Journal of Agronomy and Crop Science

J Agro Crop Sci (2013) ISSN 0931-2250

MISCELLANEOUS

Why Randomize Agricultural Experiments?

H. P. Piepho[1], J. Möhring[1] & E. R. Williams[2]

1 Bioinformatics Unit, Institute of Crop Science, University of Hohenheim Stuttgart, Germany
2 Statistical Consulting Unit, The Australian National University Canberra, ACT, Australia

Keywords
experimental design; field trial; linear model; randomization; statistics; uniformity trial

Correspondence
H. P. Piepho
Bioinformatics Unit
Institute of Crop Science
University of Hohenheim
Fruwirthstrasse 23
70599 Stuttgart
Germany
Tel.: +49 711 459 22386
Fax: +94 711 459 24345
Email: piepho@uni-hohenheim.de

Accepted March 1, 2013

doi:10.1111/jac.12026

Abstract

This study illustrates the importance of randomization using two hypothetical field trials, one with a marked systematic trend and the other with a more erratic spatial pattern. The insights from these two examples are reinforced by analysis of a uniformity trial and a small simulation study. Results illustrate that both model-based spatial analysis and randomization-based analysis assuming independent errors are valid with full randomization but may be invalidated when randomization is lacking. It is concluded that randomization provides protection against different forms of spatial trend. The examples given in the study serve as a general reminder that agricultural experiments should be randomized whenever possible.

Introduction

The purpose of a scientific experiment usually is to enable the researcher to generalize the results to similar situations in the future, not only to obtain some numerical results on measurable traits as such. With this purpose in mind, a single experiment should be regarded as merely a single sample from a potentially large population of similar experiments that could be or could have been conducted with the same experimental units. Statistical methods of analysis allow valid statistical inferences that can be generalized beyond the observed sample or experiment, provided the way that the study was designed meets some basic statistical requirements. In designed experiments, the requirement boils down to proper randomization. Randomization is particularly important in experiments under stress conditions, where spatial trends and heterogeneity in yielding ability tend to be rather pronounced (Haase et al. 2007, Leiser et al. 2012, Mühleisen et al. 2013). As will be argued in this study, without randomization, valid statistical inference is difficult if not impossible to achieve in the presence of spatial trend and heterogeneity, and biases in estimates of treatment effects are inevitable.

While the role of randomization as a basic principle in empirical research has been well understood since the work of Fisher (1925), it appears that its importance is not always fully appreciated. In our experience from reviewing articles and statistical consultation, a considerable number of agricultural experiments lack proper randomization. This study uses several examples to demonstrate that randomization is a crucial prerequisite for obtaining valid statistical inferences. To exemplify the problem and illustrate key ideas, we start with a small hypothetical variety trial comprising five varieties on 20 plots. We then present results based on a uniformity trial and a small simulation study. In both, we compare a systematic design with a completely randomized design. The benefit of spatial analysis is also explored.

Our article is essentially a reminder of well-known facts and results, which have been reviewed and emphasized by many statisticians. A study that is similar in style and spirit to ours, but considers a different application, is Greenberg (1951). We acknowledge that our exposition may appear somewhat cursory to the statistically trained. This style is intentional, and we hope that it is appealing to the intended audience. A much broader and mathematically

rigorous treatment of the subject that is not restricted to agricultural experiments can be found, for example, in Kempthorne (1977). We feel that it is valuable to reinforce important randomization principles in an agricultural context and hope that our article will be of some value for those conducting agricultural experiments and wanting to refresh their knowledge.

Three Illustrative Examples

A small hypothetical variety trial

To illustrate the importance of randomization, assume that a variety trial is to be laid out on 20 field plots arranged on a grid of five rows by four columns. Further assume that we want to test five different varieties (A–E) of wheat on four plots each. By some accident, however, all seed bags contain only seed of variety A, but nevertheless bags used for the trial are inadvertently labelled A–E. Thus, what we think are different varieties in the experiment are, in fact, copies of one and the same variety, that is, variety A is tested on all 20 plots. Figure 1 shows a hypothetical outcome of this kind of experiment. The data display a systematic spatial trend with yields decreasing from the lower left to the upper right corner. This trend could be due to heterogeneity in soil type and water holding capacity, which in turn affects yields. So for example, the plot in the lower left corner would have the largest observed yield because the soil on that plot has the highest water-holding capacity.

Note that this hypothetical experiment corresponds to what is known as a uniformity trial. Such trials, in which the same treatment is applied on each plot, are often conducted to compare alternative experimental designs and plot sizes (Smith 1938, Williams and Luckett 1988). Now, consider the allocation of variety labels to plots in a randomized experiment. The simplest experimental design is the completely randomized design, in which variety labels are allocated to plots completely at random. In this study, we focus entirely on the completely randomized design for simplicity. But it should be stressed that good experimental design usually involves some form of blocking (complete or incomplete blocks, one-way blocking or row-column designs) for efficient error control (Kempthorne 1977, John and Williams 1995). While our focus is on the completely randomized design, our conclusions regarding the merit of randomization apply equally to any form of randomized block design.

7.6	7.4	7.2	7.0
8.1	7.6	7.3	7.0
8.3	8.1	7.9	7.3
8.8	8.5	8.0	7.7
9.0	8.6	8.4	8.0

Fig. 1 Yields (in $t\ ha^{-1}$) for a hypothetical experiment with 20 plots arranged on a grid of five rows by four columns, in which five varieties labelled A–E are to be tested. By accident, however, the same variety A is tested on all 20 plots. The figure shows the 20 yields of variety A on the 20 plots. This example shows a systematic trend from lower left to upper right corner.

