and can be reported as a summary. This corresponds well with the very small vertical displacement of treatment means from the fitted quadratic regression model in Figure 3. Again, the degrees of freedom for the polynomial terms and lack of fit add up to the treatment degrees of freedom. Dropping the non-significant lack-of-fit term, the final model is,

$$y_{ij} = \mu + b_j + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}. \quad (4)$$

This model implies a block-specific intercept $\mu + b_j$. Thus, there is a regression curve for each block, but the curves run parallel, meaning there is no interaction between blocks and treatments. This example illustrates the general assumption underlying all analyses of

TABLE 4 Sequential analysis-of-variance table for model (2) fitted to the sugar beet data (Example 2)

Source	Degrees of freedom	Sum of squares	*F*-value	*p*-value
Blocks (b_j)	2	26.32	3.71	.0724
Fertilizer treatments (τ_i)	4	913.56		
Linear N ($\beta_1 x_i$)	1	651.47	183.74	<.0001
Lack of fit (δ_i)	3	262.09	24.64	.0002
Error	8	28.37		

TABLE 5 Sequential analysis-of-variance table for model (3) fitted to the sugar beet data (Example 2)

Source	Degrees of freedom	Sum of squares	*F*-value	*p*-value
Blocks (b_j)	2	26.32	3.71	.0724
Fertilizer treatments (τ_i)	4	913.56		
Linear N ($\beta_1 x_i$)	1	651.47	183.74	<.0001
Quadratic N ($\beta_2 x_i^2$)	1	260.50	73.47	<.0001
Lack of fit (δ_i)	2	1.59	0.22	.8039
Error	8	28.37		

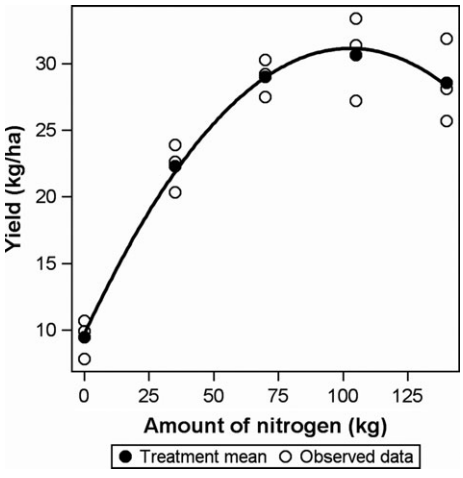

FIGURE 3 Yield of sugar beet (t/ha) plotted against five amounts of N-fertilizer (0, 35, 70, 105 and 140 kg N/ha) (Example 2). The fitted curve is Yield = 9.69 + 0.418 × N − 0.00203 × N^2

designed experiments, i.e., the absence of interaction between block and treatment factors, also known as block-by-treatment additivity (Hinkelmann & Kempthorne, 1994). Under this assumption, we can average the regression model over blocks to report an overall model. The marginal mean of regression model (4) over blocks is given by,

$$\mu + \bar{b}_\cdot + \beta_1 x_i + \beta_2 x_i^2, \tag{5}$$

where the mean intercept $\mu + \bar{b}_\cdot$ includes a term $\bar{b}_\cdot$ representing the average of the block effects. The mean curve drawn in Figure 3 is an estimate of (5). The parameter estimates, along with standard errors and confidence intervals, are shown in Table 6. A note of caution to R users: in R the first block effect is set to zero by the R default contrast setting and is not included in the output. But this block effect needs to be included (and counted) when computing the mean intercept $\mu + \bar{b}_\cdot$. However, the default contrast setting can be changed to set a sum-to-zero constraint on the block effects ($\bar{b}_\cdot = 0$) by using the `contr.sum` option in R: see `help(contrasts)` in R. With this option, the solution for μ corresponds to the overall mean intercept. The process of averaging across blocks used here is a special instance of a general algorithm for prediction in linear models, which is described, e.g., in Lane and Nelder (1982) and Welham, Cullis, Gogel, Gilmour, and Thompson (2004), and which is implemented in some linear model packages (Butler, Cullis, Gilmour, & Gogel, 2009).

Quadratic polynomials can be used to determine the optimal level of a quantitative input variable x_i. This is done by setting the first derivative with respect to x_i equal to zero and solving for x_i. For model (5), the equation for the optimum level of added nitrogen is $x_{opt} = -\beta_1/(2\beta_2)$. Plugging in the least squares solutions for β_1 and β_2 (Table 6) gives the quadratic model an estimate of x_{opt}. An approximate standard error can be computed by the delta method (Johnson, Kemp, & Kotz, 2005) using $\mathrm{var}(\hat{x}_{opt}) \approx \left(\frac{1}{4\hat{\beta}_2^2}\right)\mathrm{var}\left(\hat{\beta}_1\right) + \left(\frac{\hat{\beta}_1^2}{4\hat{\beta}_2^4}\right)\mathrm{var}\left(\hat{\beta}_2\right) - \left(\frac{\hat{\beta}_1}{2\hat{\beta}_2^3}\right)\mathrm{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)$. This yields the optimum $\hat{x}_{opt} = 102.7$ ($SE = 7.83$).

A polynomial model is an empirical approximation for an unknown theoretical model of the treatment effects and, for a general unconstrained empirical model, will normally include all polynomial terms up to and including the maximum significant polynomial of the fitted model. All polynomial terms of degree less than the degree of the maximum significant polynomial term are part of the empirical model, and significance tests on these terms are usually meaningless and irrelevant. This is the principle of "functional marginality," and except in very special situations, it is advisable to observe functional marginality and to retain all marginal (lower-degree) terms regardless of whether or not they are significant (Nelder, 2000). This marginality principle also applies when more than one quantitative treatment factor is involved, as will be discussed with Example 3.

TABLE 6 Parameter estimates, standard errors and 95% confidence limit for models (4) and (5) fitted to the sugar beet data (Example 2)

			95% Confidence limits	
Parameter	Estimate	*SE*	Lower	Upper
μ	11.52	1.13	8.99	14.04
b_1	−2.38	1.09	−4.82	0.06
b_2	−3.10	1.09	−5.54	−0.66
b_3	0	–	–	–
β_1	0.418	0.0318	0.347	0.489
β_2	−0.00203	0.000218	−0.00252	−0.00155
$\mu + \bar{b}_\cdot$	9.69	0.94	7.60	11.79

For the development of models, it is convenient to make a clear distinction between the treatment effects, which are the effects of direct interest, and the design effects which, although not of direct interest, must be accounted for in the full fitted model. Design effects represent the field layout and the randomization structures, and the model for these effects is also known as the block model or null model (Piepho, Büchse, & Emrich, 2003) or structural component (Welham et al., 2015, p. 5). In fact, the treatment and design models can be developed entirely independently due to the assumption of block-by-treatment additivity and we can emphasize the focus on the treatment model by representing equation (4) in two parts as follows:

$$y_{ij} = \mu + b_j + \tau_i + e_{ij}, \tag{6}$$

$$\tau_i = \beta_1 x_i + \beta_2 x_i^2. \tag{7}$$

Here $\mu + \tau_i$ is the qualitative factorial model for the i-th treatment, which can be modelled quantitatively by the second-degree polynomial $\mu + \beta_1 x_i + \beta_2 x_i^2$, where the treatment effect is shown in (7).

To fit models (2) and (3) with a linear model package, a qualitative factor is needed to code the treatment-specific lack-of-fit (LOF) effect δ_i (note that generally a qualitative level factor is represented by a separate effect for each level). For this purpose, we cannot use the quantitative variable that contains the amounts of nitrogen (N), because that variable is needed to fit the regression terms and therefore cannot be declared as a qualitative factor (note that the number of regression terms for a quantitative level factor is typically much lower than the number of levels). In the data set available for such an experiment, however, the quantitative amount of fertilizer (N) initially is likely to be the only variable representing the treatments. A simple way forward then is to copy the quantitative variable N into a new variable (LOF_N, say) and declare that variable as a qualitative factor. With this additional variable, model (2) can be coded as BLOCK + N + LOF_N, where BLOCK is the factor identifying the blocks. In R, the qualitative factor can also be generated from inside the model specification using the specification BLOCK + N + factor(N), but when interpreting the output one needs to be aware that "factor(N)" corresponds to the lack-of-fit term. Also, for R users, it may be more customary to fit two models, one with a quantitative factor for N (without LOF_N) and one with a qualitative factor for N. Then comparing both models using the anova() function will produce the same lack-of-fit test.

The sequential sum of squares shown in Tables 4 and 5 correspond to Type I sum of squares in some packages (e.g., SAS). We prefer these over Type III sum of squares (Searle, 1987), which would not return an *F* test for polynomial terms when a lack-of-fit effect is present. Note that some packages use Type III as the default setting, so users may need to take specific action to obtain sequential sums of squares. A very lucid discussion of the advantages of sequential (Type I) sums of squares and hypotheses is given in Nelder (1994) [also see Searle (1994) for a rejoinder].

The sequential tests given here for the linear and quadratic terms are equivalent to those obtained by using orthogonal polynomial contrasts. Computation and testing of polynomial terms via orthogonal contrasts was a convenience for polynomial regression before linear model computer packages became widely available. Nowadays, however, their use has become mostly obsolete for low-degree polynomials as it is much more convenient to fit such polynomial terms directly and to test them via the sequential *F* tests automatically furnished by any suitable computer package. The only exception to this recommendation is where it is necessary to fit a high-degree polynomial to a long time series when raw polynomials might become numerically unstable, especially when all levels are far removed from zero, as is the case with years. Then the use of orthogonal polynomial contrasts might be needed for numerical stability (Draper & Smith, 1998, Chapter 22).

To conclude this example, we re-emphasize that an unstructured analysis of quantitative level treatment factors is uninformative about any possible underlying model of treatment factor effects; treatment factors with more than two quantitative treatment levels should almost always be modelled quantitatively.

6 | EXAMPLE 1 (CONTINUED)—COMPARING SEVERAL REGRESSIONS IN AN EXPERIMENT WITH QUALITATIVE AND QUANTITATIVE TREATMENT FACTORS

Example 1 (continued): The rice experiment discussed above (Gomez & Gomez, 1984, p. 143) was laid out as a split-split-plot design, in which fertilizer was the main-plot factor, laid out in three complete blocks, management practice was the sub-plot (or split-plot) factor, completely randomized within main plots, and variety was the sub-sub-plot (or split-split-plot) factor, completely randomized within sub-plots. Thus, we will consider the model,

$$y_{ihjk} = \mu + b_k + \tau_{ihj} + f_{ik} + g_{ihk} + e_{ihjk}, \tag{8}$$

where y_{ijhk} is the yield of the *i*-th fertilizer treatment for the *h*-th management practice and *j*-th variety in the *k*-th complete block, μ is a general intercept, b_k is the effect of the *k*-th block, τ_{ihj} is the *ihj*-th treatment effect, f_{ik} is the main-plot error associated with the *k*-th block and *i*-th fertilizer level, assumed to be random with zero mean and variance σ_f^2, g_{ihk} is the sub-plot error associated with the *k*-th block, *i*-th fertilizer and *h*-th management practice, assumed to be random with zero mean and variance σ_g^2, and e_{ijhk} is a residual sub-sub-plot error with zero mean and variance σ^2. On account of the three random effects f_{ik}, g_{ihk} and e_{ihjk}, (8) is a mixed model, which we may fit using any REML package. It is particularly important to explicitly fit the main-plot, sub-plot and sub-sub-plot error terms as random to properly represent the three-stage randomization layout and hence to obtain valid inferences for the fixed effects. Each treatment effect has its own specific error term, depending on the experimental units to which the levels of the relevant treatment factor were randomly allocated. For instance, in the split-plot rice example the varieties were randomized to sub-sub-plots which means that information on variety effects is obtained from comparisons between different sub-sub-plots nested within individual sub-plots (a more formal way of saying this is that the information about these effects occurs in the sub-sub-plot stratum). So the sub-sub-plot error variance is the appropriate error term for testing the variety main effect and the interactions of variety with the other two factors. Similarly, the relevant error term for management main effects and interactions with fertilizer is the sub-plot error because management practices were allocated to sub-plots and information about these effects involves comparisons among different sub-plots nested within main plots. Finally, information about the fertilizer treatments requires comparisons between main plots, so the variance between main plots is the relevant error term for fertilizer main effects. To code the models for this example, variables are needed for the factor blocks, main plots, split plots, varieties and nitrogen fertilizer. The coding of these variables is described in Appendix 1.

Having taken care of all the design effects in (8), we can now focus on modelling the factorial treatment effects τ_{ihj}. The initial three-way analysis of variance (Table 1; also see Gomez & Gomez, 1984, p. 153), considering all treatment factors as qualitative, shows that management practice has a significant main effect, but that it does not interact with the other two treatment factors. This indicates that the slopes of the regressions on nitrogen do not depend on management and that the dependence of the regressions on management practice is only via the intercepts as represented by the management main effect. We subsequently do not explicitly write down the main effect for management practice in models for the treatment effect, and we drop the subscript *h* from the treatment effect for simplicity, but it is understood that in all analyses the main effect for management practice is fitted. It is acknowledged here that our analysis relies on the absence of nitrogen-by-management interaction based on a significance test which may not have very good power, so an alternative would be to further explore this interaction using polynomial regression, but this is not pursued here for brevity.

For this experiment, we have three separate varieties; therefore, the full model fitting separate regression models for each variety can be written as,

$$\tau_{ij} = \beta_{0j} + \beta_{1j} x_i. \tag{9}$$

Here x_i is the *i*-th amount of fertilizer, β_{0j} is an intercept term for the *j*-th variety, and β_{1j} is the linear slope for the *j*-th variety. Similarly, a quadratic model takes the form,

$$\tau_{ij} = \beta_{0j} + \beta_{1j}x_i + \beta_{2j}x_i^2, \quad (10)$$

where β_{2j} is the regression coefficient of the quadratic term for the j-th variety.

