THE STATISTICAL REASON WHY SOME RESEARCHERS SAY SOME SILVICULTURAL TREATMENTS "WASH-OUT" OVER TIME

David B. South and Curtis L. VanderSchaaf[1]

Abstract—The initial effects of a silvicultural treatment on height or volume growth sometimes decline over time, and the early gains eventually disappear with very long rotations. However, in some reports initial gains are maintained until harvest but due to statistical analyses, a researcher might conclude the treatment effect has "washed-out" by ages 10 to 18 years (even when the gain is 6 green tons per acre or more). This claim is sometimes made even when the volume gains have increased over time. Researchers who end up making Type II statistical errors do so because of the inherent variability which increases with stand age (i.e., statistical power declines over time). To avoid making Type II errors, some researchers have decided to model their data instead of applying statistics to a data set that has low power.

INTRODUCTION

"In the past, the emphasis of statistics in forestry, and other applied fields, has been on an assessment of statistical significance, or the probability that the null hypothesis will be rejected when it is true (i.e., the probability of committing a 'Type I error'). However, there is growing awareness (e.g., see Peterman 1990a, 1990b; Toft and Shea 1983) that researchers should also be concerned with the possibility that statistical methods may fail to reject a false null hypothesis (i.e., a 'Type II error' might be committed). The statistical theory and methods by which this important issue can be examined are referred to as 'power analysis' (Nemec 1991)". Power is the probability of getting a statistically significant response when a real treatment difference exists. In some cases, a lack of power explains why some researchers say a treatment will "wash-out" over time.

Sometimes landowners are told that certain silvicultural treatments are not worth pursuing since they will "wash-out" over time. We have discovered there are two definitions for the phrase wash-out (fig. 1). Definition No. 1 states that a wash-out occurs when: the absolute treatment response declines over time and, prior to harvest, the treated stand ends up with the same overall stand characteristics as the untreated stand. As a result, there are no absolute differences in stand volume between the untreated and treated plots. The second definition states that a wash-out occurs when there is an absolute gain, but the difference between treatment means is not statistically significant. As a result, some researchers say a treatment has washed-out even when field foresters realize there is a substantial economic difference in volume (e.g., 5 green tons per acre or more). When the gain is indeed a result of the treatment, then a Type II error occurs if the researcher claims the treatment effect has disappeared or has washed-out.

Several researchers have said treatment differences "disappeared" as the stand aged even though the absolute differences were substantial. For example, in 10 loblolly pine studies, the absolute volume gain from applying herbicides increased over a 3-year interval from 45 to 84 cubic feet per acre. Even so, the researchers concluded that "early growth responses declined between ages 5 and 8 years…" In another article, the statement was made that treatment "differences had disappeared upon re-examination at age 31." However, the power of the statistical test at age 31 years was so low that it could not detect a 21 percent increase in merchantable volume as significant ($\alpha = 0.05$). This difference was equal to an increase of 270 cubic feet per acre (or approximately a 5-year advance in stand development). If this experimental design could not detect a 5-year gain as statistically significant, we wonder at what stand age would a researcher be unable to declare a 1- or 3-year gain as statistically significant?

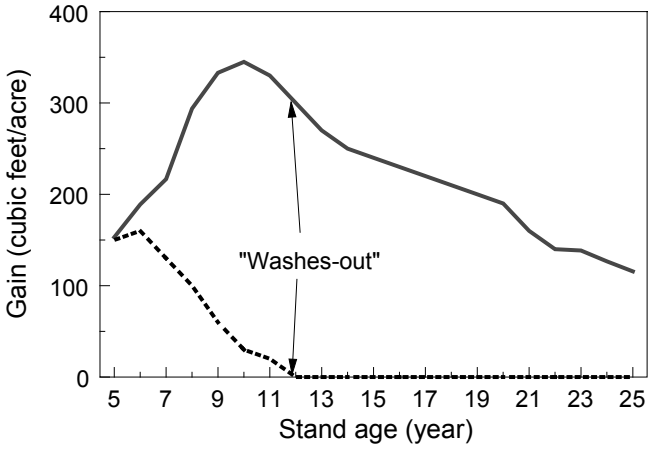

Figure 1—The effect of stand age on volume gain due to two silvicultural treatments. In one case (dashed line), a field forester says the treatment "washes-out" at age 12 years when there is no absolute difference in stand volume between treatments. In the second case (solid line), a field forester says there is a gain in volume, but the researcher says the treatment "washes-out" when the difference in volumes is no longer statistically significant ($\alpha = 0.05$). In this case, the difference between treatments at 15 years is equal to 240 cubic feet per acre.