One possible randomization for a completely randomized design is shown in Figure 2(a). Recall that despite the different variety labels allocated to plots, we assume that in reality variety A is tested on all plots. Classical analysis of variance (ANOVA), which provides an *F*-test for the global null hypothesis of no treatment difference, relies on the random allocation of treatments to plots as exemplified in Figure 2(a). Specifically, randomization-based justification of ANOVA assumes that each possible allocation is equally likely. With this assumption, the distribution of the ANOVA *F*-statistic under the null hypothesis, in short 'the null distribution', can be derived either analytically or from the randomization distribution (Pitman 1938, Good 2000). Other statistics and procedures can be handled in an analogous fashion, but here we will focus mainly on the *F*-statistic for simplicity. We will shortly study the randomization distribution in some more detail and show later that it is essentially the same as the analytical *F*-distribution.

Alternatively, consider a systematic allocation of varieties to plots as shown in Figure 2(b). Systematic designs similar to this one are sometimes used, for example, in precision farming experiments (Piepho et al. 2011). Note that the design could also be regarded as a complete block design with blocks corresponding to columns, and a systematic arrangement of treatments within columns. In principle, such an allocation could occur purely by chance when randomizing completely, but the probability of a systematic allocation occurring is very small indeed. If this systematic allocation is used, the one-way ANOVA *F*-statistic is $F = 5.63$. But in reality, there are no differences among treatments (each treatment is the same variety A), so in this case the large *F*-value results solely from the pronounced systematic spatial trend down the columns seen in the yield data in Figure 1. The resulting bias is also apparent when comparing the treatment means for the systematic design with those of the randomized design in Table 1.

(a) Randomized design

D	B	D	E
C	E	A	B
A	C	E	A
D	D	A	B
B	C	E	C

(b) Systematic design

A	A	A	A
B	B	B	B
C	C	C	C
D	D	D	D
E	E	E	E

Fig. 2 Two allocations of variety labels to plots for field layout in Figure 1. (a) One possible randomized allocation. (b) Systematic allocation.

Table 1 Means (in t ha^{-1}) for data in Figure 1 (systematic spatial trend) using randomization in Figure 2(a) and systematic design in Figure 2(b)

Treatment	Randomized design Figure 2(a)	Systematic design Figure 2(b)
A	7.73	7.30
B	7.78	7.50
C	8.20	7.90
D	8.03	8.25
E	7.73	8.50
Range of treatment means	0.47	1.20
F-value	0.46	5.63

Table 2 Means (in t ha^{-1}) for data in Figure 3 (erratic spatial pattern) using randomization in Figure 2(a) and systematic design in Figure 2(b)

Treatment	Randomized design Figure 2(a)	Systematic design Figure 2(b)
A	7.85	8.25
B	7.85	7.30
C	7.58	7.90
D	8.45	8.50
E	7.73	7.50
Range of treatment means	0.87	1.20
F-value	1.36	5.63

```
8.8  7.7  8.0  8.5
7.6  7.2  7.4  7.0
7.3  8.1  7.9  8.3
8.0  9.0  8.4  8.6
8.1  7.6  7.3  7.0
```

Fig. 3 Yields (in t ha^{-1}) for a hypothetical experiment with 20 plots arranged on a grid of five rows by four columns, in which five varieties labelled A–E are to be tested. By accident, however, the same variety A is tested on all 20 plots. The figure shows the 20 yields of variety A on the 20 plots. This example shows no systematic trend.

A popular justification for using the design in Figure 2(b) is that plots in each row can be managed in a single pass of a tractor. It is important to note, however, that in this case the experimental units (randomization units) are the rows, not the plots. Clearly, the plots must be considered as pseudo-replicates or subsamples in this case, so there is only one true replicate per treatment. The appropriate ANOVA for this design is based on a mean value (or sum) per row and has no degrees of freedom for error, which reflects the false-replicate problem.

The data in Figure 1 display a strong systematic trend with yield increasing in one particular direction. To illustrate that randomization is also useful when there is a more erratic spatial pattern with no readily discernible systematic trend, the data in Figure 3 are considered. Note that the yields in Figure 3 are exactly the same as in Figure 1, but their positions are changed relative to those in Figure 1. Specifically, the data were generated from those in Figure 1 by first permuting the rows and then permuting plots within rows, meaning that the composition of rows remained unaltered. Thus, the *F*-value for the systematic design in Figure 2(b) is $F = 5.63$, which is exactly the same as with the data in Figure 1, as are the variety means (Table 2). For the randomized design in Figure 2(a), the *F*-value is small by comparison ($F = 1.36$). The two examples (Figs 1 & 3) show that failure to randomize can be a problem both with and without clearly discernible systematic trend.

