INVENTORY

A0005 ## Design, Performance, and Evaluation of Experiments

V LeMay, University of British Columbia, Vancouver, BC, Canada
A Robinson, University of Idaho, Moscow, ID, USA

LeMay, V
Department of Forest Resources Management
University of British Columbia
2045–2424 Main Mall
Vancouver BC V6T 1Z4
Canada

Robinson, A
Department of Forest Resources
University of Idaho
PO Box 441133
Moscow ID 83843
USA

S0005 ### Introduction

P0005 Experimental design is similar to sampling and inventory design in that information about forest variables is gathered and analyzed. However, experiments presuppose intervention through applying a treatment (an action or absence of an action) to a unit, called the experimental unit. The goal is to obtain results that indicate cause and effect.

P0010 For each experimental unit, measures of the variables of interest (i.e., response or dependent variables) are used to indicate treatment impacts. Replication is the observation of two or more experimental units under identical experimental conditions. A factor is a grouping of related treatments. For example, the factor could be fertilizer, with three levels representing three treatments (e.g., none, a moderate amount, and a heavy amount) applied to plots of trees (plot is the experimental unit). For each plot, height growth measures are taken and averages are compared among the three treatments; the null hypothesis is that there are no differences among the treatment means. The sum of squared differences (termed, sum of squares) between the average for the response variable by treatment versus the average over all experimental units represents the variation attributed to a factor. Experimental error is the measure of variance due to chance causes, among experimental units that received the same treatment. The degrees of freedom, associated with a factor, are the number of treatment levels within the factor minus one. The degrees of freedom for the experimental error relate to the number of experimental units and the number of treatment levels.

P0015 The impacts of treatments on the response variables will be detectable only if the impacts are measurably larger than the variance due to chance causes. To reduce the variability due to causes other than those manipulated by the experimenter, relatively homogeneous experimental units are carefully selected. Random allocation of a treatment to an experimental unit helps insure that the measured results are due to the treatment, and not to another cause. For example, if we have applied the no-fertilizer treatment to experimental units on north-facing sites, whereas moderate and heavy fertilizer treatments are applied only to south-facing sites, we would not know if differences in average height growth were due to the application of fertilization, the orientation of the sites, or both. The results would be confounded and very difficult to interpret.

P0020 Variations in designs, issues that arise, and methods of analyses are discussed in the context of forestry experiments. References from a selection of texts are given; however, there are many books on experimental design. The further reading section also includes more recent advances in analysis of experimental data.

S0010 ### Variations in Experimental Design

S0015 #### Introduction of More than One Factor

P0025 For many forestry experiments, more than one factor is included for design efficiencies over conducting separate experiments. This also allows for examining interactions among factors, and allows for a more efficient design if there are no interactions. A treatment represents a particular combination of levels from each of the factors. For example, if we have two species and three fertilization levels, then we have six treatments that represent the two factors, using a crossed experiment. We might be interested in the effects of species and fertilization, separately, and also whether these interact, resulting in different fertilizer impacts depending upon the species. **Figure 1** illustrates this example using a completely randomized design (CRD), where the treatments are randomly assigned to the experimental units, with factor A (three levels of fertilization: A1, A2, and

A3), factor B (four species: B1, B2, B3, and B4), and four replications per treatment for a total of 48 experimental units.

P0030 If both species and fertilization are fixed effects, in that the experimenter would like to examine the mean response for each species and each fertilizer level, we obtain the analysis of variance table given in **Table 1** from the use of a general linear model and least-squares analysis.

P0035 If the assumptions of general linear models are met, in that residuals are independent, are normally distributed, and have equal variances among treatments, we can interpret the results. The null hypothesis is tested using an *F*-test for each factor and for each interaction. A type I error rate (α, significance level), the chance of rejecting a null hypothesis when it is true, must be selected; we reject an hypothesis if the probability value (*P*-value) for the test is less than the specified significance level. For this example, there is no significant interaction ($P = 0.0539$) using $\alpha = 0.05$; therefore, we can examine species and fertilizer effects separately. There are significant differences between the three fertilizer levels of factor A ($P < 0.0001$), and between the four species of factor B ($P < 0.0001$). The mean values based on these data are: A1 = 16.25, A2 = 23.38, A3 = 28.75, B1 = 17.08, B2 = 20.83, B3 = 24.17, and B4 = 29.08. Further analyses, such as Scheffé's test for multiple comparisons, could then be used to compare and contrast treatment means.

