Published online November 21, 2006

# New and Traditional Methods for the Analysis of Unreplicated Experiments

Roger W. Payne

## ABSTRACT

**This paper reviews some traditional and more recent methods for analyzing unreplicated experiments. Such experiments have presented a challenge to statisticians throughout their involvement in agricultural research. At Rothamsted this began in 1919, when R.A. Fisher was appointed to analyze the accumulated data from the classical field experiments. Fisher's experiences with the classicals, which had virtually no replication, must have contributed to his inclusion of replication as one of the key features of a well-designed experiment. Nevertheless, Fisher made good use of Rothamsted's data, for example in his study of the influence of rainfall on yields from the Broadbalk. He also devised the randomization test, which can be used to analyze unreplicated data. More recently, Broadbalk has also been used to study climate change and sustainability. Newer developments have been concerned to find alternatives to use, instead of blocking, to take account of the spatial variation within an experiment. The resulting methods for modeling spatial correlations have allowed experimenters to obtain more precise estimates of treatment effects—or to decrease numbers of replicates—and they can also provide reliable analyses of unreplicated treatments.**

The development of methods for the analysis of experiments began in 1919, when R.A. Fisher was appointed as the original statistician at Rothamsted. His remit was to study the accumulated results of the Rothamsted classical field experiments, which began in 1843 with the Broadbalk experiment on winter wheat (*Triticum aestivum* L.). This illustrates another (more unfortunate!) tradition, of the statistician being called in to do the analysis long after the design of an experiment. However, Fisher rose to the challenge and, as the 1971 Guide to Rothamsted Experimental Station notes, "he soon realized the need for improved statistical techniques over the whole range of agricultural and biological research, and the groundwork for modern statistics was laid by him during the 1920s and 1930s." Fisher recognized the importance of replication, as a way of allowing the underlying random variation to be estimated. However, he also made good use of the Rothamsted classical experiments, which essentially were unreplicated.

## RESULTS AND DISCUSSION

### Relationship of Yield to Climate

Fisher (1924) used data from Broadbalk to study the relationship of yield to climatic variation. Broadbalk was set up by Sir John Lawes and Sir Henry Gilbert in autumn 1843 to study the effects of inorganic fertilizers on winter wheat (see Leigh and Johnston, 1994). It still continues, with minor modifications generally to reflect changing practices (e.g., of varieties), and provides a resource for many research activities not anticipated by Lawes and Gilbert. The main treatments on the plots, shown in Table 1, have remained largely unchanged since 1852.

In 1926, subsequent to the period studied by Fisher, each plot was split into sections to allow fallowing as a means of weed control. More recently, some of the sections have been used to study crop rotations. Nevertheless, there are still sections that have grown wheat continuously since 1843. From a statistical point of view it is intriguing to notice that the design contains some factorial structure (e.g., plots 03, 05, 10, and 09). Thus, Lawes and Gilbert were already thinking about how one fertilizer will respond to the presence and absence of other fertilizers, although this was long before the theory of factorial experiments was devised by Fisher (1926). However, there is no replication.

The long-term experiments provided a uniquely useful resource for this purpose, giving a long series of data in controlled conditions and with the same treatments year on year. It should therefore be only climate that was affecting yields, other than perhaps a few other aspects that Fisher called "progressive changes." Specifically, Fisher was aware that there would be long-term trends in yield as well as long-term trends in climate, and that the yield trends could be caused by progressive changes in aspects unrelated to climate (e.g., improvements in husbandry or varieties). Fisher therefore fitted 5th-order Legendre polynomials of years to the annual yields to remove the long-term trends from the annual yields. The fluctuations about the Legendre model might reasonably be assumed to arise mainly from annual climatic variations, and he sought to model these by using the rainfall trends within each year. These rainfall trends were characterized by fitting further Legendre polynomials to the rainfall summarized during 6-d intervals. Fitting a multiple regression, using the Legendre rainfall coefficients of the rainfall trends, accounted for between 11 and 40% of the variance of the yield fluctuations. Furthermore, Fisher found consistent responses to rainfall according to the plots' fertilizer regimes. The work is impressive not only in terms of the computations that were required using just an electric calculator, but also in terms of current methodology. These days we would probably use smoothing splines to model the trends rather than Legendre polynomials, but the underlying ideas remain the same.

