Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture 2002 - 14th Annual Conference Proceedings

ROW-COLUNIN DESIGNS AT PIONEER HI-BRED

Kevin Wright

ROW-COLUMN DESIGNS AT PIONEER HI-BRED

Kevin Wright
Pioneer Hi-Bred, 7300 NW 62nd Ave, Johnston, IA 50131

Abstract

Pioneer Hi-Bred annually tests more than one hundred thousand new varieties of corn hybrids. Experimental designs used for these tests include CRD, RCB, nested and split-plot. A Six Sigma project at Pioneer recommended testing the use of incomplete-block (IB) designs. In 2001 some experiments were structured as row-column IB designs. This talk will discuss the experience of implementing and analyzing the row-column designs, summaries of the results, and plans for future experiments.

1 Introduction

Plant breeders have a large catalog of experimental designs available for testing new hybrids and varieties. This paper discusses the experience Pioneer Hi-Bred had with investigating, evaluating, and promoting a design type not previously widely used within the company. The paper begins with a brief introduction to Pioneer in section 2. Section 3 discusses the motivation for IB designs at Pioneer and an implementation plan. Sections 4 and 5 present results from analysis of the experiments. Section 6 reviews some of the practical issues with implementing the IB experiment design, and section 7 concludes with final conclusions.

2 Pioneer

Pioneer Hi-Bred International, Inc., a DuPont Company (hereafter referred to as "Pioneer") develops and supplies plant genetics in the form of seeds and additives to customers worldwide. Products include seed for corn, soybeans, wheat, sorghum, canola, alfalfa, millet and also include inoculants for silage. Pioneer is the leading market share in the United States for both corn and soybeans. The company headquarters are in Des Moines, Iowa, with employees worldwide at more than 100 research stations.

The most important product for Pioneer is seed corn, both from a product research perspective and financial income impact. Each year more than 100,000 candidate corn hybrids are tested. These are whittled down until about 20 products are commercialized after five years of testing. Each year, as the number of candidate hybrids is decreasing, the number of locations at which the hybrids are tested is increasing from 1 to more than 50. During the early stages of testing, yield and moisture are among the most important information collected, but as products advance toward commercialization, additional information is collected about germination, insect resistance, disease resistance, grain quality, etc.

The data collected from these experiments are entered into a database, analysis and reporting tool referred to as PRISM, the Pioneer Research Information System. The system is built on a PC architecture with client/server sides that connect various tools. For example, a client can request a data extract from a relational database that is then sent to StatServer (Bajuk, 2000) and S-Plus (Insightful, 2001) for statistical analysis. Standard statistical analysis reports that have been developed for PRISM include analysis of variance, scatter plots, and location quality for CRD, RCB, nested, and split-plot designs. For other types of statistical analysis, the data can be extracted as a text file and analyzed with any appropriate software, for example SAS, S-Plus, and Microsoft Excel. Examples of these customized analyses might include longitudinal/trend analysis, split-split plot experiments, and regression of yield against chemical dosage or irrigation amount.

3 Investigations of row-column designs

Pioneer is one of the strategic business units of DuPont. In 2000, DuPont decided to implement the Six Sigma strategy in all business units, including Pioneer. In the Six Sigma philosophy, small teams choose specific, targeted projects for quality improvement, defect reduction, and cost savings. One such team addressed optimization of early testing resources and improved methods of experimental design. From the results of the team's efforts came recommendations to explore the use of Alpha and other incomplete block designs for early stages of testing (which use few locations). See Kuehl (2000) and Kempton and Fox (1997) for general introductions to incomplete-block designs. Depending on the results of the explorations, additional design types would be added to the list of design types supported in PRISM within the standard analysis and reporting tools. The time-line for the project was given as:

- 2001 Prototyping, randomizations, modeling explorations.
- 2002 Develop PRISM support. Partial implementation.
- 2003 Full implementation in early stages of testing.

In 2001 there were eight Pioneer corn research centers chosen to participate in a test-run of the project. The research centers were chosen to represent geographically diverse areas with five centers in the United States (South, Mid, North), one center in Canada, and two centers in Europe. In total, about 400 experiments were planned, with each experiment containing between 40 and 60 hybrids and being planted at either one location with multiple reps or at three to four locations with one rep at each location. All the experiments were constructed as latinized row-column designs. In this type of design, each block of the experiment was laid out as a grid of rows and columns. The columns were incomplete blocks, as were the rows. When multiple blocks were placed contiguous to each other at a single location, the blocks were abutted so that the columns ran across all the blocks. The latinization ensured that each hybrid occurred at most once in each long column. The experiments that had only one block at each of several different locations were also latinized, since this capability existed and it was felt that this might add a (small) level of additional protection against bias. The ALPHA+ software (Williams and Talbot, 1993) was used to generate the randomizations for the experiments. These randomizations were then uploaded into the PRISM database as a 'custom' design type.