A1B1 = 10	A3B2 = 25	A3B4 = 35	A2B2 = 23	A1B2 = 14	A2B3 = 24
A1B4 = 24	A2B2 = 22	A1B2 = 15	A2B4 = 28	A3B3 = 32	A3B2 = 25
A3B2 = 27	A1B4 = 23	A3B3 = 29	A3B2 = 26	A1B3 = 17	A1B1 = 11
A3B4 = 35	A1B2 = 13	A1B4 = 22	A1B1 = 11	A2B3 = 24	A3B3 = 30
A1B3 = 19	A2B1 = 18	A2B4 = 30	A3B3 = 31	A2B3 = 23	A1B4 = 22
A3B1 = 22	A2B4 = 29	A3B1 = 23	A2B1 = 18	A1B2 = 15	A3B1 = 23
A2B2 = 25	A3B4 = 37	A1B1 = 9	A3B1 = 24	A3B4 = 36	A2B4 = 28
A1B3 = 17	A2B1 = 18	A2B2 = 20	A2B1 = 18	A2B3 = 26	A1B3 = 18

F0005 **Figure 1** Completely randomized design with two fixed-effects factors, randomly allocated to 48 experimental units, with four replications per treatment. For example, A1B1 = 10 indicates that the response variable was 10 for this experimental unit that received factor A, level 1 and factor B, level 1.

P0040 Significant interactions among factors lead to more difficult interpretations, and subsequent analyses must be based a larger group of treatment means. In the example, if the interaction were significant, the 12 means for each fertilizer/species combination would be used in interpretation and subsequent analysis, resulting in fewer experimental units used to calculate each mean value. Since factors often interact in forests, interactions are often detected.

P0045 Issues that may arise in the analysis of this type of experiment include:

1. The assumptions for the residuals are not met.
2. For subsequent analysis, care must be taken to preserve the overall type I error rate.
3. There is difficulty in randomly assigning experiments in field layouts.
4. There are difficulties in inferring results to a larger population. The spatial and temporal scale of forest management is very large, whereas experiments are often small-scale.

These issues are also relevant for other types and variations in experimental design, and are discussed later in this article.

Fixed, Random, or Mixed Effects S0020

P0050 Factors can be fixed, in that the experimenter would like to know the change that is due to the particular treatments applied (as in the CRD example), or random, in that the variance due to the factor is of interest. For example, if the impacts of species (factor) on height growth (response variable) were of interest, we could be interested in the differences among the species in the experiment, and how they rank relative to one another (fixed effect), or we could be interested in the variance in height growth due to species (random effect). Commonly, experiments in forestry include a mixture of factors, some random and some fixed (called mixed effects).

P0055 When factors are random or mixed, the default *F*-tests, as shown in the CRD example, are not appropriate. The expected mean-squares should be calculated in order to determine the correct *F*-tests. Most statistical packages allow the user to request the correct test. Alternatively, maximum-likelihood

T0005 **Table 1** Completely randomized design with two fixed factors: analysis using a general linear model

Source	*Degrees of freedom*	*Sum of squares*	*Mean squares*	F	P
A	2	1258.17	629.08	514.70	<0.0001
B	3	934.75	311.58	254.93	<0.0001
A × B	6	17.00	2.836	2.32	0.0539
Error	36	44.00	1.22		
Total	47	2253.92			

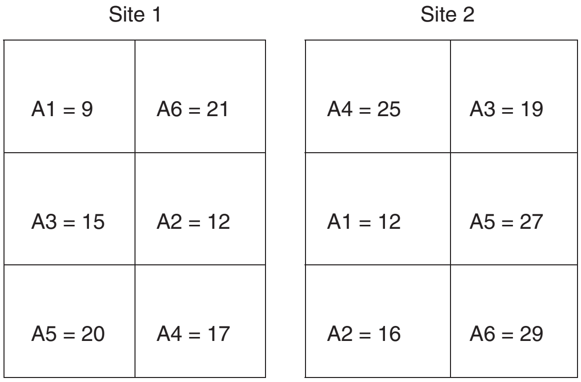
approaches may be more appropriate for mixed-effects experiments. A later section in this article presents more information on least-squares versus maximum-likelihood estimation.