To gain further insight, it is instructive to more closely consider the null distribution of the ANOVA *F*-statistic for a randomized design. We need to refer to a null distribution because the particular allocation of treatment labels to plots used in any one trial is only one instance of a usually large number of possible allocations, and all of these possible allocations need to be considered when judging the value of a single *F*-statistic. As mentioned previously, the null distribution for the *F*-test is usually taken to be the analytic *F*-distribution with suitable numerator and denominator degrees of freedom (Mead et al. 2002). This null distribution assumes that under the randomization scheme used for the trial, each possible allocation of treatments to plots is equally likely. Alternatively, this null distribution can be obtained empirically by permutation as follows: generate a large number of (if feasible the complete set of all) possible allocations of variety labels to plots and for each allocation compute the *F*-statistic. This yields an empirical version of the null distribution of the ANOVA *F*-statistic (Pitman 1938, Good 2000). The rationale of this randomization distribution is as follows: If there are no real treatment effects, that is, if each treatment has the same expected value, it should be of no consequence for the expected outcome of the experiment, when we reshuffle the allocation of treatment labels to plots. Thus, when testing the null hypothesis of no treatment effects, the *F*-value corresponding to the particular randomization used in an experiment must be viewed as just one instance of the range of possible *F*-values that can be expected under the null hypothesis of no treatment effects. Figure 4 shows the distribution for 100 000 randomizations using the data in Figure 1. Note that, assuming that all possible randomizations are used, this randomization distribution, as well as the observed *F*-value for the systematic design (Fig. 2b), is the same when using the data in Figure 3, so the exposition that follows applies equally to both examples (with and without systematic trend). It should be clear that the randomization distribution in Figure 4 is really a null distribution, because we are assuming that accidentally the same variety A was in fact tested on all plots. It is stressed, however, that the

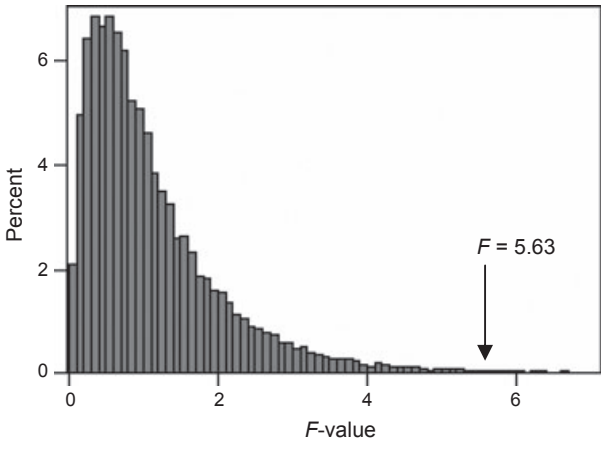

Fig. 4 Randomization distribution of ANOVA *F*-statistic using data from Figure 1 and the same as with data in Figure 3. n = 10^5.

randomization approach would also be valid had the seed bags of varieties A–E not been mixed up, because under the null hypothesis of no treatment effects, treatment labels are always exchangeable (Good 2000). A randomization approach is regarded as valid here, if the subsequent ANOVA controls the Type I error rate at the nominal level.

The general idea of a significance test is to reject the null hypothesis whenever the observed value of the test statistic is atypical compared with the null distribution. To assess how typical or atypical an observed *F*-value is, we may compute the proportion of times an *F*-value from the randomization distribution exceeds, or is equal to the *F*-value for the experiment. This proportion, or probability, is known as the P-value. For the systematic design in Figure 2(b), the randomization P-value is P = 0.0071, indicating that the observed $F = 5.63$ is a very atypical value in relation to the distribution of *F*-values expected under the null hypothesis (Fig. 4): only 0.71 % of *F*-values are expected to be as large as, or larger than the one observed for the systematic allocation. If the observed *F*-value is atypical under the assumed null hypothesis, this is reason to reject that null hypothesis. The conventional threshold is to reject the null hypothesis, when the P-value is smaller than $\alpha = 0.05$. The randomization P-value is very close to the P-value obtained by reference to an *F*-distribution with four numerator and 15 denominator degrees of freedom (P = 0.0057), the analytic reference null distribution for this example. In principle, we could always conduct *F*-tests using the randomization distribution, but it is computationally much less demanding to just use the analytical result; unless the data display gross departures from normality or homogeneity of variance, the outcome will be virtually the same in both cases.

In the two examples, the systematic design leads to a very large *F*-value, an *F*-value that would be considered very atypical compared with the randomization distribution in Figure 4. Thus, the null hypothesis would be falsely rejected with this unrandomized design, exemplifying the danger incurred when not randomizing. The crucial point to re-iterate here is that the null distribution just discussed (see Fig. 4) is the appropriate reference distribution for the *F*-test only if the trial was actually randomized. Only in this case will each randomization be equally likely and only in this case can the null distribution be derived from a classical randomization argument (Calinski and Kageyama 2000). In fact, one may argue that the randomization distribution in Figure 4 is not the appropriate reference distribution for the systematic design because this cannot be regarded as a random sample from the set of all possible fully randomized designs (Kempthorne 1977). Clearly, if the researcher decides to choose a systematic design with treatments laid out in rows of four plots as shown in Figure 2(b), the only freedom left is a permutation of treatments among rows. But each such permutation yields the exact same *F*-value, so the reference distribution in this case collapses to a single possible *F*-value! This consideration shows that no meaningful test can be based on a randomization argument for the systematic design.

It is also instructive to consider the distribution of analytical P-values associated with the *F*-statistics computed for the randomization distribution. It is an important statistical fact that for the *F*-test (or indeed any other significance test) to be valid, the randomization distribution of P-values must follow a uniform distribution. Thus, studying the randomization distribution of analytical P-values provides a particularly convenient way to assess the adequacy of any significance test. For the example, the P-values are nicely uniformly distributed (Fig. 5), so the test is valid.