It now remains to model the interaction between variety and fertilizer. If there is no interaction, the regression curves for the four varieties run parallel, i.e., the curves for the different varieties merely differ in their intercepts. This model can be written as,

$$\tau_{ij} = \beta_{0j} + \gamma_1 x_i + \gamma_2 x_i^2. \quad (11)$$

For testing the interaction, it is necessary to re-express (10) in such a way that no-interaction model (11) can be obtained from (10) as a special case by simply setting variety-specific terms to zero, thus allowing a test of the null hypothesis of interest. Hence, we re-express the linear and quadratic terms in (10) as,

$$\beta_{1j} = \gamma_1 + \alpha_{1j} \text{ and} \quad (12)$$

$$\beta_{2j} = \gamma_2 + \alpha_{2j}, \quad (13)$$

respectively, where α_{1j} and α_{2j} are interaction effects. Substituting (12) and (13) into (10) then gives the general model,

$$\tau_{ij} = \beta_{0j} + \gamma_1 x_i + \gamma_2 x_i^2 + \alpha_{1j}x_i + \alpha_{2j}x_i^2. \quad (14)$$

When there is no interaction, we have $\alpha_{1j} = \alpha_{2j} = 0$ for all varieties and (14) simplifies to parallel-curves model (11) as required.

With the factorial decomposition in (14), we can now consider the lack of fit for both in the main effect ($\gamma_1 x_i + \gamma_2 x_i^2$) and the interaction ($\alpha_{1j}x_i + \alpha_{2j}x_i^2$). Thus, the extended model can be written as,

$$\tau_{ij} = \beta_{0j} + \gamma_1 x_i + \gamma_2 x_i^2 + \delta_i + \alpha_{1j}x_i + \alpha_{2j}x_i^2 + \delta_{ij}, \quad (15)$$

where δ_i and δ_{ij} are the lack-of-fit terms for the additive and interaction effects, respectively.

Table 7 shows no evidence of any lack of fit for the main effects (p = .1949), but the variety by lack-of-fit interaction term has a p-value of p = .0791 which may indicate some evidence of a lack of fit. Sometimes when a lack-of-fit effect is small compared to the major treatment effects of a model, it can be realistic to neglect the small lack of fit and to fit a parsimonious model that includes just the major treatment effects. Here the major components are the linear and quadratic nitrogen trend components, the variety main effects and the linear and quadratic variety-by-nitrogen interaction effects. There is no evidence of any quadratic variety-by-nitrogen interaction effect, so we fit a model with variety-specific linear terms and a common quadratic trend term. Combining this with basic model (8), and adding a main effect for the h-th management practice (θ_h) as well as the block effect (b_k), the regression curve for the j-th variety and h-th management practice in the k-th block is,

$$\mu + b_k + \theta_h + \beta_{0j} + \beta_{1j}x_i + \gamma_2 x_i^2, \quad (16)$$

where the intercept is given by $\mu + b_k + \theta_h + \beta_{0j}$. For reporting the final model, it is convenient to average (16) over blocks and management practices, which only requires computing the marginal mean for the intercept as $\mu + \bar{b}_\cdot + \bar{\theta}_\cdot + \beta_{0j}$. This approach is justified, because there is no significant management practice interaction with fertilizer and variety and because block-by-treatment additivity is a standard assumption for block designs (see the "Discussion" section for options to check this assumption). The final model is reported in Table 8. Comparing the linear slopes, it emerges that the response of variety V1 is significantly different from that of V2 and V3, but V2 and V3 do not differ significantly in their responses.

The analysis discussed above fits a model based on a quadratic polynomial model assuming a common quadratic trend for all three varieties. The graphical plots in Figure 1 show that the resulting polynomial model gives quite a good fit to the treatment means of the three varieties over the five nitrogen levels. However, as discussed in the REML mixed model analysis for this example, Table 1 shows an anomalously large sub-sub-plot error mean square which suggests that there may be some inconsistencies between the replicated treatment effects in the variety stratum of the analysis. This

TABLE 7 Sequential Wald-type F tests for quadratic regression model (15) fitted to field experiment with rice (Example 1)

Source	Numerator degrees of freedom	Denominator degrees of freedom[a]	F-value[a]	p-value
Blocks (b_k)	2	8	0.66	.5439
Management practice (θ_h)	2	108	53.15	<.0001
Linear N ($\gamma_1 x_i$)	1	8	100.58	<.0001
Quadratic N ($\gamma_2 x_i^2$)	1	8	6.16	.0379
Lack of fit (δ_i)	2	8	2.02	.1949
Variety (β_{0j})	2	108	255.02	<.0001
Variety × linear N ($\alpha_{1j}x_i$)	3	108	12.21	<.0001
Variety × quadratic N ($\alpha_{2j}x_i^2$)	3	108	1.00	.3730
Variety × lack of fit (δ_{ij})	6	108	2.15	.0791

[a]Obtained using the Kenward and Roger (1997) method. Sub-plot error variance component estimated to be zero.

TABLE 8 Fitted quadratic curves for the three rice varieties in field experiment (Example 1) based on equation (16)

Variety	Intercept[a] $\mu + \bar{b}_\cdot + \bar{\theta}_\cdot + \beta_{0j}$	Linear[a] β_{1j}	Quadratic γ_2
V1	4.51 a	0.0161 a	−0.00008
V2	5.05 b	0.0258 b	−0.00008
V3	6.52 c	0.0294 b	−0.00008
SED	0.254	0.00282	

[a]Intercepts and slopes followed by a common letter are not significantly different at the 5% level of significance by a t test. Note that the intercept comparison is of limited value here because a change of origin will change the intercepts by different amounts due to the differences in linear slopes.

anomaly together with the possibly non-negligible variety by lack-of-fit interaction effects in Table 7, suggests that the data may contain anomalous results for the individual treatment replications in the variety stratum of the analysis. Due to lack of space, we cannot discuss these issues in detail here, but for the interested reader, we have provided some further in-depth analysis of the data in the Supporting information for this paper. This further analysis does indeed show some evidence of anomalous data especially for the replicated treatments for variety V1 and a complete analysis of the data set does need to take proper account of these anomalous data. See also Section 3 of Example 1 in the agriTutorial package.

In the analysis discussed above, the same nitrogen-by-variety interaction model was fitted for all the nitrogen level treatments including the zero N level treatment. Sometimes a zero level control treatment is best regarded as a separate qualitative treatment rather than as part of a quantitative treatment series. The four applied N rates are spaced at 30-unit intervals, whereas the difference between the zero control treatment and the lowest N rate is a 50-unit interval which suggests that the zero N is not a natural part of the quantitative N series. For these reasons, it may be appropriate to exclude the control from the regression. An alternative polynomial regression analysis that excludes the control from the regression, but does include it in the analysis to make full use of the data (Piepho et al., 2006), is described in the Supporting information.

7 | EXAMPLE 3—POLYNOMIAL REGRESSION MODEL FOR AN EXPERIMENT WITH TWO QUANTITATIVE LEVEL TREATMENT FACTORS

Treatment factors in combination often interact, and polynomial models for two or more quantitative treatment factors will usually require the inclusion of polynomial interaction effects. Polynomial models are built by adding individual treatment effects in a strict order of importance where the lower-degree effects of a factorial combination must be added first before the higher-degree effects of that factorial combination. Furthermore, polynomial models have a special hierarchical or model marginality structure which must be respected to ensure model invariance under shifts of origin. Suppose that the levels of a quantitative p-level factor F_1 are given by a variable X and the levels of a quantitative q-level factor F_2 are given by a variable Y, then the polynomial interaction between the r-th-degree polynomial X^r for F_1 and the s-th-degree polynomial Y^s for F_2 can be written as X^rY^s where r is an integer power between 0 and $p-1$ and s is an integer power between 0 and $q-1$. The hierarchical or model marginality principle requires that if any term X^rY^s is included in a model, then all terms X^uY^v for both $u \leq r$ and $v \leq s$ simultaneously must also be included in the model, whether significant or not, see Peixoto (1990) and Nelder (2000). In particular, the marginality principle implies that if a particular interaction term X^rY^s is included in a model then so also must be the main effects X^r and Y^s. The converse is not true, and it is theoretically possible, for example, to have a valid marginal model with separate polynomial trends for the two factors F_1 and F_2 and no-interaction terms. However, such models are probably unrealistic and some authors insist that if a model has a set of main effect terms of n-th degree, then the model should also include all interaction terms of type X^rY^s for $r + s \leq n$ even if these terms are not statistically significant: these models are sometimes called n^{th}-order response surface models, see Draper and Smith (1998, Chapter 12) and McCullagh and Nelder (1989, Chapter 3.5.4).

The principles of fitting a polynomial model with two or more quantitative level factors will be illustrated in the following example which fits and tests a polynomial regression model in two quantitative level treatment factors.

Example 3: (Gomez & Gomez, 1984, p. 401): Nitrogen uptake (g/pot) of rice was studied in a two-factor greenhouse experiment involving duration of water stress (W) and level of nitrogen application (N). The experiment had four water-stress levels (0, 10, 20 and 40 days) as main-plot treatments and four nitrogen rates (0, 90, 180 and 270 kg/ha) as sub-plot treatments. The main plots were randomized into four complete blocks. The model for design effects comprised fixed effects for the randomized complete blocks and random effects for the nested main-plot and sub-plot errors.

Examination of the studentized residuals of the replicated treatments shows that the variance increases with the predicted mean; therefore, we analysed the uptake data on a logarithmic scale, which is known to stabilize variances that have this structure (see Snedecor & Cochran, 1989, Chapter 15). On a log scale, the main-plot error variance component is estimated to be zero, while the sub-plot error variance component estimate equals 0.02531. We also note that the change in the nitrogen uptake rate between maximum and minimum is about 1.5 orders of magnitude and on this scale a log transformation often gives a simpler model than does a linear scale.

One way to fit a polynomial model is to proceed in stages by first fitting the lowest degree effects and then progressively adding significant higher-degree effects, as required. The initial model tested here was a first-degree linear regression model with two linear regression terms in two variables, w_i the i-th water-stress level and, n_j the j-th nitrogen level, also known as a first-order response surface model (Box & Draper, 2007; Dean & Voss, 1999):

$$\tau_{ij} = \gamma_{10} w_i + \gamma_{01} n_j. \tag{17}$$

Table 9 shows lack-of-fit tests for the first-order model and shows very highly significant F-values for both of the linear trend term effects. In addition, the remaining treatment degrees of freedom in the model are shown as lack-of-fit tests both as an overall set of 13 degrees of freedom and as separate factorial sets with two lack-of-fit degrees of freedom for the trend effects of water, two lack-of-fit degrees of freedom for the trend effects of nitrogen and the nine lack-of-fit degrees of freedom for the water-by-nitrogen interaction effect. There were very highly significant lack-of-fit effects for each of these classifications which showed that at least a second-degree polynomial was needed for the trend effects of water, the trend effects of nitrogen and the water-by-nitrogen interaction effects.

TABLE 9 Sequential Wald-type *F* tests for first-order model (17) fitted to the greenhouse experiment with rice (Example 3). The response was log-transformed to stabilize the variance

Source	Numerator degrees of freedom	Denominator degrees of freedom[a]	*F*-value[a]	*p*-value
Blocks	3	45	1.19	.3231
Linear W ($\gamma_{10}w_i$)	1	45	1008.30	<.0001
Linear N ($\gamma_{01}n_j$)	1	45	183.81	<.0001
Lack of fit (δ_{ij})	13	45	14.24	<.0001
Lack-of-fit W	2	45	54.10	<.0001
Lack-of-fit N	2	45	12.44	<.0001
Lack-of-fit W × N	9	45	5.79	<.0001

[a]Obtained using the Kenward and Roger (1997) method. Main-plot error variance component estimated to be zero.
N = nitrogen; W = water.

The next stage in the model building process was to add terms for the second-degree effects of W, the second-degree effects of N and the linear W-by-linear N polynomial interaction effects. The resulting model, where the γ_{pq} (p, q = 0, 1, 2) are first-degree regression coefficients if $p + q = 1$ and second-degree regression coefficients if $p + q = 2$, is a second-degree polynomial regression model, also known as a second-order response surface model (Dean & Voss, 1999):

$$\tau_{ij} = \gamma_{10}w_i + \gamma_{20}w_i^2 + \gamma_{01}n_j + \gamma_{02}n_j^2 + \gamma_{11}w_i n_j. \quad (18)$$

Note that the first-order terms are always retained in a second-degree model irrespective of their significance as they are required to preserve the functional marginality of the fitted model.

Table 10 shows highly or very highly significant *F*-values for all second-degree terms. In addition, the remaining treatment degrees of freedom in the model are shown as lack-of-fit tests both as an overall set of 10 degrees of freedom and as separate factorial sets with one lack-of-fit degree of freedom for the trend effects of water, one lack-of-fit degree of freedom for the trend effects of nitrogen and the eight lack-of-fit degrees of freedom for the water-by-nitrogen interaction effects. From this partitioning, there was evidence of a significant third-degree trend effect for nitrogen (F = 10.28, p = .025). However, with only four nitrogen treatment levels, a third-degree nitrogen model fits the nitrogen response exactly, which carries a risk of over-fitting the model. Furthermore, a graphical inspection of the full third-degree trend curve shows an implausible dose–response relationship with unrealistically wavy-shaped nitrogen response curves for the individual water-stress treatments (results not shown). For these reasons, we believe that the second-order model with second-degree nitrogen trend effects is more useful for predictions of treatment effects than a possibly over-fitted model with a third-degree term for the nitrogen response model.