S0025 Restricted Randomization Through Blocking: Randomized Block, Latin Square, and Incomplete Blocks Designs

P0060 Restricting randomization to within blocks is used when the experimental units can be grouped by another variable that may impact the results. In forestry experiments with large experimental units, blocking is often very useful in reducing error variance with only a small reduction in error degrees of freedom. Blocks (or variables that represent blocks, such as trials or sites) are most often random effects. **Figure 2** illustrates a randomized block design (RBD), with factor A (six levels of fertilization: A1 to A6), and two sites. Randomization of factor A is restricted to within sites.

P0065 Using a general linear model with fertilization as a fixed effect and sites as a random effect (mixed-effects model) gives the results in **Table 2**.

P0070 The interest with RBD is with the factor, not with the blocks; the blocks are simply used to reduce the variability among experimental units. For this example, there are significant differences among treatment means ($P = 0.0015$) As with CRD, subsequent comparisons and contrasts could be made among the treatment means.



Site 1

A1 = 9	A6 = 21
A3 = 15	A2 = 12
A5 = 20	A4 = 17

Site 2

A4 = 25	A3 = 19
A1 = 12	A5 = 27
A2 = 16	A6 = 29

F0010 **Figure 2** Randomized block design with one fixed-effect factor randomly located to six experimental units within each of two sites.

P0075 The Latin square design extends grouping of experimental units to two variables. For example, two sites may represent north-versus south-facing stands, and there might be a moisture gradient within sites.

P0080 Another variation is incomplete blocks, where not all treatments are represented in each block. Such blocks are smaller, and, therefore, cheaper, and also subject to less environmental variation, making them quite attractive for forestry applications. Relatively recent technology on the recovery of intrablock information has made the use of incomplete blocks more feasible.

P0085 As well as the issues noted for a multifactor completely randomized design, there is the concern that the blocking may not have been needed. In that case, the introduction of blocks does not result in a corresponding reduction in the experimental error. This should be addressed in the design of the experiment; variables used to group the experimental units into blocks should be those that are expected to affect the response variables.

S0030 Restricted Randomization Through Splitting Experimental Units

P0090 In many multifactor forestry experiments, the experimental unit is split, and different treatments for one factor are applied to the splits, while a single treatment from another factor is applied to the unit. For example, with six treatments representing three fertilizers and two species, we could use six small experimental units and randomly assign the six treatments to these units. However, this might result in an experimental unit that is too small for the mechanical application of fertilizer. An alternative is to apply the fertilizer treatments to three larger experimental units, and then split each unit and randomly assign the species to the split units (called split plots). This is a restriction on randomization. A further extension of this would be to split the units again, and randomly assign a third factor (e.g., particular seedling stocks for a species) to the smallest unit, resulting in split-split plots.

T0010 **Table 2** Randomized block design with one factor, randomly located with each of two blocks: analysis using a general linear model

Source	*Degrees of freedom*	*Sum of squares*	*Mean squares*	F	P
Block	1	96.33	96.33	38.03	0.0016
Fertilization	5	320.00	64.00	25.26	0.0015
Error	5	12.67	2.53		
Total	11	429.00			

P0095 Although the analysis of an experiment using split or split-split plots is very similar to a multifactor experiment where there is complete randomization of treatments to each unit, care must be taken in using the correct experimental error for the units versus the subunits, and interpreting the results.

S0035 Nesting of Factors

P0100 Treatment levels for one factor may be particular to the level of another factor, resulting in nesting of treatments. For example, for the first level of fertilizer, we might use medium and heavy thinning, whereas, for the second level of fertilizer, we might use no thinning and light thinning.