To address this question, we conducted a-priori power tests in preparation for a "year of planting" study. The objective was to determine at what stand ages a Type II error would occur. We could identify a "true" Type II error since the null hypothesis (i.e., there is no effect of planting date on stand growth) in our case was always wrong.

[1] Professor, School of Forestry and Wildlife Sciences and Alabama Agricultural Experiment Station, Auburn University, Auburn, AL 36849-5418; and Graduate Student, Virginia Polytechnic Institute and State University, Department of Forestry, Blacksburg, VA 24061, respectively.

Citation for proceedings: Connor, Kristina F., ed. 2006. Proceedings of the 13th biennial southern silvicultural research conference. Gen. Tech. Rep. SRS–92. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 640 p.

METHODS

Data were obtained from a spacing study established by the Virginia Polytechnic Institute and State University Loblolly Pine Growth and Yield Research Cooperative. The site was located on an Upper Atlantic Coastal Plain site near Roanoke Rapids, NC (Zhang and others 1996). Treatment plots were replicated three times. A spacing of 8 feet by 12 feet was selected for the untreated control. For the ages used in this paper, data from this site has had little or no impacts from windthrow, hurricanes, ice storms, beetles, etc. Thus, all variation for a particular dependent variable among replications was due to growth variation among planted trees, measurement error, and environmental variation. The variance values exhibited by these plots are reasonable approximations of the error that would be expected in non-damaged loblolly pine plantations in this region.

Four stand-level variables were examined: quadratic mean diameter (QMD), average height, basal area per acre, and total cubic foot volume per acre. Individual tree diameter at breast height (d.b.h.) was measured on all trees within a replication for stand ages of 5 to 21; total tree height (Ht) was measured annually from ages 2 to 10 and then every other year from 10 to 20 years. Individual tree total stem cubic-foot volume (stump to tip, outside-bark) was estimated using an equation developed by Tasissa and others (1997):

$$\text{cubic feet} = 0.21949+0.00238dbh^2Ht \quad (1)$$

where

cubic feet = total cubic foot volume (outside-bark) per tree.

Since d.b.h. and height are independent variables in equation (1), volume was estimated from actual data only for the ages that are common between the two predictors (e.g., ages 2 to 10, 12, 14, 16, 18, 20).

Stand-level values were obtained for each of the three replications by appropriate expansion factors. For ages 11, 13, 15, 17, and 19, stand-level average height and total cubic foot volume were interpolated. For plantation ages of 21 to 28, Chapman-Richards equations (Richards 1959) were developed separately for each of the three replications to estimate stand-level average height and total cubic foot volume (adjusted R^2s ranged from 0.9920 to 0.9985). For QMD and basal area per acre, Chapman-Richards equations were developed by replication to estimate yield for ages 22 to 28 (adjusted R^2s ranged from 0.9903 to 0.9977). Estimates were checked to insure that they reasonably extrapolated beyond the range of the actual data.

For each age, we calculated the mean and standard deviation of the three replications for each of the four stand-level variables. Three potential age-shifts were then calculated; a 1-year age shift in stand development, a 2-year age shift, and a 3-year age shift. In our study, a 1-year age shift (i.e., treated mean) was equivalent to planting seedlings 1 year earlier than seedlings in the control plots (i.e., untreated mean).

Two-sided, two-sample independent t-tests were conducted (given age constraints of the dataset for a particular variable as described above) to test for significant differences due to the age-shifts. Two α levels were used: $\alpha = 0.05$ and $\alpha = 0.10$. The power of all tests conducted was then found using a formula provided by Kirk (1995):

$$d(n-1)\text{square root}(2n) \quad (2)$$

$$2(n-1)+1.21(z_a-1.06) \quad (3)$$

$$[\text{equation (2)}/\text{equation (3)}] - z_a \quad (4)$$

where

d = [treated mean – untreated mean]/standard deviation,

n = number of replications per treatment,

z_a = z-score for a particular two-sided α level (either 1.96 for $\alpha = 0.05$ or 1.645 for $\alpha = 0.10$), and 1.06 and 1.21 are constants regardless of n or the desired α level

RESULTS AND DISCUSSION

The effect of "planting year" on height growth of loblolly pine is illustrated in figure 2. Planting year is equivalent to what some researchers call an "age-shift" (VanderSchaaf and South 2004). A 2-year age-shift in stand development can be thought of as planting a site 2 years earlier than the control. Although we know these hypothetical stands were planted in different years, after year 17 there were no statistically significant differences ($\alpha = 0.05$) in height (fig. 2), volume (fig. 3), or basal area (fig. 4). In contrast, differences in quadratic mean d.b.h. could still be detected at age 22 years (fig. 5). However, after age 26 years, there no longer was a significant difference (even when there was a 3-year age difference). Overall, Type II errors started to occur from ages 9 to 14 years. By age 20 years, Type II errors were observed for three out of four stand variables (table 1).