Further analyses of the Broadbalk yields were done by Chmielewski and Potts (1995), who retained the long-term trends so that they could study long-term climate variations in order to assess the implications of climate

VSN International, Hemel Hempstead, Hertfordshire HP1 1ES, U.K., and Biomathematics and Bioinformatics Division, Rothamsted Research, Harpenden, Herts, U.K. Received 27 Apr. 2006. Corresponding author (roger@vsn-intl.com).

Published in Crop Sci. 46:2476–2481 (2006).
Analysis of Unreplicated Experiments (Symposium)
doi:10.2135/cropsci2006.04.0273
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

**Table 1. Main treatments on Broadbalk.**

| Plot | Treatment |
|---|---|
| 01 | (Fym) N4 |
| 21 | Fym N3 |
| 22 | Fym |
| 03 | Nil |
| 05 | (P) K Mg |
| 06 | N1 (P) K Mg |
| 07 | N2 (P) K Mg |
| 08 | N3 (P) K Mg |
| 09 | N4 (P) K Mg |
| 10 | N4 |
| 11 | N4 P Mg |
| 12 | N1 + 3 + 1 (P) K2 Mg2 |
| 13 | N4 P K |
| 14 | N4 P K* (Mg*) |
| 15 | N5 (P) K Mg |
| 16 | N6 (P) K Mg |
| 17 | N1 + 4 + 1 P K Mg |
| 18 | N1 + 2 + 1 P K Mg |
| 19 | N1 + 1 + 1 K Mg |
| 20 | N4 K Mg |

change. They used grain and straw yields during 1854 to 1967 from plot 22 and plots 07 + 08 + 13 (i.e., yields totaled over plots 07, 08, and 13). They constructed multiple regression models using the even years, and found that the following *x* variables were required: precipitation October through July, average minimum temperature in July, and maximum temperature in June (for plots 07 + 08 + 13 only). They then assessed the models using data from the odd years, finding that they explained about 33% of variance. Their conclusions were that warm and wet years were disadvantageous but cold and dry years were advantageous for yields.

The key advantage of Broadbalk is that year-to-year changes were minimal, and are all recorded, so it is possible to concentrate just on the influences of climate. Furthermore, local records are available for the climate variables, so there is no need to estimate these from more distant measurements (and thus lose precision). A disadvantage is that the models relate to a few plots at one specific location. The conditions at other locations may not be the same. So any wider use, for example to generate predictions for decision support, might be questionable. Notice, however, that we are not concerned about the fact that the data are unreplicated. In fact it is worth remembering that unreplicated time-series like these are very common (and are used without question) in areas such as economics.

The concerns about the restricted location of the study were addressed in a Ph.D. project by Sabine Landau at Rothamsted and University of Nottingham; see Landau et al. (1998, 1999, 2000). The aim was first to assess the predictive ability of simulation models for winter wheat yields, using results from U.K. trials of autumn-sown fungicide-treated bread-making winter wheat varieties during 1975 to 1993. The project then aimed to identify the most important variables within the models, and use these to produce simplified models that might provide a better predictive framework.

Daily local climate information for each trial location was obtained by spatial interpolation and fed, with the relevant husbandry information, into three widely-used simulation models to generate the predictions. Unfortunately, however, these predictions were found to relate very poorly to the yields that had actually been recorded, with correlations ranging from 0.00 to 0.04 and biases ranging from 20.5 to 77.4%. More fortunately, though, the results did show that the simulation models were capable of predicting stages of growth well. So, a multiple regression model was constructed using *x* variables that were related to weather conditions around the times of various key physiological stages (as predicted by one of the simulation models). The data set was split into three subsets by stratified random sampling. A pool of *x* variables was devised based on the Broadbalk experience and the physiological and agronomic insight of the biological supervisors. The *x* variables for the model were selected from that pool using one-third of the data, the regression coefficients were estimated using a further third, and then the model was assessed by making predictions for the remaining third of the data. A correlation of 0.41 was obtained, which compared well with the correlation of 0.44 obtained for Broadbalk by Chmielewski and Potts (1995). Furthermore, the bias was only 0.078 Mg ha$^{-1}$. The conclusion was that the simulation models were needed to standardize the time-scales within the trials, and that statistical expertise was then needed to generate the predictions.