TABLE 10 Sequential Wald-type *F* tests for second-order model (18) fitted to the greenhouse experiment with rice (Example 3). The response was log-transformed to stabilize the variance

Source	Numerator degrees of freedom	Denominator degrees of freedom[a]	*F*-value[a]	*p*-value
Blocks	3	45	1.19	.3231
Linear W ($\gamma_{10}w_i$)	1	45	1008.30	<.0001
Quadratic W ($\gamma_{20}w_i^2$)	1	45	106.90	<.0001
Linear N ($\gamma_{01}n_j$)	1	45	183.81	<.0001
Quadratic N ($\gamma_{02}n_j^2$)	1	45	14.60	.0004
Linear W × linear N ($\gamma_{11}w_i n_j$)	1	45	46.09	<.0001
Lack of fit (δ_{ij})	10	45	1.76	.0972
Lack-of-fit W	1	45	1.31	.2591
Lack-of-fit N	1	45	10.28	.0025
Lack-of-fit W × N	8	45	0.75	.6497

[a]Obtained using the Kenward and Roger (1997) method. Main-plot error variance component estimated to be zero.
N = nitrogen; W = water.

The fitted equation for a second-order response model using nitrogen uptake on the log scale (base e) and assuming an averaged intercept across replicate blocks is:

$$\begin{aligned}\log_e(Y) = &-1.16 + 0.0176 \times \mathrm{W} - 0.00116 \times \mathrm{W}^2 \\ &+ 0.00680 \times \mathrm{N} - 9.38 \times 10^{-6} \times \mathrm{N}^2 \\ &- 9.07 \times 10^{-5} \times \mathrm{W} \times \mathrm{N}.\end{aligned} \quad (19)$$

Using this equation, Figures 4a shows graphical plots of the water-stress response evaluated at each observed level of nitrogen, while Figure 4b shows graphical plots of the nitrogen rate response evaluated at each observed level of water stress. The graphs in Figures 4a and 4b are back-transformed to the original N uptake scale of measurement and show the actual observed nitrogen uptake data. In particular, Figure 4a shows how the nitrogen uptake rate approaches zero asymptotically as the water stress increases. However, for a proper assessment of the fit of the quadratic regression model, equation (19) needs to be plotted against the log transformed data and graphical plots of the transformed data for this example can be seen by running the R code for Example 3 in the agriTutorial package which will generate the corresponding log transformed plots for Figs 4a and 4b. In agriTutorial, the four replicate log nitrogen uptake data values are shown for each treatment combination and display reasonable homogeneity of variance across the range of treatments.

Figure 4a shows that the nitrogen uptake depended directly on the amount of nitrogen applied, but that the rate of uptake was greatest at the lowest level of water stress and least at the highest level of water stress. Similarly, Figure 4b shows that the nitrogen uptake response curve depended on the amount of nitrogen applied, but that the response curve was higher and steeper at the lowest levels of water stress and lowest and flattest at the highest level of water stress. The fitted second-order polynomial model successfully captured these complex factorial interaction effects, and this example shows the value of empirical polynomial models for complex factorial experiments.

Note that the fitted second-order model is considerably simpler than the one reported in Gomez and Gomez (1984) who use

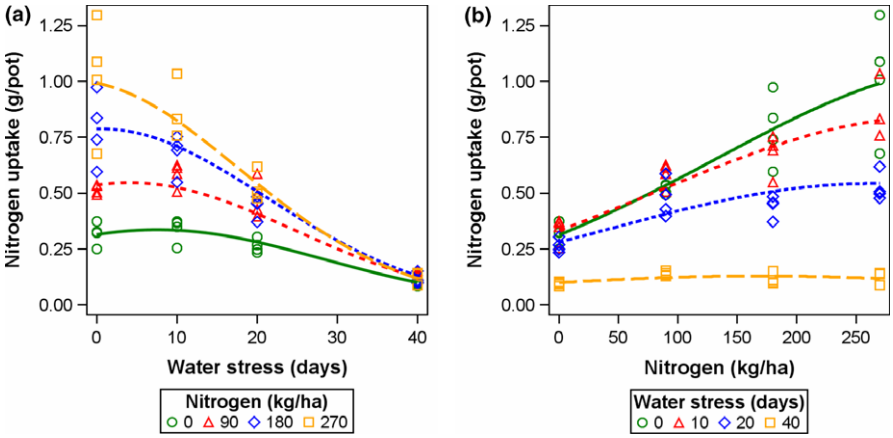

FIGURE 4 Polynomial regression (equation 19) plotting the response on the untransformed nitrogen uptake scale (Example 3). (a) The marginal model for water stress is evaluated at the observed levels of nitrogen fertilization, i.e., 0, 90, 180 and 270 kg/ha. (b) The marginal model for nitrogen is evaluated at the observed levels of water stress, i.e., 0, 10, 20 and 40 days [Colour figure can be viewed at wileyonlinelibrary.com]

orthogonal contrasts up to the highest possible third order including all interaction terms up to that order for both quantitative variates (n_i and w_j) and analysed data on the original scale. Also note that these authors seem to have inadvertently coded the highest stress level as 30 days instead of 40 days when computing orthogonal polynomial contrasts.

8 | EXAMPLE 4—AN EXPERIMENT WITH ONE QUALITATIVE TREATMENT FACTOR AND REPEATED MEASURES IN TIME

Example 4: Milliken and Johnson (1992, p. 429) describe an experiment with four sorghum varieties, in which the leaf area index was assessed in five consecutive weeks starting 2 weeks after emergence. The experiment was replicated five times and was laid out in five randomized complete blocks. Figure 5 shows repeated measurements plotted against week using different symbols for the different varieties. The figure also shows the best fitting quadratic regression curve over weeks for each variety. The following section will describe how these curves were fitted to the auto-correlated repeated measurements data.

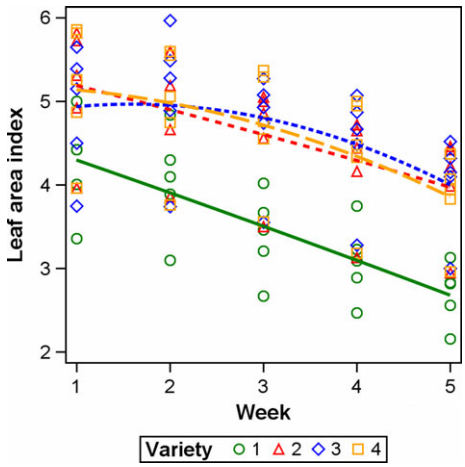

FIGURE 5 Plot of fitted quadratic polynomials for sorghum data (Example 4) [Colour figure can be viewed at wileyonlinelibrary.com]

The example is a regression problem with one qualitative factor (variety) and one quantitative factor (week), so the setting is similar to that in Example 1. However, there is a major difference as the week factor is not a treatment factor that can be randomized. Instead, repeated measurements are taken on each plot on five consecutive occasions. Successive measurements on the same plot are likely to be serially correlated, and this means that for a reliable and efficient analysis of repeated-measures data we need to take proper account of the serial correlations between the repeated measures (Piepho, Büchse, & Richter, 2004; Pinheiro & Bates, 2000).

To build a suitable model taking serial correlation into account, it is convenient to initially consider the model for a single time point:

$$y_{ij} = \mu + b_j + \tau_i + e_{ij}. \quad (20)$$

Here y_{ij} is the leaf area of the i-th variety in the j-th complete block, μ is a general intercept, b_j is the effect of the j-th block, τ_i is the i-th variety effect, and e_{ij} is a residual plot error with zero mean and variance σ^2. This type of model can be assumed to hold in each week, but it is realistic to assume that the treatment effects evolve over time and thus are week-specific. Importantly, we must also allow for the block effects to change over time in an individual manner. For example, there could be fertility or soil type differences between blocks and these could have a smooth progressive or cumulative time-based effect on differences between the blocks dependent on

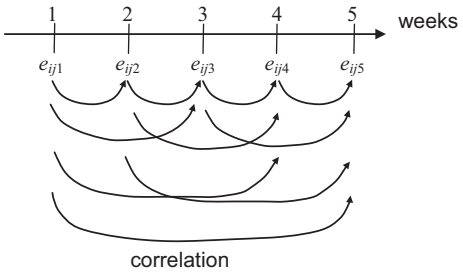

FIGURE 6 Schematic representation of correlation among errors of five repeated measurements on the same experimental unit. The arrows indicate different time lags. Examples: $e_{ij1} \rightarrow e_{ij2}$ is a lag of 1. $e_{ij1} \rightarrow e_{ij3}$ is a lag of 2. $e_{ij1} \rightarrow e_{ij3}$ is a lag of 3. $e_{ij1} \rightarrow e_{ij5}$ is a lag of 4

factors such as temperature or rainfall. Effects of this type are amenable to modelling by smooth time-based functions fitted to the individual block effects over time. Thus, to extend the model for the joint analysis of several weeks, we add a subscript t for weeks to all effects, indicating that all effects are week-specific:

$$y_{ijt} = \mu_t + b_{jt} + \tau_{it} + e_{ijt}. \quad (21)$$

The error terms e_{ijt} must now be allowed to be serially correlated (Figure 6). Before modelling the treatment effect, a variance–covariance model needs to be identified for these correlations. This is best done by using a saturated model for treatments and time, i.e., a model that considers all treatment factors and time as qualitative. It is important to note that measurements taken on the same observational unit (plot in this case) are serially correlated, but observations on different units can be considered independent because of the randomized allocation of treatments to units. Units on which repeated observations are taken are often referred to as subjects, and in the specification of these models, it is important to suitably identify the subjects. In SAS, for example, `subject=` option is used for that purpose. In R, subjects are identified after a vertical bar in the argument needed with the respective correlation function or in the random effect specification. Observations on the same subject are serially correlated, whereas observations on different subjects are independent.

There are many models for serial correlation of longitudinal data (repeated measurements over time). One popular model is the first-order autoregressive AR(1) model, under which the correlation among the observations at time points t and t' on the same plot is (dropping subscripts for plots for simplicity),

$$\mathrm{cov}(e_t, e_{t'}) = \sigma^2 \rho^{|t-t'|} \ (0 \le \rho < 1), \quad (22)$$

where ρ is the autocorrelation parameter. It is assumed here that the index t runs in temporal order, so $t = 1$ is the first time point, $t = 2$ is the second, etc. Thus,

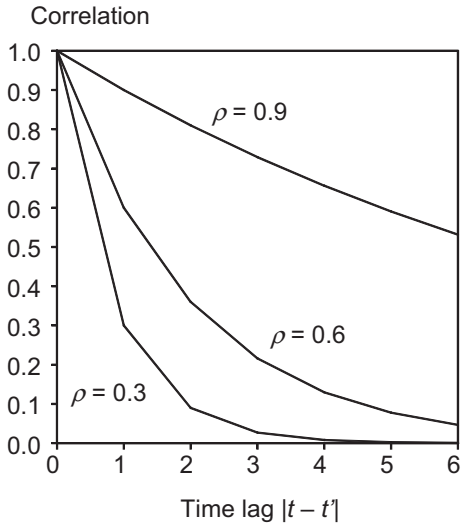

FIGURE 7 Correlation under an AR(1) model as a function of time lag for different values of the autocorrelation ρ

$$\begin{aligned} \mathrm{cov}(e_1, e_1) &= \sigma^2 \rho^{|1-1|} = \sigma^2 \\ \mathrm{cov}(e_1, e_2) &= \sigma^2 \rho^{|1-2|} = \sigma^2 \rho \\ \mathrm{cov}(e_1, e_3) &= \sigma^2 \rho^{|1-3|} = \sigma^2 \rho^2 \\ \mathrm{cov}(e_1, e_4) &= \sigma^2 \rho^{|1-4|} = \sigma^2 \rho^3 \ \text{etc.} \end{aligned} \quad (23)$$

The larger the distance in time (time lag), the lower is the covariance and correlation. This is depicted in Figure 7 for a few values of the autocorrelation ρ.

This correlation model is useful, if all time points are equally spaced. With unequal spacing, the model could be extended as follows:

$$\mathrm{cov}(e_t, e_{t'}) = \sigma^2 \rho^{d(t,t')}, \quad (24)$$

where $d(t, t')$ is the time lag between time points t and t'. This model is known as the power model of the continuous AR(1) (CAR1) model, and it can also be parameterized as an exponential model with $\rho = \exp(-1/\theta)$ (Pinheiro & Bates, 2000, p. 232). There are many alternative models, which allow the correlation to decay with distance in time (Schabenberger & Pierce, 2002). One particularly useful option is to add an independent error term (a "nugget" variance σ_N^2) to an autocorrelated model, which increases the total residual variance to $\sigma^2 + \sigma_N^2$ but uses the same spatial covariance model as without the nugget. A very flexible model for equally spaced data is the Toeplitz model (not currently available in "nlme"), which fits a separate covariance parameter for each lag distance. The only restriction of this model is that it assumes homogeneity of variance across time points. The most general model is to let both the variance at a time point and the covariances between time points vary freely. This unstructured (UN) model can be represented as,

$$\mathrm{cov}(e_t, e_{t'}) = \sigma_t \sigma_{t'} \rho_{tt'}, \quad (25)$$

where σ_t^2 is the variance at time t and $\rho_{tt'}$ is the correlation between time points t and t'. At the opposite extreme, we could postulate equal variance and correlation, which is known as the compound symmetry (CS) model or equal correlation model, such that,

$$\begin{aligned} &\mathrm{cov}(e_t, e_{t'}) = \sigma^2 \rho (t' \ne t) \text{ and} \\ &\mathrm{var}(e_t) = \sigma^2. \end{aligned} \quad (26)$$

This model is not usually a very realistic one, as the correlation function would correspond to a horizontal line in Figure 7 at correlation ρ, but nonetheless the model sometimes fits quite well.