During the 2001 growing and harvest seasons, data was collected regarding stalk count, silage yield, grain yield and moisture, etc. The data was uploaded into the PRISM database, then later extracted according to a precise format for custom analysis. Also during the growing season, a set of S-Plus functions was developed to allow each research center to analyze the data from the experiments grown at that center. The S-Plus functions provided a menu-driven interface to a suite of modeling tools. During this development period, models for RCB, RCB+Row, RCB+Column, and RCB+Row+Column were available. Tools in the software included:

- Level plots of the raw data and residuals from multiple models
- Tables of adjusted means for the hybrid performance under the different models, along with relative efficiencies, Czekanowski coefficients, and confidence intervals for estimates of variance components
- Scatter plots of adjusted means to compare two hybrids

Examples of the output from these features will be demonstrated below. Data from a single experiment could be analyzed with multiple models while data from multiple experiments could be batch processed using a single model. This allowed for an exploration to find the most appropriate model and then to apply that model to all of the experiments.

Following the example of Qiao et al. (2000), the definition of *relative efficiency* for a standard error of a difference between hybrid means is given by

$$RE_{SED} = 100\left(\frac{SED_{RCB}}{SED_{ALT}}\right)$$

where SED_{RCB} and SED_{ALT} denote the average standard error of a difference based on the RCB design or alternative design. Note that this is slightly different than the definition of relative efficiency used by some authors: the ratio of the *variances* of the difference between treatment means.

Another statistic mentioned above for comparing models is the *Czekanowski Coefficient*, also described in Qiao et al. (2000):

$$D = \frac{2a}{2a + b + c}$$

where a is the number of hybrids selected by both models, b is the number of hybrids that are selected only by the row-column analysis and c is the number of hybrids that are selected only by the RCB analysis. Note that when a common proportion of hybrids are selected by the two different methods, then $b = c$ and the Czekanowski Coefficient is the proportion of hybrids that are in common among the selected hybrids under the two different designs. It is not uncommon that experiments with higher relative efficiencies have lower Czekanowski coefficients, indicating that there is a difference in the top hybrids between the two different design types.

Using these statistics, the results of the experiments indicated that some gains in efficiency were realized, enough to justify using incomplete block designs. Some of the analysis results are discussed in more detail in the following sections.

4 Model comparisons at one research center

One of the goals for 2001 was to develop experience with various models for analysis of the data collected from these experiments and to choose the most appropriate model. For this paper, one research center was selected and the 37 experiments grown at this center are analyzed. The data presented includes yield and moisture traits. Each experiment was analyzed using four different models:

- Randomized Complete Block
- RCB + Row effect
- RCB + Column effect
- RCB + Row + Column effects

From each analysis the relative efficiency and Czekanowski coefficients were recorded. The highest median relative efficiency (108%) and lowest median Czekanowski coefficient at the 10% selection intensity (0.83) occurred with the row-column model. Figure 1 shows a parallel plot in which the relative efficiencies for each experiment are linked by straight lines. The efficiencies of the three models are calculated relative to an RCB analysis. Though not immediately obvious from the parallel plot, comparing the median relative efficiency for the three different models indicates that the row-column model (labeled as RangeRow) is more desirable than either a model without rows as incomplete blocks or without columns as incomplete blocks.

There is one experiment that shows a very high relative efficiency. To understand why this should be the case, first consider a the level plot of the raw data in figure 2. Note that row 8 of location 1 is a strip that was not planted and had no data collected. There is a very strong difference between locations with location 3 having an average yield more than 100 bushels/acre lower than the other two locations. Location 3 also has a very clear diagonal trend of increasing yield from lower right to upper left. Locations 1 and 2 show some rows and columns (ranges) that may indicate an effect due to the row or range. It would be expected that analyzing this data using an RCB model would remove the location effect, but not the effects due to rows or ranges. Figure 3 shows the residuals from the RCB model. As expected, the locations now appear to be more similar to each other in terms of the residuals, but there are still fairly strong trends and patches in the plots. A row-column analysis would be expected to reduce the structure in the residual plot. Figure 3 also shows a level plot of residuals from the row-column analysis. As compared to the RCB analysis, the residuals from the row-column analysis are generally smaller and there is almost no discernible spatial structure remaining in the residuals. With these residual plots, the high relative efficiency is understood.