## Sustainability

In the 1990s there was much concern about the future sustainability of agriculture systems. To validate methods for assessing sustainability, it was important to find good examples of sustainable and unsustainable systems. Thus, in 1993–1994, Rothamsted was among six agricultural institutes worldwide commissioned by the Rockefeller Foundation to make a study of sustainability using data from their long-term experiments. The results from the various groups are presented in a book edited by Barnett et al. (1995), with the Rothamsted findings in Chapter 9 (Barnett et al., 1995). The Rothamsted classical experiments had the advantage of extending back over a very long period, with carefully recorded data on consistently maintained sites. Broadbalk and Park Grass (another classical experiment) represented sustainable systems. Lack of sustainability was illustrated by the Woburn Continuous Wheat experiment (another unreplicated trial) which ran from 1877 to 1926 when it had to be discontinued as a result of the increasing acidity of the soil.

The aim of the study was to go beyond the previous criterion, namely that yields should be maintained, to assess economic considerations and include externalities such as effects on the local environment. For economic sustainability, a range of different measures was studied, all formed as ratios of an index of aggregate output to an index of aggregate input. Each index was calculated as a weighted mean of the various contributing factors. Factors for the input indexes included aspects like costs of fertilizers, labor, machinery, rent, seeds, pesticides, and so on; those for the output indexes included yields of grain and straw. The weights were based on estimated prices—and considerable effort and ingenuity was required to obtain values for these back into the 18th century. The measures of sustainability differed according to whether the means were formed from arithmetic or geometric averages, and from the way in which the weights were constructed.

The Rothamsted study found that, although the arithmetic and geometric indexes were on different scales, they showed a similar basic pattern which was closely related to yield. This probably arose because the weightings of the factors (the *factor shares*) did not change suddenly or dramatically during the experiments. Nevertheless, there were steady changes in shares, and our recommendation was to use an arithmetic index (called A3 in the book) in which the prices were updated every 9 yr and then chainlinked through the successive periods up to the final time. Satisfactory trends in the A3 index would demonstrate that output can be maintained across time. We recommended using the contemporary output–input ratio as well, to ensure that the system is producing output at least comparable with (and preferably in excess of) the inputs.

For externalities, we examined the effect of inflating the costs of inputs such as fertilizers, as might happen if governments were to introduce environmental taxes. In our data the trends in the indexes were very little affected by the inclusion of these effects, although at high proportionate values the farming systems did become unprofitable. In fact, this illustrates another conclusion, found particularly in the data from Oregon State University (Chapter 6), namely that the economic sustainability of agriculture depends very much on the stability of crop prices! Our recommendation here was that one should develop a measure of system health to be maintained as well as the economic indexes.