P0105 Nesting of factors will affect both the analysis and the subsequent interpretation of the experiment. An example of a nested design is given in **Figure 3**, with the subsequent analysis in **Table 3**.

P0110 When factors are nested, it is not possible to isolate the nested factor from the other factors, nor is it possible to assess interactions between nested and nonnested factors. The correct *F*-tests differ from a crossed experiment, in that the error mean-squares is not used for all *F*-tests. For factor A, there were no significant differences between the treatment means ($P = 0.1172$), using the mean-squares for factor B, nested in A for the *F*-test. The means for factor B, nested in A, were significantly different ($P < 0.0001$) using the error means-squares for the *F*-test.

A1B1 = 10	A1B1 = 11	A1B2 = 13	A2B4 = 23
A1B2 = 15	A2B3 = 18	A2B4 = 25	A1B1 = 11
A2B4 = 20	A2B3 = 18	A1B1 = 9	A2B3 = 18
A2B4 = 22	A1B2 = 15	A2B3 = 18	A1B2 = 14

F0015 **Figure 3** Nested design with two factors, where the second factor is nested in the first factor, with four replications per treatment.

P0115 Interpreting nested designs is more complicated than crossed designs. However, nesting may result in efficiencies by reducing the number of experimental units over the number that would be needed for a crossed experiment. Also, nested factors result from a hierarchical design, which is discussed next.

S0040 Hierarchical Designs and Subsampling

P0120 Commonly in forestry experiments, the experimental unit represents a group of items that we measure. For example, an experiment includes several pots in a greenhouse, each with several plants germinating from seeds. A treatment (specific level of factor A) is randomly allocated to each pot (could be more than one factor, fixed and/or random), even though measures are to be taken on plants. The three factors, which affect the measures on plants, are factor A, pots, and plants. The pots are nested within factor A treatment levels, since pot 1 receiving treatment 1 is not the same treatment as pot 7 receiving treatment 1. Similarly, plants in a pot are nested within pots and factor A treatment levels. The three factors are not all crossed in this hierarchical design; some factors are nested.

P0125 A variation on hierarchical designs is measuring a sample of items, instead of measuring all items in an experimental unit. For example, if we have 50 trees in an experimental unit, we may choose to measure only 10 of them for diameter growth.

P0130 The analysis of hierarchical designs differs from an experiment with fully crossed factors. All levels in the hierarchy must be included in the analysis. Since lower levels in the hierarchy are often random-effects factors, hierarchical models are commonly mixed-effects models. Although methods for least-squares analysis have been developed, maximum-likelihood estimators for mixed-effects models may be more appropriate, as discussed later.

S0045 Introduction of Covariates

P0135 The initial conditions for an experiment may not be the same for all experimental units, even if blocking is used to group the units. Site measures such as soil moisture and temperature, and starting conditions for individuals such as starting height, are then

T0015 **Table 3** Nested design with two fixed-effects factors, where the second factor is nested in the first factor: analysis using a general linear model

Source	*Degrees of freedom*	*Sum of squares*	*Mean squares*	*F*	*P*
A	1	256.00	256.00	7.06	0.1172
B (A)	2	72.50	36.25	23.51	<0.0001
Error	12	18.50	1.54		
Total	15	347.00			

measured (called covariates) along with the response variable, and these covariates are used to reduce the experimental error. Covariates are usually interval or ratio scale (continuous).

S0050 Issues Arising in Forestry Experiments

S0055 Failure to Meet Assumptions

P0140 When the usual assumptions of the least-squares method are not met, usual *F*-tests may not be reliable. Transformations of the response variables are commonly used, often requiring a "trial-and-error" approach until the residuals do meet the assumptions. However, results for the transformed response variable are more difficult to interpret, as mean values do not relate well to the original measurement scale. This is particularly true if a nonparametric analysis via a ranking the response variable (called rank transformation) is used. Alternatively, generalized linear models can be used if the residuals appear to follow a distribution from the exponential family, including binomial, poisson, or gamma distributions. For temporally related data, repeated measures analysis is commonly used. Analysis for spatially correlated data can be more difficult, since data can be correlated in many directions.