As this brief review indicates, for repeated-measures designs there is a range of choices of potential variance–covariance models. A standard procedure is to fit a set of candidate models and to pick the best fitting one based on the Akaike information criterion (AIC) (Burnham & Anderson, 2002), which is computed from $-2 \log L_R + 2p$, where p is the number of variance–covariance parameters and $\log L_R$ is the maximized residual log-likelihood. The term $2p$ acts as a penalty for model complexity and helps provide a balance between model realism on the one hand and model parsimony on the other. The smaller the value of AIC, the better is the fit. The AIC is standard output from most mixed model packages, and Table 11 shows the AIC values for the range of covariance models discussed above. The AR(1) model with a nugget gave the best fit because it

TABLE 11 Values of twice the negative residual log-likelihood, the Akaike information criterion (AIC) and the change in AIC compared to the best fitting model (ΔAIC) for different variance–covariance structures for plot effect with model (22) (Example 4)

Covariance model[a]	$-2 \log L_R$[b]	AIC	ΔAIC[c]
ID	−3.7	−1.7	40.8
CS	−45.6	−41.6	0.9
AR(1)	−46.0	−42.0	0.5
AR(1) + nugget	−48.5	−42.5	0
Toeplitz	−48.6	−38.6	3.9
UN	−62.9	−32.9	9.6

[a]The same covariance model was fitted to both block and error effects. ID, independent model; CS, compound symmetry; AR(1), first-order autoregressive; UN, unstructured model.
[b]L_R = maximized residual likelihood.
[c]Change in AIC relative to the best fitting AR(1)+nugget model.

TABLE 12 Fitted covariances for different lags for the Toeplitz, the AR(1), the AR(1)-plus-nugget and the CS model

	Autocorrelation			
Time lag	Toeplitz	AR(1)	AR(1)-plus-nugget	CS
0	1.0000	1.0000	1.0000	1.0000
1	0.7507	0.7490	0.7518	0.7008
2	0.6759	0.5610	0.6826	0.7008
3	0.6251	0.4202	0.6198	0.7008
4	0.5414	0.3147	0.5627	0.7008

had the smallest value of AIC. The autocorrelation estimate for the error based on the AR(1) model with nugget is 0.9080, which is substantial. Note that the best model according to the simple log-likelihood criterion was UN, but this criterion takes no account of the number of estimated variance parameters p in the variance model which, in the case of the UN model, was $p = 15$, compared to only $p = 2$ for the AR(1) model. The simple log-likelihood criterion will be biased by over-fitting which shows the need for the $2p$ penalty in the AIC criterion. Table 12 shows the fitted covariances for different lags (known as the autocorrelation function, ACF) for the Toeplitz, the AR(1), the AR(1)-plus-nugget and the CS model. The fitted covariances confirm that the AR(1)-plus-nugget model gives a very good fit, followed by the AR(1) model without nugget. Table 12 illustrates that the Toeplitz model provides a very convenient check for the fit of simple autocorrelated models that assume homogeneity of variance.

The use of AIC to select a variance–covariance structure for random effects can be contrasted to our use of significance testing for selecting fixed-effects parameters. Significance tests (i.e., likelihood ratio tests in this case) can also be used to select between variance–covariance structures that are hierarchically nested, but not all structures meet this requirement, hence our preference for AIC. Conversely, AIC could also be used to select fixed-effects model components, but this would require switching from REML to full maximum likelihood (ML) estimation. As REML is preferable to ML for variance parameter estimation (Searle et al., 1992) and good distributional approximations are available for fixed-effects hypothesis testing (Kenward & Roger, 1997, 2009), we prefer Wald-type F tests and t tests for inference on fixed-effects model terms.

TABLE 13 Sequential Wald-type F tests for linear model (27) fitted to the sorghum data (Example 4)

Source	Numerator degrees of freedom	Denominator degrees of freedom[a]	F-value[a]	p-value
Block × week (b_{jt})	24	38.0	82.61	<.0001
Variety (β_{0i})	3	12.3	93.75	<.0001
Variety × linear week ($\beta_{1i}x_t$)	3	13.2	19.10	<.0001
Lack of fit (δ_{it})	9	35.6	9.50	<.0001

[a]Obtained using the second-order method of Kenward and Roger (2009).

TABLE 14 Sequential Wald-type F tests for quadratic model (28) fitted to the sorghum data (Example 4)

Source	Numerator degrees of freedom	Denominator degrees of freedom[a]	F-value[a]	p-value
Block × week (b_{jt})	24	38.0	82.61	<.0001
Variety (β_{0i})	3	12.3	93.75	<.0001
Variety × linear week ($\beta_{1i}x_i$)	3	13.2	19.10	<.0001
Variety × quadratic week ($\beta_{2i}x_t^2$)	3	43.5	16.16	<.0001
Lack of fit (δ_{it})	6	29.0	6.20	<.0001

[a]Obtained using the second-order method of Kenward and Roger (2009).

Having selected a suitable variance–covariance model for the repeated plot errors, we are now ready to select a regression model for time trend, proceeding as with the other examples. The linear trend model for treatments with lack-of-fit effect is,

$$\tau_{it} = \beta_{0i} + \beta_{1i}x_t + \delta_{it}, \quad (27)$$

where x_t is the time variable ($x_t = t$ for the t-th week in our case). The lack of fit for this model is significant (Table 13). We have used the second-order method of Kenward and Roger for approximating the denominator degrees of freedom, which is the preferred method when the variance–covariance structure is non-linear in the parameters as is the case with the AR(1) model (23) (Kenward & Roger, 2009).

Adding the quadratic term,

$$\tau_{it} = \beta_{0i} + \beta_{1i}x_t + \beta_{2i}x_t^2 + \delta_{it}, \quad (28)$$

the lack of fit is still significant (Table 14). Even adding a cubic term does not lead to a non-significant lack of fit (not shown), so polynomial regression does not provide a fully convincing fit here.

Inspection of Figure 5 indicates, however, that the quadratic model fits reasonably well, despite the residual lack of fit, so we proceed with that model.

To test the interaction, we would normally partition the polynomial terms as was done in (15) to give the model shown in (29):

$$\tau_{it} = \beta_{0i} + \gamma_1 x_t + \gamma_2 x_t^2 + \alpha_{1i} x_t + \alpha_{2i} x_t^2. \quad (29)$$

It is noted, however, that in this particular case we have fitted fixed block-by-week effects, which are confounded with time main effects γ_1 and γ_2. Hence, the effects β_{1i} and β_{2i} in (28) are actually equivalent to the interaction effects α_{1i} and α_{2i} in (29). This can also be seen from the degrees of freedom for these two effects, which equal three instead of four in Tables 13 and 14. Hence, we can conclude from Table 14 that the interaction is significant for both the linear and the quadratic terms, so variety-specific quadratic curves need to be estimated as shown in Figure 5.

A slight subtlety remains to be clarified regarding the fitted curves in Figure 5. As with the other examples, we need to average the model $\eta_{ijt} = \mu_t + b_{jt} + \tau_{it}$ across blocks at each time point, yielding,

$$\bar{\eta}_{i\cdot t} = \mu_t + \bar{b}_{\cdot t} + \tau_{it}. \quad (30)$$

The important difference compared to examples considered so far is that the sum of the intercept μ_t and the average block effects $\bar{b}_{\cdot t}$ is time-specific and therefore is a function of the predictor variable (time) used to model the treatment effect (x_t). Thus, to present the marginal model as a smooth function of time, we need to assume a regression model for both the sum of intercept and average block effect $\mu_t + \bar{b}_{\cdot t}$ as well as the treatment effect τ_{it}. Using a quadratic polynomial for treatment effects, $\tau_{it} = \beta_{0i} + \beta_{1i} x_t + \beta_{2i} x_t^2 + \delta_{it}$, and a quadratic polynomial for block effects, $\mu_t + b_{jt} = \theta_{0j} + \theta_{1j} x_t + \theta_{2j} x_t^2$, the marginal model becomes,

$$\bar{\eta}_{i\cdot t} = (\bar{\theta}_{0\cdot} + \beta_{0i}) + (\bar{\theta}_{1\cdot} + \beta_{1i}) x_t + (\bar{\theta}_{2\cdot} + \beta_{2i}) x_t^2. \quad (31)$$

This marginal model was used to generate the curves shown in Figure 5. The corresponding coefficient estimates are shown in Table 15.

With repeated-measures data, it is vital to make proper allowance for the various sources of variability in the data. In the above example, four crop varieties were measured over time with each variety replicated five times in a complete randomized blocks design. However, the experiment was done in only 1 year and at only one place; therefore, the responses were specific to that year and that place. The standard assumption for this type of data is that site and year effects are additive effects that can vary between time points but which are constants for all plots at the same time point. This means that differencing between plots will automatically eliminate site and year effects from all treatment contrasts. While this assumption is made for any individual experiment, we emphasize it here because of the important consequence that overall time trends cannot be interpreted, whereas contrasts among trends for varieties can. Hence, in the example analysis, the differences between the time curves for the four different varieties will automatically provide unbiased estimates of the time curve treatment contrasts free of all site and year effects.

Although in this example the absolute time response curves have no particular interpretation, in some situations the individual plots might be a random sample of plots drawn from a defined population of plots. Then it could be reasonable to assume that the time trend estimates are unbiased estimates of the true time curve trend for that population of experimental units (plots). To give an alternative example, if animals or plants were drawn at random from a population of animals or plants, then it could be reasonable to assume that the samples were representative of those populations. In that situation, it would be feasible to estimate absolute time trend curves for the treatment effects together with their population variances. The analysis of repeated-measures data where plots are a true random sample from a population of plots is straightforward but will not be discussed in detail here.

Analyses presented so far for Example 4 were obtained with SAS. For comparison, we also fitted the model using the function gls() of the "nlme" package in R. There are some subtle differences in output obtained for this repeated-measures example, which are shown in Appendix 2 and explained in Appendix 3.

TABLE 15 Coefficient estimates for model (31) fitted to the sorghum data (Example 4)

Variety	Intercept $(\bar{\theta}_{0\cdot} + \beta_{0i})$	Linear $(\bar{\theta}_{1\cdot} + \beta_{1i})$	Quadratic $(\bar{\theta}_{2\cdot} + \beta_{2i})$
1	4.68	−0.379	−0.0043
2	5.46	−0.271	−0.0054
3	4.75	0.266	−0.0833
4	5.18	0.010	−0.0549
SED[a]	0.168	0.114	0.0186

[a]Obtained using the second-order method of Kenward and Roger (2009).

TABLE 16 Sequential Wald-type *F* tests for second-order polynomial regression model for turnip data using row width and density as predictor variables (Example 5)

Source	Numerator degrees of freedom	Denominator degrees of freedom	*F*-value	*p*-value
Replicates	2	38	43.62	<.0001
Linear density	1	38	360.49	<.0001
Quadratic density	1	38	157.74	<.0001
Linear row spacing	1	38	13.22	.0008
Quadratic row spacing	1	38	5.71	.0219
Linear density × linear row spacing	1	38	11.41	.0017
Lack-of-fit density	2	38	68.29	<.0001
Lack-of-fit row spacing	1	38	4.47	.0411
Lack-of-fit density × row spacing	11	38	1.28	.2740

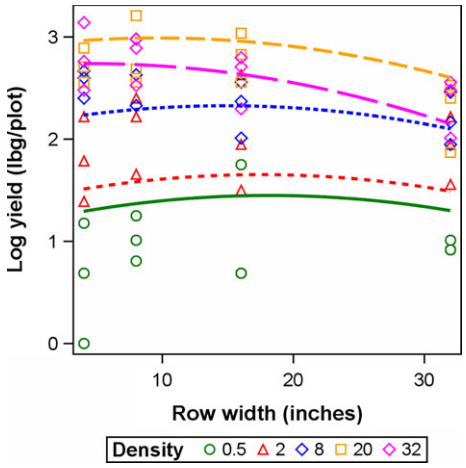

FIGURE 8 Plot of $\log_e$ yield (lb/plot) versus row width (inches) for turnip data and fitted second-order polynomial model (Example 5) [Colour figure can be viewed at wileyonlinelibrary.com]

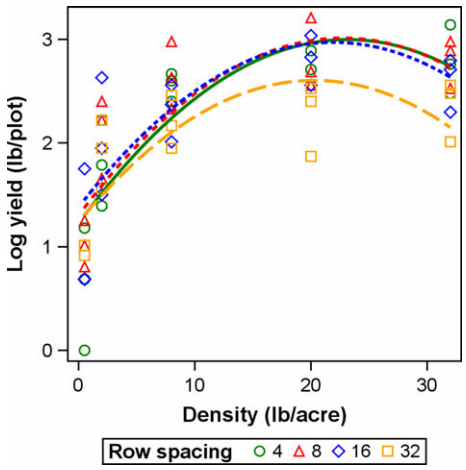

FIGURE 9 Plot of $\log_e$ yield (lb/plot) versus seeding density (lb/acre) for turnip data and fitted second-order polynomial model (Example 5) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 17 Sequential Wald-type *F* tests for second-order polynomial regression model for turnip data using log(row width) and log(density) as predictor variables (Example 5)

Source	Numerator degrees of freedom	Denominator degrees of freedom	*F*-value	*p*-value
Replicates	2	38	43.62	<.0001
Linear density	1	38	599.40	<.0001
Quadratic density	1	38	53.62	<.0001
Linear row spacing	1	38	7.16	.0110
Quadratic row spacing	1	38	15.57	.0003
Linear density × linear row spacing	1	38	19.16	<.0001
Lack-of-fit density	2	38	0.90	.4162
Lack-of-fit row spacing	1	38	0.67	.4175
Lack-of-fit density × row spacing	11	38	0.57	.8377

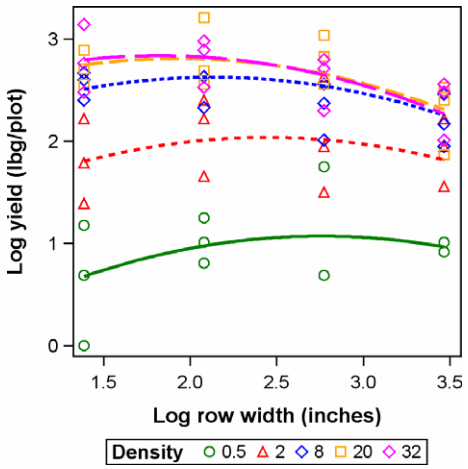

FIGURE 10 Plot of $\log_e$ yield (lb/plot) versus $\log_e$ row width (inches) for turnip data and fitted second-order polynomial model (Example 5) [Colour figure can be viewed at wileyonlinelibrary.com]

9 | EXAMPLE 5—TRANSFORMING TREATMENT LEVELS

The flexibility of polynomials can be enhanced by considering transformations of the quantitative treatment levels.