Likewise, the estimated treatment differences are unbiased with respect to the randomization distribution, as

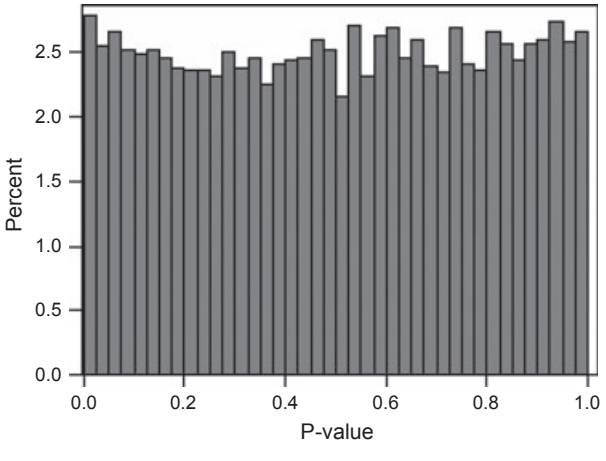

Fig. 5 Uniform distribution of P-values for ANOVA *F*-statistic under randomization distribution based on data in Figure 1. n = 10^5.

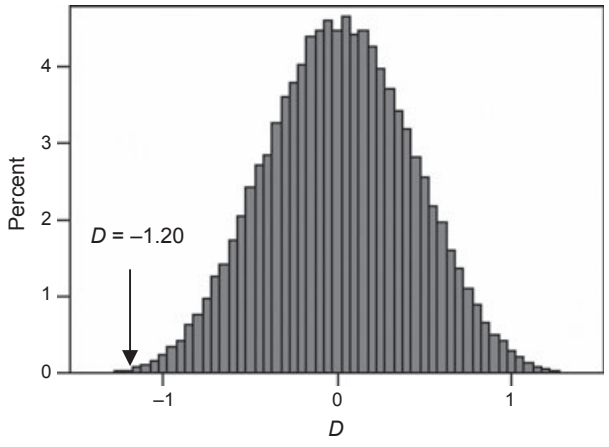

Fig. 6 Randomization distribution of the differences d (in t ha^{-1}) of means for variety labels A and E based on data in Figure 1. n = 10^5. Note that yields pertaining to both labels, in fact, come from variety A, so the expected value of d is zero. Results presented in this figure also apply for the data in Figure 3, if we consider the comparison of varieties A and B.

shown for the example of the mean difference of variety labels A and E (Fig. 6) using the data in Figure 1 (Results discussed in this paragraph are identical for the data in Figure 3 if we consider the comparison of varieties A and B; Table 2). The distribution is centred at zero, which correctly reflects the fact that yields pertaining to both labels come from variety A, so there is no treatment difference. By contrast, for the systematic design the observed difference is $d = -1.20$ (Table 1), which is far removed from the centre of the randomization distribution and so would be judged significant (P = 0.0011). This shows that severe bias of treatment effect difference estimates may be incurred without randomization. In contrast, randomization ensures the absence of the main sources of systematic error (Cox 1958:7). Moreover, the randomization distribution looks nicely normal, which is why we may safely use the t-test for comparing two treatment means. The P-value can be computed analytically by reference to a t-distribution or from the randomization distribution generated by permuting treatment labels A–E among all observations (We have assumed here for simplicity that the global null hypothesis is true. If only the partial null hypothesis holds, stating that means of A and E are identical, then only labels of these two treatments should be permuted, giving rise to a rather more limited set of permutations. When the number of permutations becomes too small, it is preferable to employ the analytical null distribution).

A uniformity trial

To further illustrate the limitations of statistical analysis in the case of improper randomization, we analysed a uniformity trial with triticale conducted at the Ihinger Hof experiment station of the University of Hohenheim in 2007 on a plot grid of 36 rows by 30 columns. The 1080 plots were planted with the same variety and managed uniformly, so there was only a single treatment. We used the same systematic design as in the previous section (Fig. 2b), that is, we overlaid the uniformity trial data with small trials of 20 plots each using a sliding 5 row × 4 column window of plots (see Figure S1 in Supplemental Material). Thus, for each 5 row × 4 column grid, the treatment labels were aligned with rows, that is, treatment A was allocated to the four plots in the first row, treatment B was allocated to the second row, etc. For the randomized design, we used a new randomization for each window. Because the designs were superimposed onto uniformity trial data, the resulting trials simulated the global null hypothesis of no treatment differences (Richter and Kroschewski 2012).

With proper randomization, the analytical P-values for the F-tests should be uniformly distributed. Figure 7(a) shows that this is approximately the case with randomization, though the histogram cannot be expected to be perfectly uniform due to the small number of tests (n = 864). With the systematic design, small P-values are strongly over-represented (Fig. 7b), indicating that the F-test is not valid. We also checked if a spatial analysis rectifies the problem. We used a separable AR1 × AR1 model (Gilmour et al. 1997). This model assumes that the correlation of plots decays exponentially with spatial distance along rows and along columns. To achieve convergence of the spatial model in almost all cases for the systematic design, we constrained the correlation estimates to be no larger

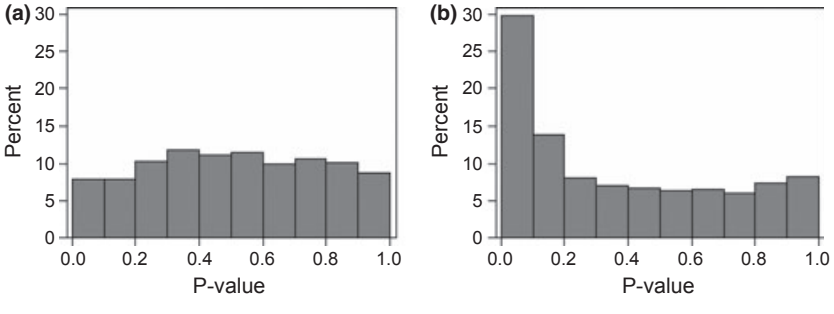

Fig. 7 Histogram of n = 864 analytical P-values of ANOVA F-test from analysis of a uniformity trial assuming independent errors. (a) Randomized design. (b) Systematic design.