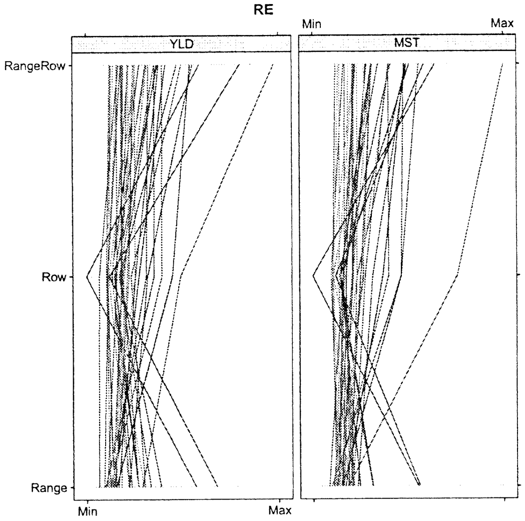

Figure 1: Relative efficiencies for 37 experiments and the traits yield and moisture using 3 different models

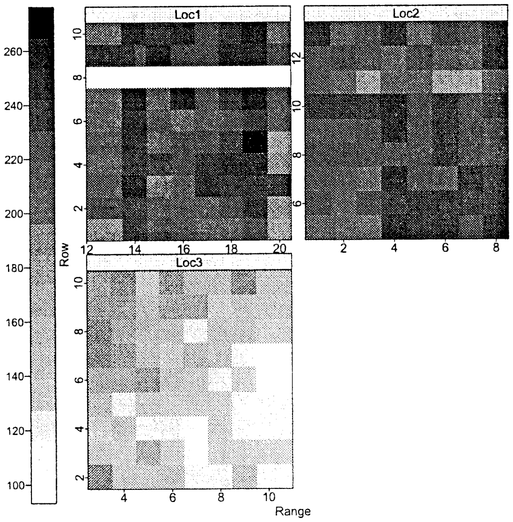

Figure 2: Level plot of yield (bushels/acre) for one experiment conducted at three locations.

5 Comparison of one model across research centers

Based on the analysis of experiments using different models, the full row-column model was chosen as most appropriate. The nested incomplete block row-column analysis is given by Kuehl (2000) as

$$y_{ijlm} = \mu + loc_m + row_{j(m)} + col_{l(m)} + hyb_i + e_{ijlm}$$

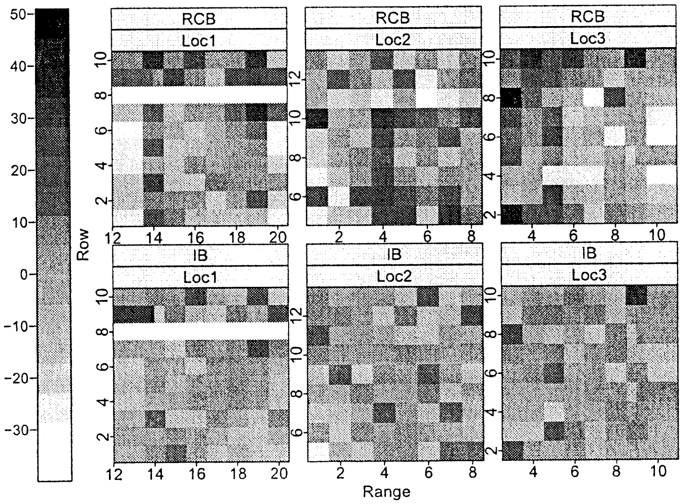

Figure 3: Level plot of residuals from RCB (top row) and row-column IB (bottom row) models.

where μ is the overall mean, $hyb_i(i = 1, \dots, t)$ is the hybrid effect, $loc_m(m = 1, \dots, r)$ is the replicate group, $row_{j(m)}(j = 1, \dots, p)$ is the row effect nested in the replicate group, $col_{l(m)}(l = 1, \dots, q)$ is the column effect nested in the replicate group, and e_{ijlm} is the random error.

Using this as the standard model, data from all 285 experiments at U.S. research stations were analyzed. Figure 4 shows that experiments analyzed with the row-column model typically have a relative efficiency greater than 100% for both moisture and yield as compared to an RCB analysis. Different research stations experienced differing amounts of improvement, with only one station (Station4) showing a median relative efficiency less than 100% for yield. Possible reasons for the difference between stations include plot management techniques, environment differences, and differences in genetic material.