Example 5: Mead (1988, p. 323) describes an experiment on spacing effects with turnips, which was laid out in three complete blocks. Five different seed rates (0.5, 2, 8, 20 and 32 lb/acre) were tested in combination with four row widths (4, 8, 16 and 32 inches), giving rise to a total of 20 treatments. There were three replicates of each treatment and the design was arranged in three complete randomized blocks. Turnip yields (in lb per plot) were logarithmically transformed for analysis because this stabilized the variance (Mead, 1988; also see Figure 12).

We first fitted a second-order polynomial model based on the actual seed rates and row widths (Table 16), and Figures 8 and 9 show plots of the log yield versus row spacing (row widths) and plant density (seed rates), respectively, where the fitted curves are based on the fitted model. It appears from these plots that a polynomial based on the actual row width and density variables will not provide a satisfactory fit. The analysis of variance of the second-order polynomial model based on the actual untransformed predictor variables shows a very highly significant lack of fit for the density effect and significant lack of fit for the spacing effect at the 5% level (Table 16), confirming the visual impression from Figures 8 and 9.

We next fitted a second-order polynomial model based on the log-transformed seed rates and the log-transformed row widths (Table 17). Figures 10 and 11 show plots of the log yield versus log row spacing (log row widths) and log plant density (log seed rates), respectively, where the fitted curves are based on the fitted model. These plots suggest that a quadratic function fits reasonably well on the log row width scale (Figure 10) and, similarly, that a quadratic

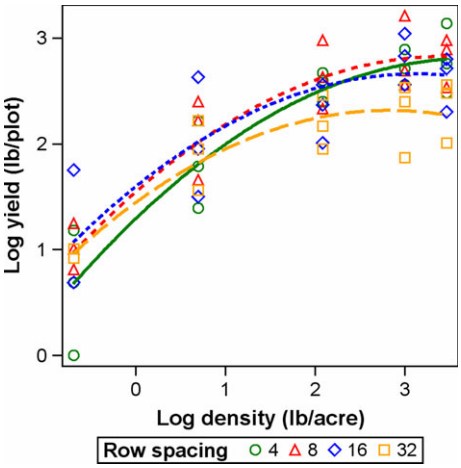

FIGURE 11 Plot of $\log_e$ yield (lb/plot) versus $\log_e$ seeding density (lb/acre) for turnip data and fitted second-order polynomial (Example 5) [Colour figure can be viewed at wileyonlinelibrary.com]

model fits reasonably well on the log density scale (Figure 11). The lack of fit for this model was non-significant for both of the main effects terms and for the interaction term (Table 17), and all polynomial terms were significant, demonstrating that this model fits the data well. We therefore conclude that a second-order polynomial model as shown in (18) using logarithmic row width and logarithmic density for the predictor variables gives a significantly better fit to the data than does a model based on the original untransformed predictor variables.

We note that the choice of model is of critical importance for this example, especially for the prediction of density effects on yield. In the first (untransformed) analysis, the fitted model predicts a maximum crop yield at around 20–25 lb of seed per acre for all four row widths (Figure 9), whereas in the second (transformed) analysis, the interpretation is more complex with only the widest row width showing evidence of a maximum at crop yield at the very high end of the seed rate scale (Figure 10). Hence, the choice of a model could have a very major impact on the correct interpretation of this experiment. Results on the transformed scale suggest that the optimal row spacing decreases with increasing seeding density (Figure 10) and the optimal seeding density decreases with increasing row spacing (Figure 11). This can be explained by competition within rows, which increases with the number of plants per row. Thus, for any given seeding density, the wider the row spacing the more seeds need to be sown per row, and hence the more within-row competition there will be.

The preferred model uses a logarithmic transformation of the response variable, and this model has studentized residuals that meet the assumptions of normality and homogeneity of variance quite well (Figure 12). The logarithmic transformation is helpful in stabilizing the variance because the observed yields have a wide range due to the wide range of seeding densities (Snedecor &

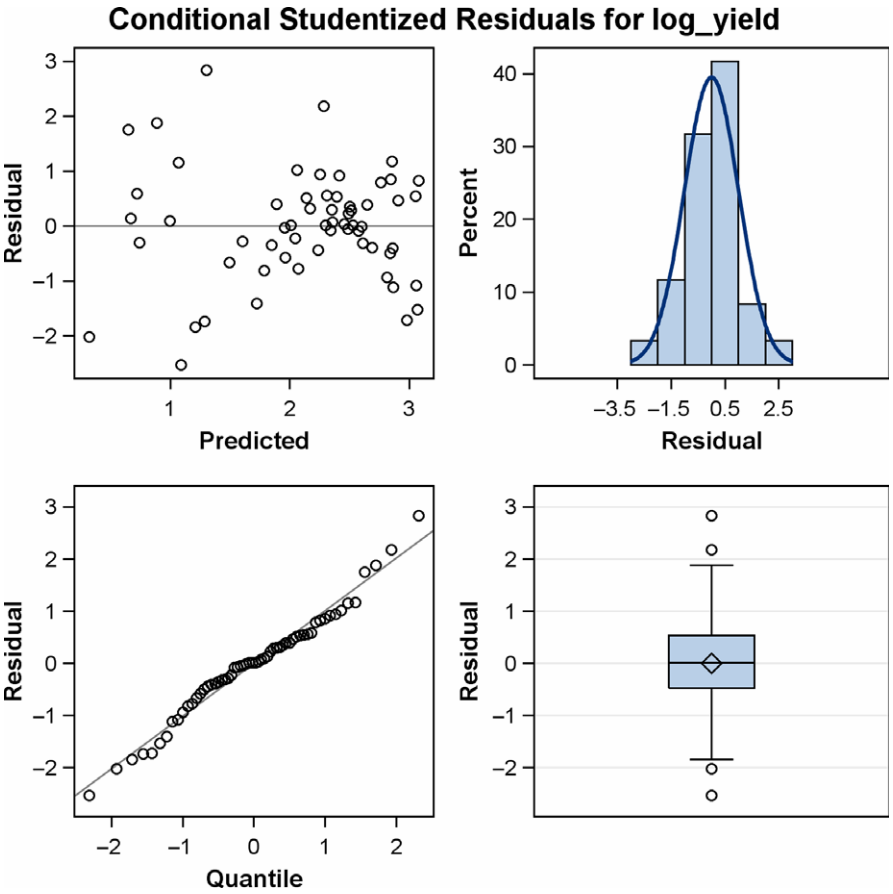

FIGURE 12 Studentized residual plots from a second-order polynomial model of $\log_e$(yield) of turnip data fitted to $\log_e$(row width) and $\log_e$(seeding density) (Example 5) [Colour figure can be viewed at wileyonlinelibrary.com]

Cochran, 1989, Chapter 15). When the response has a more narrow range, a transformation is usually not critical.

10 | DISCUSSION

Regression analysis is a powerful methodology for the analysis of experimental designs with quantitative level treatment factors. Although the analysis of such designs is often based on a comparison of the treatment means using multiple comparisons tests, such comparisons take no account of the ordered structure of the observations and are totally inappropriate for structured treatment comparisons. A suitable regression model gives insight into the functional relationships underlying the quantitative level treatment factors and also provides a powerful parameterization of the data that allows for inference and prediction of treatments effects (Pearce, 2005).

Regression analysis can also be used for the control of quantitative nuisance factor effects such as spatial position (Federer & Crossa, 2005) or initial size or initial weight or initial fertility of each experimental unit (Mead et al., 2012). Sometimes regression on nuisance factors is called analysis of covariance (Milliken & Johnson, 2002), but, apart from the need to fit the nuisance factors first before the treatment factors in an analysis of variance, there is no real distinction between regression and covariance analysis. Both methodologies are best regarded as different applications of basic regression analysis methodology.

A further use of regression analysis is to find the levels of two or more quantitative treatment factors minimizing or maximizing the response, as illustrated with Example 2. This is polynomial response surface modelling and is similar to ordinary polynomial regression analysis except that a response surface usually includes all polynomial terms up to the degree of the response surface, irrespective of the significance of the individual terms (Dean & Voss, 1999; Nelder, 2000). Also, there are specialized designs for response surface regression for efficiently locating optima (Box & Draper, 2007). In agronomic research, the emphasis is usually on improving understanding of factor effects rather than on finding maxima or minima; therefore, for agronomic experiments, response surface design and analysis based on factorial models (Edmondson, 1991; Piepho, Herndl, Pötsch, & Bahn, 2017) may be more useful than highly specialized designs aimed at locating the optima of a response surface.

Discussions in the literature rarely explain in sufficient detail and with suitable examples how to provide a full analysis of both qualitative and quantitative treatment factors in designed experiments. For this reason, this paper has presented a detailed set of examples illustrating various features of the practical analysis of data from real experiments. It is impossible to show a representative example of every possible type of analysis for a designed experiment, but the examples in this paper do cover some of the more common types that occur in practice. It is important to appreciate that modern computer software, properly used, is now capable of providing a full and efficient analysis of some quite complex experimental designs. It is also true, however, that modern computer software provides a black-box approach to analysis for many users and the use of examples for illustrating the methodology of the analysis and the interpretation of the output will be essential for many non-specialists.

The initial step of any regression analysis is to plot the raw data, as well as the treatment means, against the levels of the quantitative treatment variables to identify a suitable model. As the five examples considered in this paper show, a full analysis of a factorial experiment produces substantial output for the various steps. The most crucial piece of information is the final fitted model, including parameter estimates and associated standard errors and confidence intervals, and an analysis-of-variance table or table of Wald-type *F* tests substantiating the model choice. The most effective means to report the fitted model is a plot of the curves as shown in figures for all examples.

A complete analysis of data from designed experiments should always be accompanied by careful inspection of model assumptions. These assumptions pertain to effects for treatments, blocks and residual error terms. Residual error effects are usually assumed to follow a normal distribution with constant variance. These two assumptions can be checked by inspection of residual plots (Kozak & Piepho, 2018) such as those shown in Figure 12. In case of violations, the simplest remedy is to identify a data transformation that restores validity of the assumptions (Atkinson, 1987), as was done in Examples 3 and 5. Other options include the use of generalized linear mixed models (GLMM) (Lee, Nelder, & Pawitan, 2006; McCullagh & Nelder, 1989) or non-parametric methods (Shah & Madden, 2004), but these methods are beyond the scope of this paper.

A further common assumption in blocked experiments is that residual errors within the same error stratum are independent of each other. This assumption of independence can, however, be violated in a number of ways, and one especially important case of non-independence of observations occurs with repeated measures as discussed in Example 4 above. Mixed modelling, allowing for a serial correlation structure for observations taken on the same experimental unit, as illustrated with Example 4, relaxes this assumption and allows for serially correlated observations to be suitably modelled using a properly defined error structure.

Finally, a key assumption is the additivity of block and treatment factors, i.e., the absence of any interaction between these factors. This assumption is often accepted without further examination, mostly because there are few easy-to-use diagnostic procedures available for routine use (Malik, Möhring, & Piepho, 2016). We have not considered this assumption in great detail, but the Supporting information for Example 1 demonstrates that the assumption can be conveniently tested in factorial experiments by assuming block effects are fixed factor effects and by including the blocks in a complete analysis of factorial effects. For a complete factorial design, it then becomes straightforward to compare the different factorial components of the block-by-treatment interaction effects. If all the block-by-treatment interaction components appear homogeneous, it can be assumed that a valid estimate of error variance can be obtained by pooling all the individual block-by-treatment interaction

components. If, however, the components show any evidence of heterogeneity, as seems to be the case in Example 1, further modelling and analysis may be needed to provide an adequate understanding of the data.