Power Analysis

Forest researchers are familiar with α levels (for Type I errors), but few understand the importance of β levels for Type II errors (Di Stefano 2001, Foster 2001). Although reviewers of forestry manuscripts sometimes object if an author uses an α level of 0.15 (which increases statistical power), most reviewers do not ask that β levels be listed (Bennett and Adams 2004). As a result, readers of forestry journals are rarely told the power of the tests even when no statistical differences are detected. Some suggest experiments should be designed to

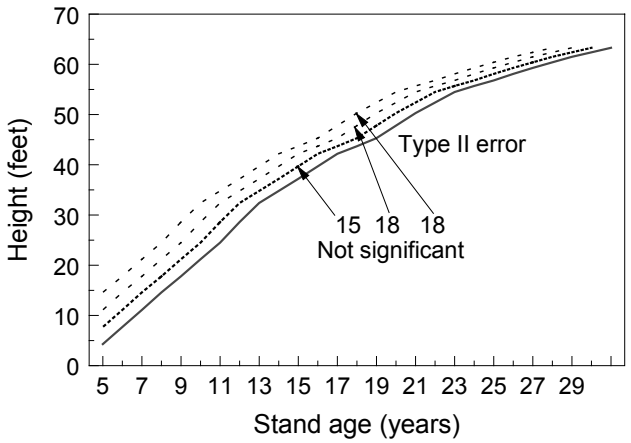

Figure 2—The effect of planting year on height growth of four identical stands. Dashed lines represent stands that were planted 1, 2, and 3 years before the control stand (solid line). The null hypothesis is: there is no difference in stand development due to stand age. Type II errors begin at ages 15, 18, and 18 years (for stands that are 1-, 2-, and 3-years older respectively).

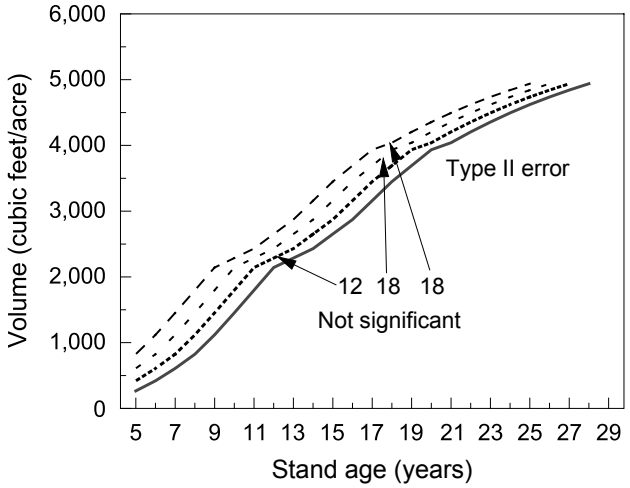

Figure 3—The effect of planting year on volume growth per acre of four identical stands. Dashed lines represent stands that were planted 1, 2, and 3 years before the control stand (solid line). The null hypothesis is: there is no difference in stand development due to stand age. Type II errors begin at ages 12, 18, and 18 years for stands that are 1-, 2-, and 3-years older, respectively.

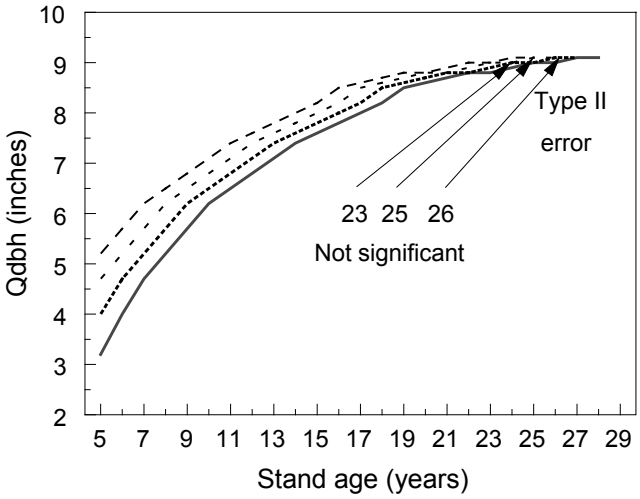

Figure 5—The effect of planting year on quadratic mean d.b.h. (Qdbh) of four identical stands. Dashed lines represent stands that were planted 1, 2, and 3 years before the control stand (solid line). The null hypothesis is: there is no difference in stand development due to stand age. Type II errors begin at ages 23, 25, and 26 years for stands that are 1-, 2-, and 3-years older, respectively.