## Spatial Analysis of Field Experiments

The second major theme in this paper is to describe the recently developed spatial methods for analyzing field experiments, and explain how they enable experiments to be analyzed where treatments may have little or no replication. The methods will be illustrated using analyses from the GenStat statistical system (Payne et al., 2005a; www.genstat.com, verified 8 Sept. 2006) of a variety trial at State Hall farm (Kempton et al., 1994, Gilmour et al., 1997). The design plan, in Fig. 1, has a rectangular layout with 25 treatments (varieties of wheat) arranged in six replicates each containing a 5 × 5 array of plots.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 3 | 5 | 19 | 23 | 2 | 6 | 15 | 18 | 25 | 9 | 11 | 2 |
| 6 | 7 | 9 | 8 | 10 | 8 | 12 | 16 | 25 | 4 | 5 | 7 | 16 | 23 | 14 |
| 21 | 22 | 24 | 23 | 25 | 11 | 20 | 24 | 3 | 7 | 6 | 13 | 22 | 4 | 20 |
| 11 | 12 | 14 | 13 | 15 | 22 | 1 | 10 | 14 | 18 | 24 | 1 | 15 | 17 | 8 |
| 16 | 17 | 19 | 18 | 20 | 5 | 9 | 13 | 17 | 21 | 12 | 19 | 3 | 10 | 21 |
| 3 | 18 | 8 | 13 | 23 | 16 | 24 | 10 | 13 | 2 | 10 | 4 | 17 | 11 | 23 |
| 1 | 16 | 6 | 11 | 21 | 12 | 20 | 1 | 9 | 23 | 12 | 6 | 24 | 18 | 5 |
| 5 | 20 | 10 | 15 | 25 | 4 | 7 | 18 | 21 | 15 | 19 | 13 | 1 | 25 | 7 |
| 2 | 17 | 7 | 12 | 22 | 25 | 3 | 14 | 17 | 6 | 21 | 20 | 8 | 2 | 14 |
| 4 | 19 | 9 | 14 | 24 | 8 | 11 | 22 | 5 | 19 | 3 | 22 | 15 | 9 | 16 |

**Fig. 1. Design of variety trial at Slate Hall Farm.**

To illustrate the advantages of the spatial methods, the experiment is first analyzed conventionally, treating it as a randomized block design. Remember that, in this design, the aim is to group the units (i.e., plots) into *blocks* in such a way that the plots in the same block are more similar than those in different blocks. Each treatment occurs an identical number of times in each block (usually once), and the allocation of treatments is randomized independently within each block. The analysis estimates and removes between-block differences so that treatment effects can be estimated more precisely. This is demonstrated for the Slate Hall Farm experiment by the fact that replicates in the randomized-block analysis have a variance ratio of 7.69 compared with plots within replicates; see Table 2.

The fertility in the field has been modeled here by fitting a single parameter (block-effect) for each block (i.e., replicate), which generates an equal correlation between the plots in each block. Essentially, the analysis models the fertility trends by a step function, and this may not work well when there are many plots within each block. However, the analysis is built on well-established theory (Fisher, 1925, 1926), which can be used with confidence. The issue is our ability to take sufficient account of the underlying variation to estimate the effects precisely. The solution is either to find more sophisticated designs or more sophisticated analyses.

In fact, the Slate Hall Farm experiment was really set up as a row–column design with a structure of rows crossed with columns within each replicate. So, in GenStat terminology, the ANOVA will have the following random terms: replicates (i.e., differences between replicates), replicates.rows (rows within replicates), replicates.columns (columns within replicates), and replicates.rows. columns (the final residual term). The treatments were allocated in such a way that the design was balanced, so that identical amounts of information are available on every comparison between pairs of varieties in the stratum corresponding to each of these error terms. (For a more rigorous definition of balance, see Payne and Tobias, 1992.) In fact, the design is a lattice square. The analysis models the variation more effectively (Table 3). The residual mean square is now 0.8097 compared with 3.466 in randomized-block analysis, and the standard error for differences between variety means is 0.6363, compared with 1.075 for the randomized-block analysis. Furthermore, analysis is still built on established theory. You can use a standard ANOVA table to assess the

**Table 2. Randomized-block ANOVA for Slate Hall farm.**

| Source of variation | df | ss | ms | Variance ratio | *P* |
|---|---|---|---|---|---|
| **Replicates stratum** | **5** | **133.327** | **26.665** | **7.69** | |
| **Replicates.plots stratum†** | | | | | |
| **Variety** | **24** | **254.808** | **10.617** | **3.06** | **<0.001** |
| **Residual** | **120** | **415.976** | **3.466** | | |
| **Total** | **149** | **804.110** | | | |