It is important to keep in mind that a data transformation affects several assumptions simultaneously. On the one hand, distributional assumptions for the random effects, such as normality and homogeneity of variance, need to be met. On the other hand, the systematic part of the model, represented by the fixed effects, should take a simple and easily interpretable form. This includes not only a need to meet the assumption of block-by-treatment additivity, but also a need to find a simple form of regression model. Ideally, the hope is that a suitable transformation will achieve all of these objectives simultaneously and sometimes there is, indeed, a natural scale on which all assumptions are likely to hold simultaneously (Piepho, 2009). The classical example is the exponential growth model where the variance increases proportional to the response Y and the regression model can be approximated by an exponential function of the form $Y = \exp(P + e) = \exp(P) \times \exp(e)$, where P is a polynomial in the quantitative predictor variables and e comprises all design effects, including error. Then a log transformation can both stabilize the variance and linearize the fitted response model simultaneously, as is the case in Examples 3 and 5. For counts, square root transformations often work well (McCullagh & Nelder, 1989), and for proportions, the logit transformation provides a natural scale (Warton & Hui, 2011). For a very accessible discussion of the topic, see Mead and Curnow (1983, p. 171, Section 8.5). The transformed model can be back-transformed to the original scale (including confidence intervals), but it needs to be kept in mind that the back-transformation yields estimates of medians rather than expected values (Piepho, 2009) and the standard errors cannot be so easily back-transformed (the delta method needs to be applied; Johnson et al., 2005). Where a transformation that meets all the required assumptions simultaneously cannot be found, more complex methods that relax one or several of the assumptions, such as weighted regression (Carroll & Ruppert, 1988) or GLMM, may need to be considered.

Polynomial models often provide a satisfactory representation of curvilinear relationships, and it should be emphasized that a polynomial model can be considered as an approximation to an unknown non-linear function based on a Taylor series expansion, the order of the approximating polynomial depending on the order of the Taylor approximation (Schabenberger & Pierce, 2002, p. 193). That is why polynomial models often provide a very good empirical fit for an arbitrary unknown quantitative model and is also the fundamental reason why all polynomial model terms up to the term of highest fitted degree must be included in a linear model approximation, in accordance with the marginality principle alluded to earlier. Otherwise, the approximation is not a Taylor series approximation and there can then be no particular reason to expect a good empirical model fit.

Care should be exercised in choosing the degree of the polynomial model. Lack-of-fit testing provides one mechanism to avoid over-fitting. Generally, polynomials higher in degree than quadratic or cubic should be avoided because they usually indicate an over-fitted or miss-specified model. If a lower-degree polynomial does not fit well, it may be preferable to consider an alternative approach. In particular, polynomials can fail when the relationship has an asymptote for very low or very high observed levels of a quantitative treatment factor. Polynomials cannot successfully capture such asymptotic responses because they tend to either plus or minus infinity as the quantitative level factor increases indefinitely. However, as shown in Example 3 Fig. 4, a log transformation of data with a zero asymptote can sometimes be successful in fitting a low-degree polynomial model to the transformed data. Example 5 has demonstrated that a transformation of the quantitative treatment levels can enhance the flexibility of a polynomial model.

For simplicity, we have restricted attention here to polynomial models, but the general principles, including lack-of-fit testing, can also be applied to intrinsically non-linear models with or without asymptote such as the sigmoidal growth curves (logistic, Gompertz, etc.) (Schabenberger & Pierce, 2002; Chapter 5; Ritz, Kniss, & Streibig, 2015). Also, one may apply a generalized linear model framework (Lee et al., 2006; McCullagh & Nelder, 1989), where inverse polynomials provide a particularly convenient way to model non-linear relationships, also when the treatment design entails qualitative factors. Further very flexible options are fractional polynomials (Sauerbrei, Meier-Hirmer, Benner, & Royston, 2006) or splines and other smoothing methods (Green & Silverman, 1994; Wood, 2017; Wood & Scheipl, 2017).

Most of the examples in this paper are basic, but the example on the analysis of repeated-measures data is more advanced. Repeated-measures data are common in agronomic research but are frequently misinterpreted and analysed inappropriately. The analysis in Example 4 shows how modern software can be used to fit an AR(1) correlation model for repeated observations assuming a fixed blocks effects model with arbitrary block effects at each time point. The gls() function of the R package "nlme" fits models to the repeated-measures data by the method of generalized least squares, which is equivalent to fitting an ordinary linear model but with a general covariance matrix for the correlated observations (see Rao, Toutenburg, Shalabah, & Heumann, 2008; Chapter 4). Pinheiro and Bates (2000, Chapter 5) discuss methods for the construction of extended linear mixed-effects models and models for serially correlated error structures. They recommend gls() for autoregressive models with a single error stratum where all model effects are assumed fixed, but for models with a hierarchical data structure or with more than a single error stratum, they recommend the lme() function of the "nlme" package. The use of lme() for repeated-measures data is complex and beyond the scope of this paper, and we refer the readership to Chapter 5 of Pinheiro and Bates (2000) and to Gałecki and Burzykowski (2013) for further discussion.

For the purposes of this paper, the examples have concentrated on randomized complete block designs or simple complete split-plot block designs. However, the generalization of the examples to include block designs with incomplete blocks is straightforward in most cases. The use of mixed model software means that, provided the incomplete blocks are specified as a random factor, the analysis

will automatically provide a properly weighted block analysis and properly weighted combined estimates of treatment effects (Mead et al., 2012; Section 9.2; Möhring, Piepho, & Williams, 2015). In case of repeated measures, it is necessary to allow for serial correlation among random incomplete block effects over time or indeed among any other random design effects (main plots, sub-plots, etc.) over time. Fitting serial correlation structures to multiple random effects is straightforward in most mixed model packages, but capabilities in R packages "nlme" and "lme4" are somewhat limited in this respect. The package "lme4" cannot fit serial correlation models except the unstructured model and, via simple random effects, the compound symmetry model, but it can do so for multiple random effects. The "nlme" package can fit several serial correlation models for the residual error, and it can also fit some such models for additional random effects using the pdMat construct (Bates, Mächler, Bolker, & Walker, 2015; Pinheiro & Bates, 2000, p. 157), although the syntax required to do so is somewhat complex. For R users, the commercial "ASReml-R" package (http://www.vsni.de/de/software/ASReml/) is an interesting alternative with great flexibility to fit a large number of covariance structures (Butler et al., 2009).

We have used the methods proposed by Kenward and Roger (1997, 2009) for approximating the denominator degrees of freedom. These methods are available in most mixed model packages. In R, one can use the packages "pbkrtest" (Halekoh & Højsgaard, 2014) and "lmerTest" (Kuznetsova, Brockhoff, & Christensen, 2017). The use of these methods can be controversial because *F*- and *t*-statistics in mixed models do not generally have an exact *F*- or *t*-distribution [although simulation evidence suggests that usually these approximations work quite well; see, e.g., Richter, Kroschewski, Piepho, and Spilke (2015) and references cited therein, as well as the two papers by Kenward and Roger]. For this reason, some statisticians advocate alternative simulation-based parametric bootstrap procedures (Bates et al., 2015).

11 | RESOURCES

Full data sets and all computer code (R, SAS) used to produce the results in this paper are provided in the Supporting information. The data and the R code used to run our analyses have been incorporated in an R package at https://CRAN.R-project.org/package=agriTutorial. After installing and loading the agriTutorial package, the start page can be accessed by typing help(agriTutorial) and further information can be accessed by typing vignette(''agriTutorial''). All figures appearing in the paper were generated using the SGPLOT procedure of SAS. In our R code for Example 3, we show how similar figures can be generated using the "lattice" and "ggplot2" packages (Kabacoff, 2015; Sarkar, 2008; Wickham, 2009). Good references for fitting of mixed models in R include Pinheiro and Bates (2000), Venables and Ripley (2002), Faraway (2006), Gałecki and Burzykowski (2013) and Bates et al. (2015). For mixed models in SAS, the reader is referred to Schabenberger and Pierce (2002), Littell, Milliken, Stroup, Wolfinger, and Schabenberger (2006) and Stroup (2015).

ACKNOWLEDGEMENTS

We are very grateful to Christel Richter (Humboldt Universität Berlin, Germany) and Andrea Onofri (University of Perugia, Italy) for careful reading of an earlier version of this paper. Three reviewers are thanked for their very constructive comments.

ORCID

H. P. Piepho https://orcid.org/0000-0001-7813-2992

REFERENCES

Agresti, A. (2010). *Analysis of ordinal categorical data*, 2nd ed. New York, NY: Wiley. https://doi.org/10.1002/9780470594001

Atkinson, A. C. (1987). *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*. Oxford, UK: Oxford University Press.

Bailey, R. A. (2008). *Design of comparative experiments*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511611483

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Box, G. E. P., & Draper, N. P. (2007). *Response surfaces, mixtures, and ridge analyses*. New York, NY: Wiley. https://doi.org/10.1002/0470072768

Bretz, F., Hothorn, T., & Westfall, P. (2011). *Multiple comparisons using R*. Boca Raton, FL: CRC Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*, 2nd ed. New York, NY: Springer.

Butler, D., Cullis, B., Gilmour, A., & Gogel, B. J. (2009). *ASReml-R reference manual, release 3*. Technical Report, Queensland Department of Primary Industries.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York, NY: Chapman & Hall. https://doi.org/10.1007/978-1-4899-2873-3

Dean, A., & Voss, D. (1999). *Design and analysis of experiments*. New York, NY: Springer. https://doi.org/10.1007/b97673

Dixon, P. (2017). *Are blocks fixed or random? Annual Conference on Applied Statistics in Agriculture*. Retrieved from http://newprairiepress.org/agstatconference/2017/ (to appear)

Draper, N. R., & Smith, H. (1998). *Applied linear regression analysis*, 3rd ed. New York, NY: Wiley.

Edmondson, R. N. (1991). Agricultural response surface experiments based on four-level factorial designs. *Biometrics*, *47*, 1435–1448. https://doi.org/10.2307/2532397

Edmondson, R. N., & Piepho, H. P. (2018). *agriTutorial: Tutorial analysis of some agricultural experiments. R package version 0.1.0*. Retrieved from https://CRAN.R-project.org/package=agriTutorial

Faraway, J. J. (2006). *Extending the linear model with R. Generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: Chapman and Hall/CRC.

Federer, W. T., & Crossa, J. (2005). Designing for and analyzing results from field experiments. *Journal of Crop Improvement*, *14*, 29–50. https://doi.org/10.1300/J411v14n01_04

Gałecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R. A step-by-step approach*. Berlin, Germany: Springer.

Gertheiss, J. (2014). ANOVA for factors with ordered levels. *Journal of Agricultural, Biological and Environmental Statistics*, *19*, 258–277. https://doi.org/10.1007/s13253-014-0170-5

Giesbrecht, F. G., & Gumpertz, M. L. (2004). *Planning, construction and analysis of comparative experiments*. New York, NY: Wiley. https://doi.org/10.1002/0471476471

Gomez, K. A., & Gomez, A. A. (1984). *Statistical procedures for agricultural research*, 2nd ed. New York, NY: Wiley.

Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models A roughness penalty approach*. London: Chapman & Hall.

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – The R package pbkrtest. *Journal of Statistical Software*, *59*(9), 1–30.

Hinkelmann, K., & Kempthorne, O. (1994). *Design and analysis of experiments. Volume I: Introduction to experimental design*. New York, NY: Wiley.

Hocking, R. R. (1985). *The analysis of linear models*. Monterey, CA: Brooks/Cole.

Hsu, J. C. (1996). *Multiple comparisons. Theory and methods*. London, UK: Chapman & Hall. https://doi.org/10.1007/978-1-4899-7180-7

John, J. A., & Williams, E. R. (1995). *Cyclic and computer generated designs*, 2nd ed. London, UK: Chapman & Hall. https://doi.org/10.1007/978-1-4899-7220-0

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions*, 3rd ed. New York, NY: Wiley. https://doi.org/10.1002/0471715816

Kabacoff, R. I. (2015). *R in action: Data analysis and graphics with R*, 2nd ed. Greenwich, UK: Manning Publications.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997. https://doi.org/10.2307/2533558

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis*, *53*, 2583–2595. https://doi.org/10.1016/j.csda.2008.12.013

Kozak, M., & Piepho, H. P. (2018). What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. *Journal of Agronomy and Crop Science*, *204*, 86–98.

Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis*, 2nd ed. Pacific Grove, CA: Duxbury Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). *lmerTest: Tests in linear mixed effects models*. R package version 2.0-34. Retrieved from https://CRAN.R-project.org/package=lmerTest.

Lane, P. W., & Nelder, J. A. (1982). Analysis of covariance and standardisation as instances of prediction. *Biometrics*, *82*, 613–621. https://doi.org/10.2307/2530043

Lee, Y., Nelder, J. A., & Pawitan, Y. (2006). *Generalized linear models with random effects: A unified analysis via h-likelihood*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/CHMONSTAAPP

Lenth, R. V. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.1.. https://CRAN.R-project.org/package=emmeans

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models*, 2nd ed. Cary, NC: SAS Institute Inc.

Malik, W. A., Möhring, J., & Piepho, H. P. (2016). A clustering-based test for nonadditivity in an unreplicated two-way layout. *Communications in Statistics – Simulation and Computation*, *45*, 660–670. https://doi.org/10.1080/03610918.2013.870196

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*, 2nd ed. London, UK: Chapman & Hall. https://doi.org/10.1007/978-1-4899-3242-6

Mead, R. (1988). *The design of experiments. Statistical principles for practical application*. Cambridge, UK: Cambridge University Press.

Mead, R., & Curnow, R. N. (1983). *Statistical methods in agriculture and experimental biology*. London, UK: Chapman & Hall. https://doi.org/10.1007/978-1-4899-2951-8

Mead, R., Gilmour, S. G., & Mead, A. (2012). *Statistical principles for the design of experiments*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139020879

Milliken, G. A., & Johnson, D. E. (1992). *Analysis of messy data. Volume I: Designed experiments*. Boca Raton, FL: CRC Press.

