Available online at www.sciencedirect.com
ScienceDirect
Forest Ecology and Management 255 (2008) 2022–2023
Forest Ecology and Management
www.elsevier.com/locate/foreco

Editorial

Three key points in the design of forest experiments

Unfortunately, the editors of *Forest Ecology and Management* reject dozens of manuscripts each year because of fundamental design flaws that undermine the value these projects might hold for readers. We would like to suggest three basic guidelines to avoid some too-common pitfalls in the design of research in forest ecology and management. Authors should:

1. Clearly state the population of interest, and how the experimental findings will be extrapolated to the population.
2. Based on the population of interest, be clear on what is a true replicate, and what is a subsample.
3. Analyze quantitative treatments quantitatively and not as separate categories.

An example of the one of the most commonly used experimental designs can illustrate these guidelines. A typical forest fertilization experiment might use 4 replicated blocks within a single stand of red pine on sandy glacial outwash near Toronto, Canada, with four levels of nitrogen fertilizer added to randomly assigned plots in each block. The attention of the researchers could focus on choosing which levels of fertilizer to add (100 kg/ha? or 200 kg/ha?), and the measurements to use for characterizing the response to fertilization (diameter, height, volume, mortality?). In some cases, the post-treatment response might be more precise if pre-treatment differences among plots were considered (such as basal area, or growth rates within plots). We stress that all of these questions need to be preceded by a clear definition of the population the researchers plan to address with the experimental results.

The design of this experiment may be strong for inferring the response of trees in this single stand to fertilization during a single period in time. In fact, this design would be suitable for determining the response of all red pine forests to fertilization, if the response of all forests to fertilization was not affected by factors such as soils, microclimate, and management. The design would be very weak, however, in providing insights about growth responses of any other stands because these other factors dramatically confound the response to fertilization. Even if the researchers were interested only in the fertilizer response of all red pine stands that occur on sandy glacial outwash near Toronto, this design would provide 0 degrees of freedom for understanding the variability among sites within this narrowly defined population (with 1 site, $n - 1 = 0$). A similarly poor design might aim to determine whether the response to fertilization would differ between sandy outwash and glacial-till derived soils by placing a well-replicated trial in a single stand on each soil type. Perhaps the response would be significantly higher in the stand on the rocky till soil, but this design would have no statistical power to test whether the response would differ between the population of sandy soils and the population of rocky soils.

If the population for this experiment were defined as "this red pine stand," then the block design would indeed have 4 true replicates. If the population of interest was "red pine stands on sandy soils near Toronto," the design would have a single replicate (one stand) with an estimated responses based on 4 subsamples (actually, split subsamples) of the single replicate. A much better design would omit any replication of treatments within a single stand, and apportion the experimental work across 4 independent stands within the population of interest. The dispersed approached would have 3 degrees of freedom for interpreting the likely responsiveness of the whole population, despite having 0 degrees of freedom for interpreting the response within any single stand. Replicate subsamples within a single stand are only useful if they improve the representativeness of the response for that stand. Efforts invested in more true replicates are (almost) always more profitable than the same effort applied to obtaining more subsamples that contribute no degrees of freedom for extrapolating to the population.

Would this classic 4-replicate-block design (within a single site) ever be useful for forest research? The answer always depends on the population of interest, and how one might want to extrapolate the experimental findings to the population. A scientist may want to know how variable the response to fertilization is within a single stand, and the classic design would provide an answer that could be extrapolated across this single stand. More commonly, the results from a replicated study at a single site might be extrapolated to a population of sites based on an inference about the processes (or mechanisms) behind the response. For example, this experiment could be used to examine the relationship between the magnitude of growth response and the change in canopy leaf area. The results might show a linear relationship between increases in stand area increase and volume growth, and the scientists might infer that this process relationship would apply across all red pine stands of this general type. Inferences in science can be based on