**† In GenStat terminology, replicates = differences between replicates; replicates.plots = plots within replicates.**

**Table 3. Lattice-square ANOVA for Slate Hall farm.**

| Source of variation† | df | ss | ms | Variance ratio | *P* |
|---|---|---|---|---|---|
| Replicates stratum | 5 | 133.327 | 26.665 | | |
| Replicates.rows stratum | | | | | |
| Variety | 24 | 215.905 | 8.996 | | |
| Replicates.columns stratum | | | | | |
| Variety | 24 | 229.809 | 9.575 | | |
| Replicates.rows.columns stratum | | | | | |
| Variety | 24 | 166.767 | 6.948 | 8.58 | <0.001 |
| Residual | 72 | 58.301 | 0.809 | | |
| Total | 149 | 804.110 | | | |

† In GenStat terminology, the ANOVA will have the following random terms: replicates = differences between replicates; replicates.rows = rows within replicates; replicates.columns = columns within replicates; replicates.rows.columns = the final residual term.

variety term. Also, if you form the variety means using information in replicates.rows.columns stratum only, these can be compared using *t* statistics. Finally, you can also form means that combine information from all strata (Payne and Tobias, 1992). The disadvantage is that balanced designs, analyzable by ANOVA, are not available for most numbers of treatments. Furthermore, spatial analysis may provide a still better representation of the variation, and thus better precision of estimation.

These traditional ANOVA is based on Fisher's three *R*'s (Fisher, 1935): (i) Replication—usually replicate all treatment combinations (improves reliability of their estimates, guards against aberrant plots); (ii) Randomization—guarantees validity of analysis (avoids bias and sensitivity to unrecognized sources of variation); (iii) Blocking (originally called local control)—group similar plots together, and fit a random term to model the differences between the groups (eliminates variability to give more precise estimates of treatment effects).

In contrast, in spatial analysis, sometimes only the control treatments are replicated. You should randomize where possible, but the design may constrain which treatments appear on some of the plots in order to allow good estimates to be obtained for the parameters in the spatial model. You take account of variation by fitting models to describe how the correlation between each plot and its neighbors changes according to their relative locations, and the analysis is by REML [residual (or restricted) maximum likelihood] (see Patterson and Thompson, 1971; Gilmour et al., 1995).

The traditional mixed model is as follows:

$$\mathbf{y} = \mathbf{X}\beta + \textstyle\sum_i \mathbf{Z}_i\mathbf{u}_i + \varepsilon$$

where $\mathbf{y}$ is the vector of observations; $\mathbf{X}$ is the design matrix for the fixed effects; $\beta$ is the vector of fixed effects; $\mathbf{Z}_i$ is the design matrix for random term *i*; $\mathbf{u}_i$ is the vector of effects of random term *i*; and $\varepsilon$ is the vector of residuals. The assumption required for the *F* tests in the ANOVA is that each element of the residual vector $\varepsilon$ follows a normal distribution with mean zero and variance $\sigma^2$. An equivalent way of looking at this, which opens the door to spatial modeling, is that the vector $\varepsilon$ follows a multivariate normal distribution with mean = 0 and variance–covariance matrix $\sigma^2\mathbf{I}$. Likewise, each element of the vector of random effects $\mathbf{u}_i$ is assumed to follow a normal distribution with mean 0 and variance $\gamma_i\sigma^2$, where $\gamma_i$ is the variance component for error term *i*. Or again, equivalently, the vector $\mathbf{u}_i$ follows a multivariate normal distribution with mean = 0 and variance–covariance matrix $\gamma_i\sigma^2\mathbf{I}$.

In correlation modeling, the mixed model is the same, but now the residual vector $\varepsilon$ follows a multivariate normal distribution with mean = 0 and variance $\sigma^2\mathbf{R}$, where $\mathbf{R}$ can be defined using a correlation model (or $\mathbf{R}$ is the identity matrix $\mathbf{I}$ if the effects are independent, as in the traditional model). Similarly, the random effects $\mathbf{u}_i$ follows a multivariate normal distribution with mean = 0 and variance–covariance matrix $\gamma_i\sigma^2\mathbf{G}$, where $\mathbf{G}$ again can be defined using a correlation model. Usually, in field experiments, there are no correlation models for the random effects (so $\mathbf{G}$ is the identity matrix $\mathbf{I}$). For the matrix $\mathbf{R}$, a separable correlation structure, in which the correlation between plots at locations (*i*, *j*) and (*k*, *l*) is defined as the product of a correlation model for rows (relating to rows *i*–*k* apart) and a correlation model for columns (relating to columns *j*–*l* apart), is generally found to be appropriate; see Gilmour et al. (1997). The available models in GenStat for the row or column correlations include autoregressive structures of Order 1 or 2 for designs on a regular grid, and power-distance models for irregularly-spaced designs (which are equivalent to autoregressive structures of Order 1); see Chapter 5 of Payne et al. (2005b) for more details.