Milliken, G. A., & Johnson, D. E. (2002). *Analysis of messy data. Volume III: Analysis of covariance*. Boca Raton, FL: CRC Press.

Möhring, J., Piepho, H. P., & Williams, E. R. (2015). Inter-block information: To recover or not to recover it? *Theoretical and Applied Genetics*, *128*, 1541–1554. https://doi.org/10.1007/s00122-015-2530-0

Nelder, J. A. (1954). The interpretation of negative components of variance. *Biometrika*, *41*, 544–548. https://doi.org/10.1093/biomet/41.3-4.544

Nelder, J. A. (1994). The statistics of linear models: Back to basics. *Statistics and Computing*, *4*, 221–234. https://doi.org/10.1007/BF00156745

Nelder, J. A. (2000). Functional marginality and response surface fitting. *Journal of Applied Statistics*, *27*, 109–112. https://doi.org/10.1080/02664760021862

Pearce, S. C. (2005). The factorial field experiment. *Experimental Agriculture*, *41*, 109–120. https://doi.org/10.1017/S0014479704002364

Peixoto, J. L. (1990). A property of well-formulated polynomial regression models. *The American Statistician*, *44*, 26–30.

Petersen, R. G. (1994). *Agricultural field experiments. Design and analysis*. New York, NY: Marcel Dekker.

Piepho, H. P. (2009). Data transformation in statistical analysis of field trials with changing treatment variance. *Agronomy Journal*, *101*, 865–869. https://doi.org/10.2134/agronj2008.0226x

Piepho, H. P., Büchse, A., & Emrich, K. (2003). A hitchhiker's guide to the mixed model analysis of randomized experiments. *Journal of Agronomy and Crop Science*, *189*, 310–322. https://doi.org/10.1046/j.1439-037X.2003.00049.x

Piepho, H. P., Büchse, A., & Richter, C. (2004). A mixed modelling approach to randomized experiments with repeated measures. *Journal of Agronomy and Crop Science*, *190*, 230–247. https://doi.org/10.1111/j.1439-037X.2004.00097.x

Piepho, H. P., Herndl, M., Pötsch, E., & Bahn, M. (2017). Designing an experiment with quantitative treatment factors to study the effects of climate change. *Journal of Agronomy and Crop Science*, *203*, 584–592. https://doi.org/10.1111/jac.12225

Piepho, H. P., & Spilke, J. (1999). Anmerkungen zur Analyse balancierter gemischter Modelle mit der SAS Prozedur MIXED. *Zeitschrift für Agrarinformatik*, *7*, 39–46.

Piepho, H. P., Williams, E. R., & Fleck, M. (2006). A note on the analysis of designed experiments with complex treatment structure. *HortScience*, *41*, 446–452.

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

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-0318-1

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2017). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-131. Retrieved from https://CRAN.R-project.org/ package=nlme.

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511806384

Rao, C. R., Toutenburg, H., Shalabah, , & Heumann, H. (2008). *Linear models. Least squares and alternatives*, 3rd ed. New York, NY: Springer.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Richter, C., Kroschewski, B., Piepho, H. P., & Spilke, J. (2015). Treatment comparisons in agricultural field trials accounting for spatial correlation. *Journal of Agricultural Science*, *153*, 1187–1207. https://doi.org/10.1017/S0021859614000823

Ritz, C., Kniss, A. R., & Streibig, J. C. (2015). Research methods in weed science: Statistics. *Weed Science*, *63*, 166–187. https://doi.org/10.1614/WS-D-13-00159.1

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-75969-2

SAS Institute Inc. (1999). *SAS/STAT user's guide*, Version 8. Cary, NC: SAS Institute Inc.

Sauerbrei, W., Meier-Hirmer, C., Benner, A., & Royston, P. (2006). Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Computational Statistics and Data Analysis*, *50*, 3464–3485. https://doi.org/10.1016/j.csda.2005.07.015

Schabenberger, O., & Pierce, F. J. (2002). *Contemporary statistical models for the plant and soil sciences*. Boca Raton, FL: CRC Press.

Searle, S. R. (1987). *Linear models for unbalanced data*. New York, NY: Wiley.

Searle, S. R. (1994). Comments on J.A. Nelder 'The statistics of linear models: Back to basics'. *Statistics and Computing*, *4*, 103–107.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York, NY: Wiley.

Shah, D. A., & Madden, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology*, *94*, 33–43. https://doi.org/10.1094/PHYTO.2004.94.1.33

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods*, 8th ed. Ames, IA: Iowa State Univ. Press.

Spilke, J., Hu, X., & Piepho, H. P. (2005). A simulation study on tests of hypotheses for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological and Environmental Statistics*, *10*, 374–389. https://doi.org/10.1198/108571105X58199

Stroup, W. W. (2015). *Generalized linear mixed models. Modern concepts, methods and applications*. Boca Raton, FL: Chapman & Hall/CRC.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*, 4th ed. New York, NY: Springer. https://doi.org/10.1007/978-0-387-21706-2

Verdooren, L. R. (1982). How large is the probability for the estimate of a variance component to be negative? *Biometical Journal*, *24*, 339–360. https://doi.org/10.1002/(ISSN)1521-4036

Warton, D., & Hui, F. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, *92*, 3–10. https://doi.org/10.1890/10-0340.1

Welham, S., Cullis, B., Gogel, B., Gilmour, A., & Thompson, R. (2004). Prediction in linear mixed models. *Australian and New Zealand Journal of Statistics*, *46*, 325–347. https://doi.org/10.1111/j.1467-842X.2004.00334.x

Welham, S. J., Gezan, S. A., Clark, S. J., & Mead, A. (2015). *Statistical methods in biology*. Boca Raton, FL: CRC Press.

West, B. T., Welch, K. B., & Gałecki, A. T. (2014). *Linear mixed models. A practical guide using statistical software*, 2nd ed. New York, NY: CRC Press.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag. https://doi.org/10.1007/978-0-387-98141-3

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, *22*, 392–399. https://doi.org/10.2307/2346786

Wood, S. N. (2017). *Generalized additive models: An introduction with R*, 2nd ed. Boca Raton, FL: CRC Press.

Wood, S. N., & Scheipl, F. (2017). *gamm4: Generalized additive mixed models using 'mgcv' and 'lme4'*. R package version 0.2-5. Retrieved from https://CRAN.R-project.org/package=gamm4.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Piepho HP, Edmondson RN. A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. *J Agro Crop Sci*. 2018;204:429–455. https://doi.org/10.1111/jac.12267

APPENDIX 1

CODING THE MODEL FACTORS FOR EXAMPLE 1

To code the models for this example, variables are needed for the factors blocks, main plots, split plots, varieties and nitrogen fertilizer. First consider the model for treatment effects (treatment model). Care is needed in the choice of the variable for the factor nitrogen. This factor is actually needed for two distinct purposes: (i) to fit the polynomial terms, N must be declared as a quantitative factor; (ii) for specifying the lack-of-fit effects δ_i and δ_{ij}, N must be declared as a qualitative factor and crossed with variety. As with the single-factor Example 2, this can be achieved by copying the quantitative variable for nitrogen (N) into another variable (LOF_N), which is then declared as qualitative. Adding a lack-of-fit effect, the right-hand side of model (15) is coded as,

$$\text{REP} + \text{MGMT} + \text{VAR} + \text{N} + \text{N} \cdot \text{N} + \text{LOF\_N} + \text{VAR} \cdot \text{N} + \text{VAR} \cdot \text{N} \cdot \text{N} + \text{VAR} \cdot \text{LOF\_N},$$

where the dot (·) indicates crossing of two factors (Piepho et al., 2003), REP is a factor for replicates (complete blocks), MGMT is a factor for management practices, VAR is the variable identifying the four varieties, and the terms can be equated as follows: $b_k \equiv$ REP, $\theta_h \equiv$ MGMT, $\beta_{0j} \equiv$ VAR, $\gamma_1 x_i \equiv$ N, $\gamma_2 x_i^2 \equiv$ N·N, and $\delta_i \equiv$ LOF_N, $\alpha_{1j} x_i \equiv$ VAR·N, $\alpha_{2j} x_i^2 \equiv$ VAR·N·N, and $\delta_{ij} \equiv$ VAR·LOF_N.

To code the model for design effects (block model) representing the field layout, we need variables uniquely identifying all block units, i.e., complete replicates (REP; three levels), main plots (MAINPLOT; 15 levels) and split plots (SPLITPLOT; 45 levels). Random effects need to be fitted for each of these design factors to represent the randomization layout. In our view, it is good policy to use entirely different sets of variables and factors to code the treatment model and the block model (Piepho et al., 2003; Wilkinson & Rogers, 1973). Some authors use a combination of block and treatment factors to code effects for, e.g., main plots and split plots, but we think this is potentially confusing and misleading (sometimes it even fails to produce the correct analysis, e.g., when there is extra replication of a control level of a main-plot factor) and we would recommend against it. For example, the main-plot error could be coded by N·BLOCK, provided N is declared as a factor, but this effect looks like an interaction even though it is not. It is therefore much clearer to just represent the effect by the single-factor MAINPLOT with 15 unique levels.

When fitting polynomial models, occasionally numerical difficulties occur when the levels of the quantitative treatment factor are large in absolute value. It may then help to standardize the levels, e.g., by subtracting the mean or a value close to the mean. In the present case, the Kenward–Roger method failed when testing the quadratic model for lack of fit. Subtracting 80 from the x_i values solved the problem.

APPENDIX 2

REPEATED-MEASURES ANALYSIS FOR EXAMPLE 4 USING R

The generalized least squares gls() function of the "nlme" package of R (Pinheiro, Bates, DebRoy, & Sarkar, 2017) can be used to fit repeated-measures data assuming a fixed block effects model. Users new to this package frequently try to fit models having only one random error term using the lme() function instead, but this does not work because the lme() function requires at least one additional random effects term to be fitted. However, unlike SAS, the gls() function will not fit a factor both as a qualitative factor and as a set of polynomial effects in the same model. For example, a model that includes both factorial block-by-week effects and polynomial variety-by-week effects cannot be fitted by gls(). Instead, both sets of effects have to be modelled explicitly by using polynomial terms. The R code to fit week-by-treatment and block-by-week polynomial effects in the same model for the repeated-measures analysis using the gls() function is shown in the examples in the R package agriTutorial.

Initially, we used corCAR1() to fit the model correlation structure but subsequently found that corExp() was more flexible as it allowed the inclusion of a "nugget" term, which seemingly cannot be done with corCAR1(). We note that it is essential that the variable weeks in the specification of the correlation structure, `corr = corExp (form = ~weeks|Plots)`, is a numeric variable with values representing the time values of the repeated measurements. Note that this code will also work for repeated observations that are not necessarily equally spaced. It is also necessary that each plot has a unique level for the variable `Plots`. Each term in this model is a single term in the corresponding analysis of variance, and Table A1 shows the analysis of variance for this model.

Table A2 shows the goodness-of-fit tests for a fitted model assuming a quadratic model for both the treatment-by-week interaction and the block-by-week interaction. The method used to code the model in R is quite general, and any feasible polynomial models, not necessarily of the same degree, can be fitted for the treatment-by-week model and the block-by-week model separately. Once a suitable treatment model has been selected, the coefficients of the model must be estimated by fitting the required model terms alone and without the lack-of-fit terms. Models of different degree can be fitted for blocks and varieties separately provided the same formulation is used for all polynomial terms common to both models.

For significance testing and model selection, orthogonal polynomials are more stable than raw polynomials and give the same results. For model fitting, however, orthogonal and raw polynomials give different model parameter estimates and, where feasible, raw polynomials will normally be preferred as they have the simplest and most direct interpretation. Here the maximum polynomial power is four in Example 4; therefore, raw polynomials should be stable and will be the polynomials of choice for this analysis. For designs with larger polynomial powers, however, raw polynomials may be unstable and then orthogonal polynomials will need to be used both for significance testing and model selection and for model fitting. Although orthogonal polynomials are not difficult to understand, they are beyond the scope of this paper and the interested reader is referred to Draper and Smith (1998) for further discussion.

TABLE A1 Sequential Wald-type *F* tests for quadratic model fitted to the sorghum data (Example 4). Analysis obtained with the gls() function of the "nlme" package in R using the exponential model, which is equivalent to AR(1) for equally spaced data with $\rho = \exp(-1/\theta)$. $\hat{\rho} = 0.908$[a], $\hat{\sigma}^2 = 0.00383$, $\hat{\sigma}^2_N = 0.01867$[b], $-2 \log L_R = -131.509$[c]

	Numerator degrees of freedom	Denominator degrees of freedom[d]	*F*-value	*p*-value
(Intercept)	1	60	22278.12	<.0001
P (full polynomial)	4	60	447.08	<.0001
Blocks	4	60	97.68	<.0001
Varieties	3	60	88.74	<.0001
P × blocks	16	60	4.22	<.0005
P × varieties	12	60	12.58	<.0001

[a]$\hat{\theta} = 10.36$, from which $\hat{\rho} = \exp\left(-1/\hat{\theta}\right) = 0.908$.
[b]The "nlme" package reports a "nugget" $\hat{c}_0 = 0.1702$ and a "residual standard error" $\hat{\phi} = 0.1500$, from which we may obtain $\hat{\sigma}^2 = (1 - \hat{c}_0)\hat{\phi}^2 = 0.00383$ and $\hat{\sigma}^2_N = \hat{c}_0\hat{\phi}^2 = 0.01867$, which agrees closely with the output from SAS (Table A3).
[c]The log-likelihood does not agree with that obtained with SAS, the difference being due to different parameterizations of the polynomial terms, giving rise to different values of the term X^TX in the log-likelihood, where X is the design matrix for the fixed effects.
[d]Residual degrees of freedom = number of observations minus number of fixed effects fitted.