than 0.99. The significance of *F*-values was approximated by the first-order method of Kenward and Roger (2009). Simulation runs in which the spatial model did not lead to convergence of the REML algorithm was discarded. Figure 8 indicates that there is some improvement, but small P-values are still over-represented and the distribution is far from uniform when randomization is lacking. In case of randomization, the spatial analysis tended to be slightly on the liberal side, that is, the null hypothesis was rejected too frequently. Of course, we do not claim here to have used the most appropriate spatial model for this data set, so there may be a spatial model that would have led to a somewhat more even distribution of P-values. But the example does demonstrate that a spatial analysis cannot generally be expected to salvage an unrandomized experiment (Bailey et al. 1995, Cox 2009). Or as Kempthorne (1977:23) put it: 'If one … does not use the randomization framework of design and inference, one is at the mercy of the vagaries of whatever model search procedure or model assumptions one uses'.

While there is some merit in considering the uniformity trial data, our findings are somewhat limited in scope, because of the relatively small number of tests and because tests for the sliding window of overlaid designs are not strictly independent as there is a partial overlap of the data used to compute the test statistic and its associated P-value. We could have tried to avoid the problem by only looking at non-overlapping windows, but this would have considerably reduced the total number of tests, thus rendering the histogram uninformative, so we preferred the overlapping and thus correlated windows. Also, even non-overlapping windows are not entirely independent if spatial correlation exists between windows. In the next section, we will use simulation to overcome these problems. The results of the simulation study are more robust with respect to the main issues associated with randomization.

A small simulation study

To further study the effect of lacking randomization and to what extent a spatial analysis can salvage the analysis, we simulated field plot data following a spatial error model. The advantage of a simulation is that the model generating the data is known, so adverse effects of model misspecification can be ruled out. Here, errors were simulated from a separable AR1 × AR1 model with varying values of the autocorrelation parameters ρ_r and ρ_c for correlation over rows and columns, respectively, and spatial variance $\sigma_s^2 = 1$. In a further simulation study, we overlaid the spatial errors with a random row effect with variance $\sigma_r^2 = 0.5$. In this case, the spatial variance was reduced to $\sigma_s^2 = 0.5$. The number of treatments (t) was set equal to the number or rows (r), while the number of replicates per treatment was set equal to the number of columns (c). We set $c = 4$ and $r = t = 5$, 10. Two scenarios were considered for the allocation of treatments to plots:

- Systematic allocation of treatments to rows of plots, that is, all replicates of a treatment were located in the same row.
- Completely random allocation (the design was randomized afresh for each run).

The simulated yields had no treatment effect, so the data reflect the global null hypothesis of no treatment differences. Three methods for the analysis of variance were used with regard to modelling the error distribution:

- Classical ANOVA assuming independent errors.
- Analysis assuming the AR1 × AR1 model (with or without random row effect) that was used to simulate the data, supplying the variance–covariance structure used to simulate the data as a known quantity that was held fixed in analysis (a Wald-type *F*-test was used to test the global null hypothesis and the method of generalized least squares was used to estimate treatment differences).
- Same analysis as with second method, but with the parameters of the error model being estimated from the data. For estimating the spatial model and determining the significance of *F*-values, we used the same specifications as with the uniformity trial.

We assessed the empirical Type I error rate at a significance level of $\alpha = 5$ %. Moreover, we assessed the mean

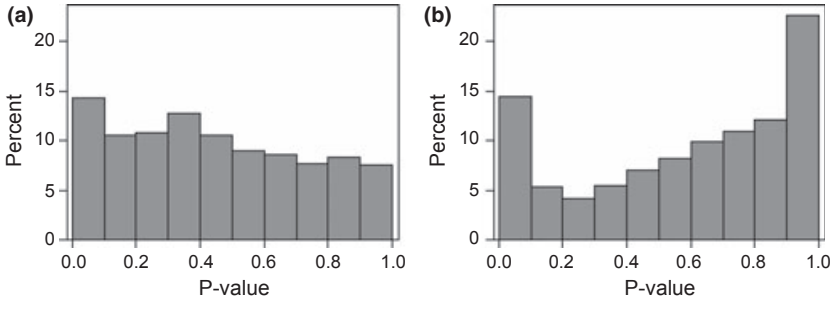

Fig. 8 Histogram of n = 864 analytical P-values of ANOVA *F*-test from analysis of a uniformity trial assuming spatially correlated errors. AR1 × AR1. (a) Randomized design. (b) Systematic design.

squared error of estimated treatment differences. Thus, for each treatment comparison and simulation, we computed the difference of true and estimated difference, squared this discrepancy and averaged these squared differences over all treatment comparisons and simulations.