In the analysis, the full fixed model should be fitted (here there is just the variety term), while the correlation models are studied. These can be assessed using the deviance, which is defined as −2 times the log-likelihood for the model. Once the appropriate correlation models have been established, the treatment model can be assessed to see whether there are any unnecessary fixed terms by using Wald tests instead of the *F* tests previously available in the ANOVA table. The Wald statistics would have an exact chi-square distribution if the variance parameters were known but, as these must be estimated, the statistics are only asymptotically distributed as chi-square. In practical terms, the chi-square values will be reliable if the residual degrees of freedom for the fixed term is large compared with its own degrees of freedom. Alternatively, Kenward and Roger (1997) show how to modify the chi-square degrees of freedom to counteract the upwards bias of the Wald statistics.

For the Slate Hall data, the best correlation model is one with an AR1 ⊗ AR1 structure for the matrix $\mathbf{R}$, and an additional random term to represent a measurement error from each plot (in the context of Geostatistics, this would be known as the *nugget variance*). See Table 4 for the deviances from GenStat. To improve the efficiency of the calculations, these omit some constants which depend only on the fixed model (however, the interest is only in differences between deviances).

**Table 4. Deviances from the Slate Hall Farm analyses.**

| Design | Deviance | df |
|---|---|---|
| Randomized-complete block | 35.39 | 123 |
| Lattice square | 264.28 | 121 |
| AR1 ⊗ AR1 | 249.35 | 122 |
| AR1 ⊗ AR1 with measurement error | 242.35 | 121 |

The conclusion is therefore that spatial methods can provide better ways for modeling fertility trends than traditional blocking when there are many treatments to assess. Note that traditional blocking factors can be included too to allow for step-changes in fertility (or additional correlation within groups of plots). These might arise from husbandry differences, such as time of harvest or planting, or use of different operators or observers. In the Slate Hall example, however, the replicate differences arose only from fertility trends, and those are described sufficiently effectively by the AR1 $\otimes$ AR1 structure. For more details of the GenStat analyses of the Slate Hall farm data, see Chapter 7 of Payne et al. (2005a).

A further important point is that, if we can model fertility trends effectively, it may be viable to use designs with unreplicated treatments. In variety trials, the standard (or control) varieties are generally replicated to enable the parameters of the spatial model to be estimated effectively. However, the test varieties may have little or no replication because (i) there may be many of them, and (ii) the available seed may be limited. For example, in a *p*-rep design (Cullis et al., 2006), a percentage ($p$) of the designs are replicated while the remainder occur only once. Figure 2 shows the allocation of the replicated varieties in a *p*-rep design grown at Wagga Wagga in New South Wales with 1001 test lines and four standard varieties: 189 of the test lines were replicated twice while the remaining 812 were in single plots (so $p = 18.9$). Three of the standard varieties were on 14 plots each, while the fourth had 16 plots. The 1248 plots were arranged in a 104 row by 12 column array, organized so that replicated test lines occurred once in the top and once in the lower half of the field. Subject to that constraint, the design was formed using A-optimality, assuming an AR1 $\otimes$ AR1 spatial model with random row and column effects, taking $\gamma = 1$ for rows and columns and parameters 0.6 and 0.4 for the AR1 $\otimes$ AR1 structure. The design search, performed by the program Digger, used the modified Tabu search method described by Coombes et al. (2002). The algorithm allocated the replications approximately in a grid arrangement to enable the spatial parameters to be estimated effectively. The use of this design assumes that fertility trends are predicted sufficiently well on the plots occupied by unreplicated test varieties for these to be evaluated correctly. However, there may still be some aberrant plots. The breeder must be prepared to lose some good test lines but, with 1001 test varieties available in the example, this is an acceptable risk. Also, you may accidentally select some bad varieties, but these should be discarded later in the selection process. Thus, unreplicated trials can provide an acceptable means of screening large numbers of varieties in the initial stages of selection, although they should not be relied on for definitive conclusions.