Comparison of SAS versus R for Example 4

The sequential *F* tests for equivalent terms in Table A2 do not agree exactly with those in Table 14, which was obtained with the MIXED procedure of SAS. Note that in Table 14 we used the second-order method of Kenward–Roger, which not only involves a different approximation of the denominator degrees of freedom but also a correction to the *F*-statistic as described in Kenward and Roger (2009). By contrast, in Table A2 we used the residual degrees of freedom as denominator degrees of freedom in all *F* tests and the computation of the *F*-statistics did not involve the Kenward–Roger adjustment. The residual degrees of freedom are defined as the total number of observations minus the number of fitted fixed-effects parameters. Reproducing this analysis without Kenward–Roger adjustment in SAS, we still did not obtain perfect agreement (Table A3). Only the tests of effects involving the lack-of-fit factor agreed perfectly. This is not a numerical problem, because the estimates of the autocorrelation ρ, the spatial residual variance σ^2 and the "nugget" variance σ^2_N agree very well (Tables A1, A2 and A3), although "nlme" uses a different

TABLE A2 Sequential Wald-type *F* tests for quadratic model fitted to the sorghum data (Example 4). Analysis obtained with the gls() function of the "nlme" package in R using the exponential model, which is equivalent to AR(1) for equally spaced data with $\rho = \exp(-1/\theta)$. $\hat{\rho} = 0.908$[a], $\hat{\sigma}^2 = 0.00383$, $\hat{\sigma}^2_N = 0.01867$[b], $-2 \log L_R = -131.509$[c]

Source	Numerator degrees of freedom	Denominator degrees of freedom[d]	*F*-value	*p*-value
Linear week	1	60	1687.51	<.0001
Quadratic week	1	60	62.68	<.0001
Lack of fit (LoF2)	2	60	19.07	<.0001
Blocks	4	60	97.68	<.0001
Blocks × linear week	4	60	12.449	<.0001
Blocks × quadratic week	4	60	1.694	.1632
Blocks × lack of fit (LoF2)	8	60	1.368	.2292
Varieties	3	60	88.74	<.0001
Varieties × linear week	3	60	21.19	<.0001
Varieties × quadratic week	3	60	16.62	<.0001
Varieties × lack of fit (LoF2)	6	60	6.25	<.0001

[a]$\hat{\theta} = 10.36$, from which $\hat{\rho} = \exp\left(-1/\hat{\theta}\right) = 0.908$.
[b]The "nlme" package reports a "nugget" $\hat{c}_0 = 0.1702$ and a "residual standard error" $\hat{\phi} = 0.1500$, from which we may obtain $\hat{\sigma}^2 = (1 - \hat{c}_0)\hat{\phi}^2 = 0.00383$ and $\hat{\sigma}^2_N = \hat{c}_0\hat{\phi}^2 = 0.01867$, which agrees closely with the output from SAS (Table A3).
[c]The log-likelihood does not agree with that obtained with SAS, the difference being due to different parameterizations of the polynomial terms, giving rise to different values of the term X^TX in the log-likelihood, where X is the design matrix for the fixed effects.
[d]Residual degrees of freedom = number of observations minus number of fixed effects fitted.

TABLE A3 Sequential Wald-type *F* tests for quadratic model (29) fitted to the sorghum data (Example 4). Analysis obtained using the MIXED procedure of SAS using the option DDFM = residual. $\hat{\rho} = 0.9080$, $\hat{\sigma}^2 = 0.01862$, $\hat{\sigma}^2_N = 0.00387$, $-2 \log L_R = -42.98$

Source	Numerator degrees of freedom	Denominator degrees of freedom[a]	*F*-value	*p*-value
Linear week	1	60	1705.28	<.0001
Quadratic week	1	60	59.72	<.0001
Lack of fit	2	60	19.06	<.0001
Blocks	4	60	98.80	<.0001
Blocks × weeks	16	60	4.22	.0005
Variety	3	60	92.20	<.0001
Variety × linear week	3	60	21.03	<.0001
Variety × quadratic week	3	60	16.49	<.0001
Variety × lack of fit	6	60	6.25	<.0001

[a]Residual degrees of freedom = number of observations minus number of fixed effects fitted.

parameterization (Tables A1 and A2). Also, the weighted (generalized) least squares estimates of fixed effects agree perfectly when re-coding factors in SAS so that the first factor level becomes the last in alpha-numerical order, thus yielding identical parameterizations with both packages (results not shown). The difference arises because slightly different strategies are used to compute sequential *F* tests in SAS and in the R packages "nlme" and "lme4." Sequential *F*-statistics in these two R packages (as in many other packages such as GenStat and ASReml) are based on reductions in weighted residual sums of squares, where the weighting is done using the inverse of the fitted variance–covariance matrix of the data (Rao et al., 2008; Chapter 4; Bates et al., 2015). By contrast, SAS procedures for mixed models construct coefficient matrices for Type I tests (these are also denoted as *sequential* in the online documentation) using the same method that is employed for fixed-effects models using reductions in ordinary (unweighted) residual sums of squares (Type I sums of squares). This method operates only on the design matrix for fixed effects, whereas tests based on the sequential reduction of the weighted residual sum of squares also operate on the fitted variance–covariance matrix for the data. Technical details of this discrepancy are discussed in Appendix 3. It is important to note that with certain models and data structures, *sequential* or Type I *F* tests produced by alternative packages do not necessarily agree (although they usually agree in simple cases such as those in the other examples considered in this paper), although both provide valid tests. The most important fact to be remembered with either sequential method is that only the last term added in the sequence is guaranteed to be tested with an adjustment for all other terms in the model. It is noted for completeness that the R package "lmerTest" (Kuznetsova et al., 2017) can produce SAS-like Type III tests based on output from the function lmer() of the package "lme4."

Note that we fitted blocks after the main effects for time in Table A3 to obtain an *F* test for blocks nested within weeks. The log-likelihoods in Tables A2 and A3 do not agree because of the different parameterizations for the block-by-week effects (polynomials in Table A2, simple effects in Table A3), causing a difference in the design matrix for fixed effects and hence in the residual log-likelihood.

APPENDIX 3

WHY DO SAS AND R SOMETIMES SHOW DIFFERENT SEQUENTIAL *F*-STATISTICS?

Throughout the paper, we are presenting the output generated using the MIXED procedure of SAS. Model fits agree with those obtained using R for all examples without exception, as do the variance component estimates. *F*-statistics from both packages usually agree for all our examples, with minor differences in *p*-values that are mainly due to differences in the denominator degrees of freedom approximations used, as can be verified by running the SAS and R code we are presenting as Supporting information. There is one notable exception. With Example 4, several of the *F*-statistics differ, as we have detailed in Appendix 2. We here sketch a brief explanation for the cause of these differences.

In matrix form, the expected value of the data vector y in a mixed model can be written as $E(y) = X\beta$, where β is a vector of fixed-effects parameters and X is a corresponding design matrix. The variance–covariance matrix, which depends on the assumed variance–covariance structures for the random effects, can be written as $\text{var}(y) = V$. The parameters in V are estimated by REML. With the REML estimate $\hat{V}$, the weighted least squares estimator of β is given by $\hat{\beta} = \left(X^T\hat{V}^{-1}X\right)^{-}X^T\hat{V}^{-1}y$, where M^T, M^{-1} and M^{-} denote the transpose, the inverse and a generalized inverse of the matrix M, respectively. The estimator $\hat{\beta}$ minimizes the weighted residual sum of squares $\text{RSS} = \left(y - X\hat{\beta}\right)^T\hat{V}^{-1}\left(y - X\hat{\beta}\right)$.

For sequential *F* tests, consider a partitioning of the linear model as $E(y) = X_1\beta_1 + X_2\beta_2 + \ldots$, where β_i ($i = 1,\ldots, p$) are the successively fitted (possibly vector-valued) effects and X_i are the corresponding design matrices. Obviously $\beta = \left(\beta_1^T, \beta_2^T, \ldots, \beta_p^T\right)^T$ and $X = (X_1|X_2|\ldots|X_p)$.

TYPE I TESTS IN SAS

In SAS procedures for mixed models (MIXED, HPMIXED, GLIMMIX), Type I *F* tests of a null hypothesis of the form,

$$H_0 : K\beta = 0, \tag{A1}$$

are computed as,

$$F = \frac{\hat{\beta}^T K^T\left[K\left(X^T\hat{V}^{-1}X\right)^{-}K^T\right]^{-}K\hat{\beta}}{\text{rank}(K)}, \tag{A2}$$

where rank(K) is the rank of the coefficient matrix K and represents the numerator degrees of freedom (SAS Institute, 1999, Chapter 41, p. 2165). Coefficient matrices K for Type I tests are constructed from a Cholesky decomposition of the matrix $X^TX = L^TL$, where L is an upper triangular Cholesky square root of X^TX (McCullagh & Nelder, 1989, p. 86). Skipping over zero rows, the non-zero rows of L associated with X_i provide a matrix K for testing a Type I hypothesis pertaining to β_i (SAS Institute Inc, 1999, Chapter 12, p. 166). Due to the construction of the hypotheses from a Cholesky decomposition of X^TX, the resulting test for β_1 is not adjusted for effects β_2 to β_p. The test for β_2 is adjusted for β_1, but not for β_3 to β_p, and so forth. Only the test for the last term β_p will be adjusted for all other terms (β_1 to β_{p-1}).

SEQUENTIAL SUMS OF SQUARES AND TESTS IN R PACKAGES

One way to compare the Type I tests of SAS with sequential *F* tests in R packages for mixed models (nlme, lme4) is to think of the coefficient K matrices in (A6) as being derived from a decomposition of $X^T\hat{V}^{-1}X$ instead of X^TX. Denote the decomposition as $X^T\hat{V}^{-1}X = R^TR$, where R is an upper triangular matrix. Skipping over zero rows in R, the partition of rows corresponding to X_i can be denoted as R_i. As will be shown here, setting in $K = R_i$ in (A2), we obtain the sequential *F* tests for β_i in R packages "nlme" and "lme4." Thus, the difference between SAS and R is essentially this: in SAS, the Type I tests are obtained by generating K from a Cholesky decomposition of X^TX, while sequential *F* tests in R can be obtained from a Cholesky decomposition of $X^T\hat{V}^{-1}X$.

To support the claim just made, we make use of the fact that the Moore–Penrose inverse of $X^T\hat{V}^{-1}X$ can be written as $\left(X^T\hat{V}^{-1}X\right)^{-} = R^T\left(RR^T\right)^{-2}R$ (Searle, 1987, p. 294). Thus, we have $R\left(X^T\hat{V}^{-1}X\right)^{-}R^T = RR^T\left(RR^T\right)^{-2}RR^T = I$ and hence $R_i\left(X^T\hat{V}^{-1}X\right)^{-}R_i^T = R_iR^T\left(RR^T\right)^{-2}RR_i^T = I_{\text{rank}(R_i)}$. Then setting $K = R_i$, we obtain,

$$F = \frac{\hat{\beta}^TR_i^T\left[R_i\left(X^T\hat{V}^{-1}X\right)^{-}R_i^T\right]^{-}R_i\hat{\beta}}{\text{rank}(R_i)} = \frac{\hat{\beta}^TR_i^TR_i\hat{\beta}}{\text{rank}(R_i)}. \tag{A3}$$

This is exactly the form of the *F*-statistic computed in "lme4" (and "nlme") for the effect β_i (Bates et al., 2015).

It is also helpful to observe that the total model weighted sum of squares, given by,

$$y^T\hat{V}^{-1}X\left(X^T\hat{V}^{-1}X\right)^{-}X^T\hat{V}^{-1}y = \hat{\beta}^TX^T\hat{V}^{-1}X\hat{\beta}, \tag{A4}$$

can be partitioned as,

$$\hat{\beta}^TX^T\hat{V}^{-1}X\hat{\beta} = \hat{\beta}^TR^TR\hat{\beta} = \sum_{i=1}^{p}\hat{\beta}^TR_i^TR_i\hat{\beta}, \tag{A5}$$

where (up to a possible scaling by the residual variance) $SS_i = \hat{\beta}^TR_i^TR_i\hat{\beta}$ is the weighted sum of squares reported for β_i in ANOVA tables by the R packages "nlme" and "lme4" [Bates et al., 2015; Equations (54) and (67)]. These sums of squares are identical to the sequential reductions in weighted residual sums of squares due to sequentially adding terms β_i and fixing the variance–covariance matrix V at the REML estimate obtained for the full model with all terms fitted. Other mixed model packages such as GenStat and ASReml compute *F*-statistics explicitly deriving these sequential reductions of sums of squares by sweeping the mixed model equations for successive terms. Many text books describe these

operations for the method of ordinary least squares, which operate on X^TX, e.g., McCullagh and Nelder (1989). For mixed models, essentially the same machinery can be applied replacing X^TX with $X^T\hat{V}^{-1}X$ and extending the equations for random effects.

FINAL REMARK

As was shown in this appendix, *F*-statistics reported by SAS and R are computed somewhat differently. It should be stressed, however, that with many balanced designs and simple random effects models, both types of *F*-statistics agree essentially because ordinary and weighted least squares estimates of β agree and tested hypotheses involve a single error stratum (Hocking, 1985, Chapters 9 and 10), as was the case with Example 1. Differences are likely to occur with unbalanced data and when non-linear variance–covariance structures are fitted. Differences were found for Example 4 as explained in Appendix 2. A further difference was found for the general lack-of-fit tests for the first-order and second-order models in Example 3, because these tests simultaneously involve the main-plot and subplot error strata.