When the trials were randomized, classical ANOVA and spatial analysis with known covariance structure were valid (Table 3), as is expected from theory. Spatial analysis using the correct model, but when parameters needed to be estimated, was approximately valid. Particularly, in small samples, Type I error rates can deviate somewhat from nominal rates. In simulations, the tests tended to be on the liberal side. Without randomization, classical ANOVA was clearly invalid, because spatial effects associated with rows could not be dissected from treatment effects. If the correct spatial model was used and variance components were fixed at true values, the significance test was still valid. When parameters had to be estimated, however, the test was not generally valid. In particular, empirical significance levels were liberal when there was a random row effect, because this was confounded with treatment effects. Without randomization, the mean squared error of differences was about the same for classical ANOVA and spatial analysis (Table 4). For randomized trials, the mean squared errors were much reduced in most cases, most notably when spatial analysis was used. This indicates that randomization can improve precision and that spatial analysis can be more efficient than classical ANOVA provided the trial is appropriately randomized.

The following conclusions emerge from this simulation study:

- Classical ANOVA can be severely invalidated when randomization is lacking.
- When proper randomization is used, classical ANOVA is valid even if the plot trend follows some spatial model. This is essentially because randomization breaks any spatial correlation.
- Spatial analysis is not guaranteed to salvage the analysis of an unrandomized experiment, even if the model is correctly specified, because there is always a danger of confounding of treatment effects and field trend.
- Lack of randomization can result in considerable loss of precision of contrast estimates.
- With proper randomization, spatial analysis is approximately valid in the case we considered. Moreover, spatial analysis can considerably improve precision compared with classical ANOVA when the model is correctly specified.

Concluding Remarks

We hope that this article has convinced or reassured the reader that randomization is a key component of well-designed experiments. For simplicity, our focus was on the completely randomized design. There may be good reasons, however, to restrict randomization in order to account for any known or suspected trend. Most importantly, blocking of experimental units, meaning that the set of treatments allocated to a block must remain unaltered after randomization, provides an efficient way to account for spatial trend and reduce experimental error (Edmondson 2005). Efficient systematic designs without blocking and randomization can sometimes be found when the form of trend is known and of a simple parametric form (Edmondson 1993), but validity of statistical analysis strongly depends on the validity of the assumed model. Unless the trend is very well known, which will rarely be the case in agricultural experiments, it is usually better to use blocking in combination with randomization, because this design

Table 3 Empirical Type I errors at $\alpha = 5$ % in simulation

					Not randomized			Randomized		
$r = t$	c	ρ_r	ρ_c	σ_r^2	ANOVA	AR × AR fixed	AR × AR estimated	ANOVA	AR × AR fixed	AR × AR estimated
5	4	0.2	0.8	–	0.852	0.049	0.086	0.053	0.054	0.067
		0.5	0.5	–	0.304	0.052	0.094	0.054	0.051	0.069
		0.8	0.2	–	0.032	0.050	0.054	0.054	0.052	0.063
10	4	0.2	0.8	–	0.984	0.051	0.061	0.050	0.054	0.052
		0.5	0.5	–	0.561	0.052	0.071	0.047	0.053	0.054
		0.8	0.2	–	0.079	0.051	0.062	0.050	0.054	0.051
5	4	0.2	0.8	0.5	0.964	0.049	0.105	0.055	0.055	0.078
		0.5	0.5	0.5	0.831	0.049	0.178	0.055	0.053	0.070
		0.8	0.2	0.5	0.680	0.051	0.376	0.054	0.052	0.061
10	4	0.2	0.8	0.5	1.000	0.052	0.103	0.051	0.055	0.060
		0.5	0.5	0.5	0.977	0.051	0.216	0.051	0.053	0.056
		0.8	0.2	0.5	0.894	0.051	0.576	0.049	0.053	0.047

10 000 simulation runs. r = no. of rows; t = no. of treatments; c = no. of columns; ρ_r, ρ_c = serial correlation for rows and columns, respectively; σ_r^2 = variance of row effects. Spatial variance $\sigma_s^2 = 0.5$ when there was a random row effect, otherwise $\sigma_s^2 = 1.0$

Table 4 Mean squared error of mean difference estimates in simulation

					Not randomized			Randomized		
$r = t$	c	ρ_r	ρ_c	σ_r^2	ANOVA	AR × AR fixed	AR × AR estimated	ANOVA	AR × AR fixed	AR × AR estimated
5	4	0.2	0.8	–	1.407	1.364	1.377	0.410	0.150	0.180
		0.5	0.5	–	0.718	0.696	0.716	0.405	0.247	0.308
		0.8	0.2	–	0.233	0.231	0.241	0.404	0.141	0.171
10	4	0.2	0.8	–	1.476	1.431	1.443	0.459	0.152	0.165
		0.5	0.5	–	0.848	0.822	0.841	0.446	0.245	0.270
		0.8	0.2	–	0.339	0.336	0.345	0.421	0.131	0.142
5	4	0.2	0.8	0.5	1.706	1.685	1.691	0.413	0.079	0.097
		0.5	0.5	0.5	1.362	1.352	1.361	0.410	0.143	0.175
		0.8	0.2	0.5	1.121	1.120	1.125	0.410	0.098	0.116
10	4	0.2	0.8	0.5	1.741	1.719	1.725	0.463	0.079	0.089
		0.5	0.5	0.5	1.428	1.415	1.425	0.457	0.141	0.158
		0.8	0.2	0.5	1.174	1.173	1.178	0.444	0.091	0.098

10 000 simulation runs. r = no. of rows; t = no. of treatments; c = no. of columns; ρ_r, ρ_c = serial correlation for rows and columns, respectively; σ_r^2 = variance of row effects. Spatial variance $\sigma_s^2 = 0.5$, when there was a random row effect, otherwise $\sigma_s^2 = 1.0$

strategy does not require any assumptions about the form of trend and provides randomization protection. In factorial experiments, there is often a need to restrict randomization of one or several factors for technical reasons. Split-plot designs allow accommodating such needs. But despite any such restrictions on randomization, to obtain valid conclusions, (restricted) randomization can and always should be used with these types of design.