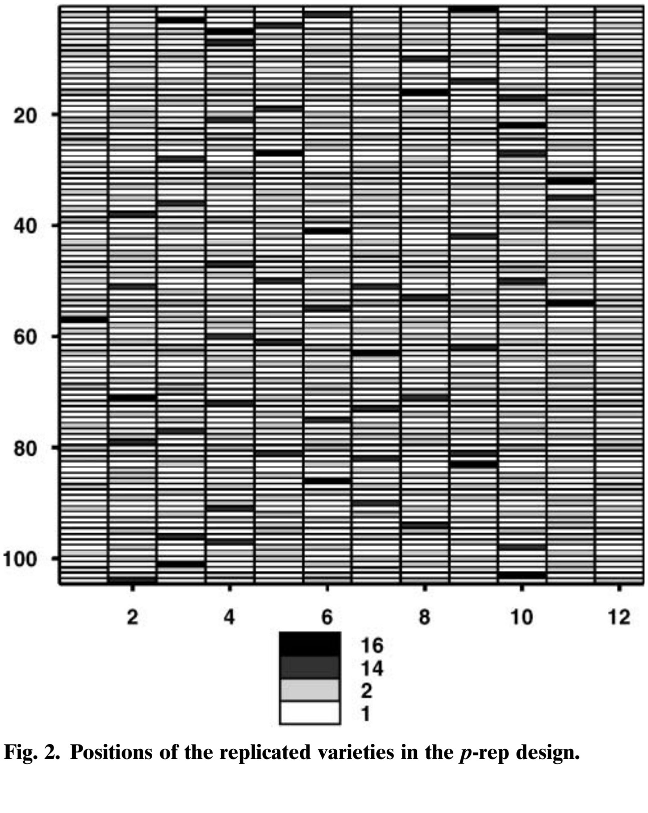

**Fig. 2. Positions of the replicated varieties in the *p*-rep design.**

## Other Methods

Another of Fisher's contributions, relevant to the analysis of unreplicated data, is the randomization test (see Fisher, 1935). For this test, you do the ordinary analysis and calculate the test statistics (e.g., variance ratios or $t$ statistics) in the usual way. You then permute the data values (preferably a large number of times), using the same randomization technique that was used originally to set up the design, and calculate the test statistic(s) for each permuted data set. The randomization test treats the test statistics as a sample from all the possible sets of results that might have been obtained from this particular set of experimental units (i.e., experimental plots if this is a field experiment). The probability value for each test statistic is then determined by seeing where it lies within its distribution over the data sets. The method is useful when the data do not satisfy the distributional assumptions (e.g., of normality) required for the standard analysis. It can also be used if there are few residual degrees of freedom, in fact even if there are no residual degrees of freedom. However, it will not work when there are few observations, as there will then be few permutations; for example, you would need at least 20 permutations for a significance level of 5% to be obtainable. Note, though, that if it is feasible to make all the permutations, you obtain an exact test—another concept established in Fisher (1935). The randomization test is an early example of a method that involves resampling of the data. More recent methods include the jackknife and the bootstrap (e.g., Efron and Tibshirani, 1993).

### ACKNOWLEDGMENTS

This paper is based on a talk presented at the Conference at Salt Lake City in November 2005. The author would like to thank Stephen Machado for the invitation to speak, and Oregon State University and VSN International for the funding to attend. He is grateful also to VSN and the Lawes

Trust for their support for the research, and to Dr Brian Cullis, Dr Alison Smith, and Dr Neil Coombes of New South Wales Agriculture for the example of a *p*-rep design.