S0060 Preservation of Overall Error Rate in Subsequent Analyses

P0145 The use of a particular type I error rate to test for differences among treatment means within a factor should be preserved in subsequent analyses. For example, if an *F*-test is used with a type I error rate of 0.05, appropriate subsequent pairwise tests should use the type I error rate of 0.05 over all tests.

S0065 Difficulty in Randomly Allocating One or More Treatments

P0150 Although randomizing the allocation of treatments to experimental units is fundamental to removing confounding of treatments with other impacts, sometimes randomization of all treatments is not possible. For example, the impact of burning as a site preparation method prior to planting is difficult to randomize; burning may necessarily need to be confined to one side of the experimental area, resulting in a restriction in randomization. As noted, the results are then subject to confounding, since there may be another factor in the burned area that influences the response. Experimenters often use the analysis appropriate for unrestricted randomization; however, caution must be used in interpreting results.

Missing Information S0070

For some circumstances, particular combinations of factors may be missing, because of a lack of experimental units, because some of the experimental units are damaged, or because of the nature of the treatments. For example, all trees with the high fertilizer, species 1, die because of a failure in one section of a greenhouse sprinkler system. Analysis of the experiment as a nested experiment may be possible, allowing for different levels of one factor within a level of another factor. Imputation methods may be used to find estimates for missing data. However, at some point, statements of statistical inference may not be possible, if too much of the experimental data is missing. P0155

Size of Experimental Units and Time Scale S0075

For studies of young trees and plants, experimental units can be relatively small, and may be conducted in greenhouses with many experimental units. For larger trees, large experimental units are needed to reflect the scale of processes impacting growth. Difficulties arise in finding homogeneous units. As a result, the number of experimental units is often small, resulting in low power. This becomes more pronounced in studying wildlife habitat and watershed processes, where the scale of some processes is even larger. For these very large-scale processes, often a number of case studies are conducted. Results are more difficult to interpret, since unknown or known confounding may have occurred. P0160

Another complication of forestry experiments is that long time scales are often needed to study forest changes meaningfully. As a result, missing information is more common, measurement standards may change over time, and measures might not be taken at regular time intervals, due to changes in funding. These long-term experiments are difficult to analyze and interpret. Models and graphs are commonly used to interpret trends. P0165

Inferences Made from Experimental Results S0080

Since the aim of experimental design is that results indicate cause and effect, experimental units are carefully selected for homogeneity. Results of experiments can, therefore, be somewhat artificial, since the usual heterogeneity of the biological system has been removed from the experiment. Often researchers include observational studies to attempt to model the biological system, and experimental components to isolate causes and effects. The results of the two types of studies are then combined for a more thorough interpretation. P0170

S0085 ### Power of Experiments

P0175 The power of a test is the ability to reject a null hypothesis when it is false. An experiment may have too little power to detect an important difference among treatment means, or conversely, too much power, resulting in detection of significant differences that are of no practical importance.

P0180 The ability to detect differences between treatment means increases as the size of the experiment increases, where size is defined as the number of replicates and the number of treatments. Power analysis is the assessment of the power of test for the planned experiment, given the size of differences that have practical importance, and an estimate of the expected variation.

P0185 The method of determining the size of the difference that will be detected by an experiment will vary with the design of the experiment. For example, if a randomized block design is used, then more experimental units per block could be used to increase power (sometimes called generalized randomized block design), or more blocks could be established. For split-plot experiments, the power for the factor assigned to the split plot (subunit) is higher than for the factor randomly assigned to the whole plot (experimental unit). Careful design of the experiment should allow for varying sizes of differences for different factors. If power analysis is done following the experiment, the correct analysis given the experimental design must be followed.

S0090 ## Least-Squares versus Maximum-Likelihood Estimation

P0190 Many forestry experiments require and benefit from a mixture of fixed and random effects. These different effect types simplify the analysis of hierarchical designs as well as correlations in time and space. Analysis of these models using least-squares techniques can be complicated. Analysis using maximum-likelihood estimators and their variants (restricted maximum-likelihood estimators), is often much more straightforward and flexible. Furthermore, the statistical properties of the maximum-likelihood estimation-style estimators can be superior.