As we indicated in the introduction, our exposition of randomization does not provide a detailed and in-depth discussion of the underlying mathematical theory, which can be found in pertinent textbooks, for example, Calinski and Kageyama (2000) and Hinkelmann and Kempthorne (1994). A few issues will be briefly mentioned here. (i) A fundamental assumption underlying all randomization theory is that of unit-treatment additivity, that is, that the effects of the experimental unit and of the treatment applied to it affect the response additively. This assumption implies that there is no interaction between treatment and the environmental conditions prevailing on the experimental unit. It must be realized that this assumption is a strong one and that it is not usually easy to verify. If experimental units are relatively homogeneous, then the assumption is often tenable. But with increasing heterogeneity, the likelihood of unit-treatment interaction increases. Depending on the trait, a data transformation may be needed to meet the additivity assumption, for example, for count data (Mead et al. 2002). (ii) Our examples may have raised the impression that the F-statistic is somehow fundamental to the null randomization distribution but it is not. Randomization theory applies to any statistic or procedure used for inference, such as standard errors, t-tests and confidence intervals. Moreover, randomization ensures unbiasedness of treatment contrast estimates. (iii) When analysing randomized experiments, we are accustomed to rely on a linear model with an independently and normally distributed error term. This model can be developed from randomization theory, whereby the process of randomization is represented by design random variables, which induce the stochastic properties of the residual error term (Kempthorne 1977). These properties hold irrespective of the particular effects or response levels associated with the experimental units used. In particular, the randomization-induced derivation of a linear model with independent error terms is in no way invalidated by any spatial correlation we detect in real data collected from a randomized experiment. The permutation test is valid irrespective of the underlying error distribution (Good 2000), while reference to the analytical F- and t-distributions requires approximate normality, so linear model residuals should be scrutinized for any departures from normality in case the latter are used. When the data are approximately normally distributed, permutation and analytical null distributions will be very similar. (iv) We have considered only the simplest case of a single randomization step. Complex experiments, such as split-plot experiments, can involve several randomization steps, giving rise to multiple 'error strata'. Such experiments can be analysed by mixed models with simple variance structures representing several error terms, which can be justified from the fact that all possible permutations of labelling experimental units define a multivariate distribution (Nelder 1965a,b).

One may consider optimizing the experimental design with a spatial analysis in mind. The performance of designs optimized for a purely spatial analysis can be very good if the model is correctly specified. The main advantage of

randomization is that it provides protection against all types of underlying experimental structure (spatial trend, effects from technical equipment such as tyre tracks, differences between plots in time of harvest, etc.). There can be a price to pay in terms of efficiency, but this price is usually cheap compared with the advantage of repeatability of error estimates. Also, the loss of efficiency from randomization can be minimized by using an efficient blocking structure (incomplete blocks, row-column designs). If a blocked design is to be optimized for a particular spatial analysis, while still providing randomization protection, randomization must be restricted. Thus, the randomization set, that is, the set of possible randomizations from which to randomly choose the allocation of treatments to plots, will be reduced compared with the full set under complete randomization (Williams et al. 2006). The reduced set comprises the allocations that are known to yield the most efficient treatment estimates under the assumed spatial model. While there is no full randomization theory when the randomization set is reduced (Williams 1986, Monod et al. 1996), analysis using a spatial model can usually be expected to be both efficient and approximately valid provided the spatial model is well chosen (These comments are in just relation to spatial analysis. For example, one can, in fact, have strong validity with a reduced randomization set in other contexts as demonstrated by Grundy and Healy (1950)). In practice, however, it is mostly difficult to predict the most suitable spatial model. A conservative strategy to spatial modelling is to just use the full randomization set for an efficient blocked design, try various spatial add-on models for analysis, but to keep a purely randomization-based analysis assuming independent errors as a fall-back option in case spatial add-on components do not lead to an improved fit (Piepho and Williams 2010).

Some factors of interest are not amenable to randomization, most notably the time factor in growth course analyses. In this case, a repeated measures problem arises. Repeated measures experiments allow for a valid statistical analysis, provided that the other treatment factors have been properly randomized. If an experiment involves repeated measures, then a very simple litmus test for availability of a randomization-based analysis is as follows: Is a valid statistical analysis possible for a single time point? If the answer is 'yes', the experiment has been appropriately randomized, and hence, standard procedures may be used for statistical analysis. In a repeated measures mixed model analysis, observations from different randomization units (plots) are modelled as independent, while repeated observations on the same unit are modelled as serially correlated (Piepho et al. 2004, Loughin et al. 2007, Brien and Demetrio 2009). Ignoring such correlations invalidates statistical inference.

The examples considered in this study all concerned field experiments, but it should be stressed that randomization is a universally valid and useful principle in experimental design. For example, in our experience, systematic designs are quite frequently used in laboratory work, where biological replicates from the field, greenhouse, pen, etc. are processed in systematic order. For example, all replicates of treatment A would be analysed first, then all replicates of treatment B, etc. A further example, lucidly discussed by Greenberg (1951), is given by systematic designs, where two treatments are tested in alternate order. If there is any time trend, for example, due to gradual changes in the settings or conditions of laboratory equipment, the same kinds of bias as in field experiment must be expected. Such bias can be very easily avoided by proper randomization.