## REFERENCES

Barnett, V., A.E. Johnston, S. Landau, R.W. Payne, S.J. Welham, and A.I. Rayner. 1995. Sustainability, the Rothamsted experience. p. 171–206. *In* V. Barnett et al. (ed.) Agricultural sustainability: Economic, environmental and statistical considerations. Wiley, Chichester, U.K.

Barnett, V., R. Payne, and R. Steiner (ed.) 1995. Agricultural sustainability: Economic, environmental and statistical considerations. Wily, Chichester, U.K.

Chmielewski, F.-M., and J.M. Potts. 1995. The relationship between crop yields from an experiment in southern England and long-term climate variations. Agric. For. Meteorol. 73:43–66.

Coombes, N.E., R.W. Payne, and P. Lisboa. 2002. Comparison of nested simulated annealing and reactive tabu search for efficient experimental designs with correlated data. p. 249–254. *In* COMPSTAT 2002. Physica-Verlag, Heidelberg, Germany.

Cullis, B., A. Smith, and N. Coombes. 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11:(in press).

Efron, B., and R. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, Boca Raton, FL.

Fisher, R.A. 1924. The influence of rainfall on the yield of wheat at Rothamsted. Philos. Trans. R. Soc. London. Ser. B 213:89–142.

Fisher, R.A. 1925. Statistical methods for research workers. Oliver and Boyd, Edinburgh, U.K.

Fisher, R.A. 1926. The arrangement of field experiments. J. Min. Agric. 33:503–513.

Fisher, R.A. 1935. The design of experiments. Oliver and Boyd, Edinburgh, U.K.

Gilmour, A.R., B.R. Cullis, and A.P. Verbyla. 1997. Accounting for natural extraneous variation in the analysis of field experiments. J. Agric. Biol. Environ. Stat. 2:269–273.

Gilmour, A.R., R. Thompson, and B.R. Cullis. 1995. Average Information REML, an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51:1440–1450.

Kempton, R.A., J.C. Seraphin, and A.M. Sword. 1994. Statistical analysis of two-dimensional variation in variety trials. J. Agric. Sci. (Cambridge) 122:335–342.

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

Landau, S., R.A.C. Mitchell, V. Barnett, J.J. Colls, J. Craigon, K.L. Moore, and R.W. Payne. 1998. Testing winter wheat simulation models' predictions against observed UK grain yields. Agric. For. Meteorol. 89:85–99.

Landau, S., R.A.C. Mitchell, V. Barnett, J.J. Colls, J. Craigon, K.L. Moore, and R.W. Payne. 2000. A parsimonious, multiple-regression model of wheat yield response to environment. Agric. For. Meteorol. 101:151–166.

Landau, S., R.A.C. Mitchell, V. Barnett, J.J. Colls, J. Craigon, and R.W. Payne. 1999. Response to "Comments on 'Testing winter wheat simulation models predictions against observed UK grain yields by Landau et al. [Agric. For. Metereol. 89 (1998) 85−99]' by Jamieson et al. [Agric. For. Metereol., this issue]". Agric. For. Meteorol. 96:163–164.

Leigh, R.A., and A.E. Johnston (ed.). 1994. Long-term Experiments in Agricultural and Ecological Sciences. CABI, Wallingford, U.K.

Patterson, H.D., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554.

Payne, R.W., S.A. Harding, D.A. Murray, D.M. Soutar, D.B. Baird, S.J. Welham, A.F. Kane, A.R. Gilmour, R. Thompson, R. Webster, and G. Tunnicliffe Wilson. 2005b. The guide to GenStat Release 8, Part 2: Statistics. VSN Int., Oxford.

Payne, R.W., D.A. Murray, S.A. Harding, D.B. Baird, and D.M. Soutar. 2005a. Genstat for Windows. 8th ed. Introduction. VSN Int., Oxford.

Payne, R.W., and R.D. Tobias. 1992. General balance, combination of information, and the analysis of covariance. Scand. J. Stat. 19:3–23.