6 Implementation issues

During the planning, implementation and analysis of the incomplete block experiments, a number of practical issues arose and needed to be addressed:

1. Some breeders historically used nested designs for grouping similar hybrids within each rep. These breeders expressed a desire to use nesting groups with incomplete block designs. Due to the restrictions with randomization, nests are not readily available with row-column designs.

2. Row-column designs should have rectangular reps, but there is some flexibility in the shape of the rectangle. The number of rows that can be planted by a machine is an important consideration for the rep size. The number of hybrids in an experiment, the planter size, and the goal of using incomplete blocks that are homogeneous, all are determining factors in the shape of the reps in the experiment.

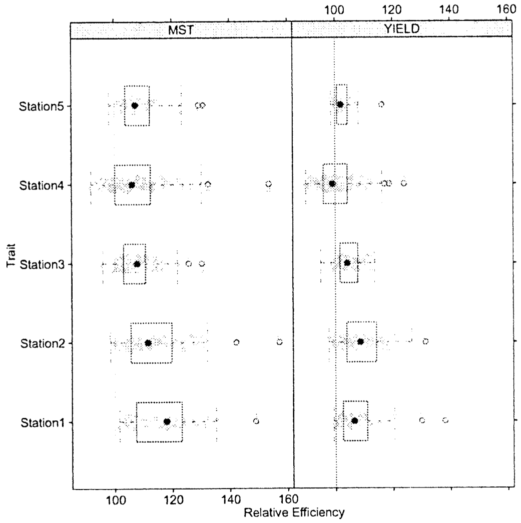

Figure 4: Distributions of relative efficiencies for two traits for all 285 experiments conducted at five U.S. stations. All data is presented as jittered dot plots and is summarized by the box plots.

3. The analysis results from each experiment include the relative efficiency compared to an RCB analysis. If the relative efficiency is less than 100%, it may be preferable to use the results from the RCB analysis.

4. With mixed-effects models, main effects can be specified as random or fixed. For example, it is common to specify hybrids as a fixed effect and to specify effects that are not of interest as random effects. In this case the effects of interest include location, row and column. However, the experiments described above used only three locations and treating location as a random effect may give a poor estimate of the variance component for location. Mixed-effects models are generally more computationally intensive and require longer processing times than fixed-effects models. Furthermore, convergence of the model-fitting algorithm can be less certain with mixed-effects models. For these reasons, it may be better to specify some of the conceptually random effects as fixed effects. One additional feature of mixed-effects models is that when some plots within reps are missing, estimability of effects is not so problematic as in fixed-effects models.

5. RCB models have the nice feature that when no data is missing, the least-squares means are the same as the simple means. IB models are always unbalanced with respect to the incomplete-blocks. Some education was necessary to help researchers understand and feel comfortable with adjusted means as opposed to simple means.

7 Conclusion and future plans

Based on the successful experience with row-column designs in 2001, these designs will be promoted for wider use within Pioneer. During 2002 and 2003 an increasing number of experiments within the early stages of hybrid testing will be set-up as incomplete block designs. In addition to row-column designs, the more flexible alpha designs will be available for use. In the future, the use of incomplete block experiments for crops other than corn will also be explored.

8 Acknowledgments

The material in this paper is presented due to the efforts of many individuals. A partial list of participants in the project includes (in no particular order): Brett Callaway, Mark Hoffbeck, Carlos Loeffler, Mike O'Leary, Arlo Thompson, Gustavo Garcia, Deanne Wright, Jean-Louis Laffont, Mark Cooper, Hui-Rong Qian, Rolland Bagby and Pioneer research center managers. An anonymous referee suggested several changes that led to improvements in this paper.

References

Bajuk, L. (2000). *StatServer 2000 Administrator's Guide.* Data Analysis Products Division, MathSoft.

Insightful (2001). *S-PLUS 6 for Windows User's Guide.* Insightful Corporation.

Kempton, R. A. and Fox, P. N. (1997). *Statistical Methods for Plant Variety Evaluation.* Chapman and Hall.

Kuehl, R. O. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis.* Duxbury.

Qiao, C. G., Basford, K. E., DeLacy, I. H., and Cooper, M. (2000). Evaluation of experimental designs and spatial analyses in wheat breeding trials. *Theor Appl Genet.*

Williams, E. R. and Talbot, M. (1993). *ALPHA+: Experimental designs for variety trials.* CSIRO, Canberra and SASS.