0378-1127/$ – see front matter
doi:10.1016/j.foreco.2008.01.047

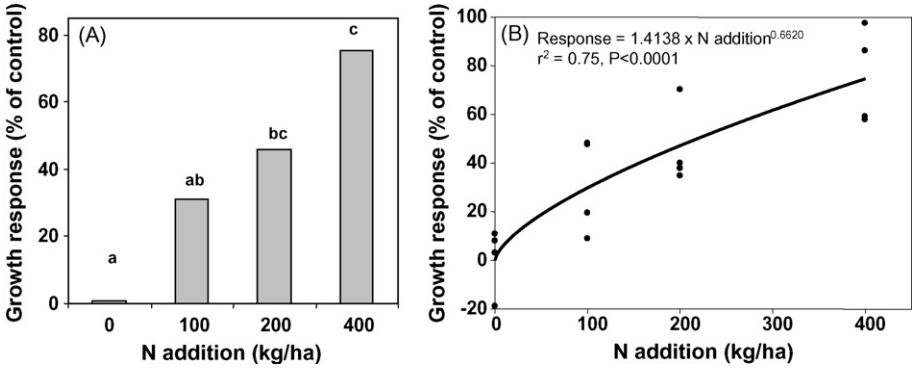

Fig. 1. (A) Analyzing the fertilization experiment with an analysis of variance and multiple range test (unplanned comparison of all pairs of means) showed that the 200 and 400 kg/ha treatments differed from the control (bars with the same letter do not differ at $P < 0.05$), and the 400 kg/ha treatment differed from the 100 kg/ha treatment. (B) The stronger curve-fitting analysis takes advantage of the quantitative relationship inherent in the levels of the single type of treatment (N addition) to support stronger inferences about the relationship between treatment level and response.

professional judgment as well as on blind statistical extrapolations, but authors need to describe clearly the approach for leaping from experimental results to the population of interest.

The third guideline is illustrated in Fig. 1 with a comparison of two methods that might be used to analyze the same data from the hypothetical fertilization experiment. The weak approach (Fig. 1A) analyzed treatment effects with an analysis of variance, demonstrating a significant effect for fertilization ($P = 0.001$). The ANOVA was followed by Tukey's Honestly Significantly Different test of all pairs of means. The statistical analysis indicated that the 200 and 400 kg/ha treatments differed from the control, and the 400 kg/ha treatment differed from the 100 kg/ha treatment; other pairs were not significantly different at the 0.05 (or even 0.10) level of confidence. The ANOVA with means comparisons treated the fertilizer levels as though they were different types (categories) of treatments, rather than quantitatively scaled levels of a single type of treatment. The degrees of freedom in the means comparison were dispersed evenly (blindly) across all possible pairs of means, and this is weak for at least two reasons. Some pairs of means might be very interesting to compare (such as 100 kg N/ha vs. 200 kg N/ha), but others might be less interesting (such as 100 kg N/ha vs. 400 kg/ha). The researchers may also have a prior expectation that a higher rate is likely to provide a higher response (not a lower response), so the power of the tests would be stronger if *a priori* expectations are stated that allow use of 1-tailed distributions rather than 2-tailed. In many cases, the power of this sort of analysis can be increased if a subset of all possible comparisons is chosen in advance.

A stronger (and more appropriate) analysis is shown in Fig. 1B. This approach uses the quantitative information present in the treatments; 100 kg/ha is quantitatively smaller than 200 kg/ha, which is smaller than 400 kg/ha. About 75% of the variation among plots in growth response to fertilization was explained by the amount of fertilizer added ($P < 0.0001$). The relationship shows that confidence is warranted in the assertion that increasing levels of fertilizer (up to 400 kg/ha) are likely to increase growth.

How confident should a manager be that 200 kg N/ha will increase growth more than 100 kg N/ha in this stand? Neither of these statistical analyses answered this question, though the design does provide information that can be explored with other methods for decision making under uncertainty. Consulting with statistician during the design (and analysis) of experiments is always a productive idea. We also note that the analyses in Fig. 1 remain handicapped by the original design of the experiment; nesting all the treatments within a single stand means that the analysis of fertilizer response applies just to this single stand. Other stands may or may not show similar patterns; the design simply allows no statistical evaluation of the likelihood of responses in other stands.

We hope that our three guidelines will help contributors to *Forest Ecology and Management* to design stronger research projects, and lead to success in the review process for publication.

Dan Binkley*
Colorado Forest Restoration Institute,
Warner College of Natural Resources,
Colorado State University, Fort Collins, CO 80523, USA

*Tel.: +1 970 491 6519; fax: +1 970 491 6754
E-mail address: Dan.Binkley@Colostate.edu