P0195 Although maximum-likelihood estimation allows for greater model flexibility, it requires a search algorithm to find a global maximum (overall maximum), unlike generalized least-squares models. For very complex models, only a local maximum may be found, or there may be no convergence. Many statistical packages have built-in procedures for mixed-linear or nonlinear models, allowing for easier application of these relatively new procedures.

Overall Considerations in Designing and Analyzing Forestry Experiments S0095

In order to obtain results that can be interpreted with little or no confounding, experimental units should be carefully selected to remove factors that are not of interest to the experimenter, but would affect the variables of interest. Random allocation of treatments is also needed to equalize the impacts of any remaining factors that were not removed through careful selection. Identifying factors as fixed versus random and using the appropriate design is essential to correct interpretation of results. Also, the correct analysis of hierarchical designs should be stressed; incorrect analyses sometimes appear in literature. For least-squares analysis, expected mean-squares should be calculated to determine appropriate *F*-tests. Power analysis is strongly recommended, during the design of the experiment, to ensure that statistically significant results indicate differences of practical importance. P0200

Because of the large time and spatial scale of many forest processes, experimental units often are large and long-term, in order to have meaningful results. This leads to problems with traditional designs, in that experimental units are large and very heterogeneous, and some are lost over time. Also, there is low power as there are few experimental units. Assumptions of least-squares analysis are commonly not met, resulting in difficulties in analysis and interpretation. P0205

New technologies using maximum-likelihood methods allow greater variability in the analysis of data. These methods have improved our ability to conduct analyses when the assumptions of least-squares analysis are not met, and have increased the flexibility in the design of forestry experiments. P0210

List of Technical Nomenclature S0100

See also: **Forest Environment**: Environmental Impacts (00113). **Forest Health**: Diagnosis, Monitoring and Evaluation (00109). **Forest Products**: Biological Improvement of Wood Properties (00038); Effect of Growth Conditions on Wood Properties (00039). **Forest Recreation**: Inventory, Monitoring and Management - Concepts, Approaches and Methods (00165). **Human Influences on Tropical Forest Wildlife** (00012). **Inventory**: Biometric Research (00151); Forest Inventory and Monitoring (00154); Modeling (00147); Spatial Information (00160);

Statistical Methods (Mathematics and Computers) (00152); Yield Tables, Forecasting, Modelling and Simulation (00148). **Silviculture**: Species Choice (00213); Stand Establishment, Treatment and Promotion - European Experience (00224). **Soils and Site**: Soil Contamination and Amelioration (00244).

Further Reading

Box GEP and Cox DR (1964) An analysis of transformations. *Journal of the Royal Statistical Society Series B* 26: 211–252.

Cochran WG and Cox GM (1957) *Experimental Designs*. New York: John Wiley.

Cressie NAC (1993) *Statistics for Spatial Data*. revd. edn. Toronto, Canada: John Wiley.

Hurlbert SH (1984) Pseudoreplication and the design of ecological experiments. *Ecological Monographs* 54(2): 187–211.

John JA and Williams ER (1995) *Cyclic and Computer Generated Designs*. London: Chapman & Hall.

Kirk RE (1982) *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, CA: Brooks/Cole.

McCullagh P and Nelder JA (1991) *Generalized Linear Models*. New York: Chapman & Hall.

Meredith MP and Stehman SV (1991) Repeated measures experiments in forestry: focus on analysis of response curves. *Canadian Journal of Forestry Research* 21: 957–965.

Neter J, Kutner M H, Nachtsheim C J, and Wasserman W (1996) *Applied Linear Statistical Models*, 4th edn.

Schabenberger O and Pierce FJ (2002) *Contemporary Statistical Models for the Plant and Soil Sciences*. New York: CRC Press.

Scheffé H (1959) *The Analysis of Variance*. Toronto, Canada: John Wiley.

Sheskin DJ (1997) *Handbook of Parametric and Nonparametric Statistical Procedures*. New York: CRC Press.