Acknowledgements

Jens Möhring was supported by DFG grant PI 377/13-1.

References

Bailey, R. A., J. Azaïs, and H. Monod, 1995: Are neighbour methods preferable to analysis of variance for completely systematic designs? 'Silly designs are silly!' Biometrika 82, 655–659.

Brien, C. J., and C. G. B. Demetrio, 2009: Formulating mixed models for experiments, including longitudinal experiments. J. Agric. Biol. Environ. Stat. 14, 253–280.

Calinski, T., and S. Kageyama, 2000: Block Designs: A Randomization Approach. Volume I: Analysis. Springer, Berlin.

Cox, D. R., 1958: Planning of Experiments. Wiley, New York, NY, USA.

Cox, D. R., 2009: Randomization in the design of experiments. Int. Stat. Rev. 77, 415–429.

Edmondson, R. N., 1993: Systematic row-and-column designs balanced for low order polynomial interactions between rows and columns. J. R. Stat. Soc. Series B 55, 707–723.

Edmondson, R. N., 2005: Past development and future opportunities in the design and analysis of crop experiments. J. Agric. Sci. 143, 27–33.

Fisher, R. A., 1925: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Gilmour, A. R., B. R. Cullis, and A. R. Verbyla, 1997: Accounting for natural and extraneous variation in the analysis of field experiments. J. Agric. Biol. Environ. Stat. 2, 269–293.

Good, P., 2000: Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypotheses. Springer, Berlin.

Greenberg, B. G., 1951: Why randomize? Biometrics 7, 309–322.

Grundy, P. M., and M. J. R. Healy, 1950: Restricted randomization and quasi-Latin squares. J. R. Stat. Soc. Series B 12, 286–291.

Haase, T., C. Schüler, H. P. Piepho, H. P. Thöni, and J. Heß, 2007: The effect of preceding crop and presprouting on crop

growth, N use and tuber yield of organic maincrop potatoes for processing under conditions of N stress. J. Agron. Crop Sci. 193, 270–291.
Hinkelmann, K., and O. Kempthorne, 1994: Design and Analysis of Experiments. Volume 1: Introduction to Experimental Design. Wiley, New York, NY, USA.
John, J. A., and E. R. Williams, 1995: Cyclic and Computer Generated Designs, 2nd edn. Chapman and Hall, London.
Kempthorne, O., 1977: Why randomize? J. Stat. Plan. Inference 1, 1–25.
Kenward, M. G., and J. H. Roger, 2009: An improved approximation to the precision of fixed effects from restricted maximum likelihood. Comput. Stat. Data Anal. 53, 2583–2595.
Leiser, W., H. F. Rattunde, H. P. Piepho, and H. K. Parzies, 2012: Getting the most out of sorghum low-input field trials in West Africa using spatial adjustment. J. Agron. Crop Sci. 198, 349–359.
Loughin, T. M., M. Poehlman-Roediger, G. A. Milliken, and J. P. Schmidt, 2007: On the analysis of long-term experiments. J. R. Stat. Soc. 170, 29–42.
Mead, R., R. N. Curnow, and A. M. Hasted, 2002: Statistical Methods in Agriculture and Experimental Biology. CRC Press, Boca Raton, FL, USA.
Monod, H., J. M. Azaïs, and R. A. Bailey, 1996: Valid randomization for the first-difference analysis. Aust. J. Stat. 38, 91–106.
Mühleisen, J., J. Reif, H. P. Maurer, J. Möhring, and H. P. Piepho, 2013: Visual scorings of drought stress intensity as covariates for improved variety trial analysis. J. Agron. Crop Sci. 199, 321–330.
Nelder, J. A., 1965a: The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. Proc. R. Soc. Series A 283, 147–162.
Nelder, J. A., 1965b: The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. Proc. R. Soc. Series A 283, 163–178.
Piepho, H. P., and E. R. Williams, 2010: Linear variance models for plant breeding trials. Plant Breeding 129, 1–8.
Piepho, H. P., A. Büchse, and C. Richter, 2004: A mixed modelling approach to randomized experiments with repeated measures. J. Agron. Crop Sci. 190, 230–247.
Piepho, H. P., C. Richter, J. Spilke, K. Hartung, A. Kunick, and H. Thöle, 2011: Statistical aspects of on-farm experimentation. Crop Pasture Sci. 62, 721–735.
Pitman, E. J. G. 1938: Significance tests which may be applied to samples from any populations. III. The analysis of variance test. Biometrika 29, 322–335.
Richter, C., and B. Kroschewski, 2012: Geostatistical models in agricultural field experiments: investigations based on uniformity trials. Agron. J. 104, 91–105.
Smith, M. F. C. 1938: An empirical law describing heterogeneity in the yields of agricultural crops. J. Agric. Sci. 28, 1–23.
Williams, E. R. 1986: Neighbor analysis of uniformity data. Aust. J. Stat. 28, 182–191.
Williams, E. R., and D. J. Luckett, 1988: The use of uniformity data in the design and analysis of cotton and barley variety trials. Aust. J. Agric. Res. 39, 339–350.
Williams, E. R., J. A. John, and D. Whitaker, 2006: Construction of resolvable spatial row- column designs. Biometrics 62, 103–108.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Figure S1: Sketch of a sliding window of 5 rows × 4 columns. Two consecutive positions super-imposed over uniformity trial of 36 rows by 30 columns of plots.