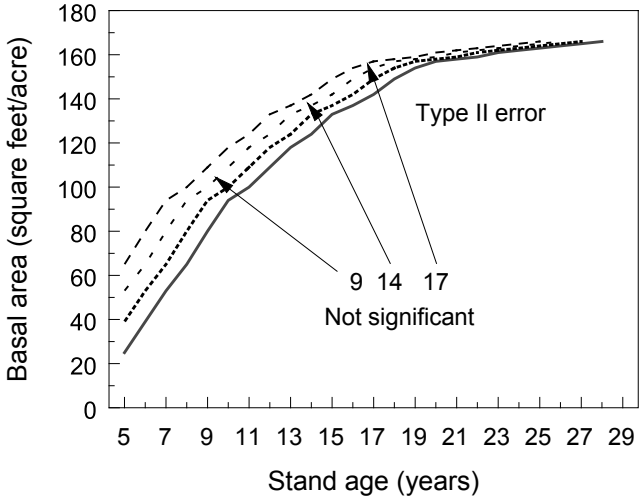

Figure 4—The effect of planting year on basal area per acre of four identical stands. Dashed lines represent stands that were planted 1, 2, and 3 years before the control stand (solid line). The null hypothesis is: there is no difference in stand development due to stand age. Type II errors begin at ages 9, 14, and 17 years for stands that are 1-, 2-, and 3-years older, respectively.

produce a power of 0.8 (i.e., β value of 0.2). This might be rather expensive for most long-term silvicultural trials. A target β value of 0.5 would likely be more practical (Zedaker and others 1993).

There are two basic types of power tests: a-priori and post-priori (Nemec 1991). An a-priori power test helps during the experimental design process and is calculated using the desired sample effect size (i.e., effect size is fixed) and an estimate of the error variance (either an educated guess or error variance from a similar study in the past). A post-priori power test is conducted after the experiment has been measured and is calculated using the observed difference between treatment means (i.e., observed rather than desired sample effect size). Some see little value in post-priori power tests because power level is inversely related to the observed p-value (Haywood and others 1998, VanderSchaaf and others 2003). We therefore propose a "hybrid" power test where the error variance is derived after the study has been analyzed, but the sample effect size is predetermined (i.e., fixed). We suggest calculating power using a "fixed" sample effect size equal to 10 percent of the control mean (but this percentage difference might not be economically or biologically appropriate for all stand variables).

Table 1—Effect of a 2-year difference in planting date on four stand variables. Probabilities associated with a t-test of treatment means (α = 0.05; two tailed test; three samples per mean; completely randomized design)

Variable	Age 18	Age 20	P > t value	Power (1-b)	LSD	α level required to detect a 10% increase (3 samples) b = 0.2	Samples needed to detect a 10% increase b = 0.2	Standard deviation
volume	3,449.00	3,935.00	0.1561	0.282	773.000	0.6930	16.1	338.000
height	50.30	54.50	0.1308	0.468	6.100	0.1220	4.2	2.150
basal area	149	158	0.3179	0.173	21.900	0.2640	6.3	8.530
Qdbh	8.08	8.46	0.0015	1.000	0.251	0.0001	NA	0.062

If a particular test is not statistically significant, is it because there is no effect or because the study does not have enough replications to produce a small enough error term? A hybrid power analysis can be useful in answering this question. A hybrid power test can be calculated using web-based calculators (Thomas and Krebs 1997) or from our Excel program (www.sfws.auburn.edu/south/power.xls). Use of these programs can help researchers gain experience understanding how the α level, number of replications, and use of a one-sided t-test might affect the power level.

In our example, statistical power remained high for at least 8 years. The power (i.e., ability to detect a 1-year difference in stand development) dropped below 0.5 around ages 13 to 16 for basal area and volume, respectively (fig. 6). For height, power did not drop below 0.5 until age 16 years. Quadratic d.b.h. was the least variable, and power remained high for about 20 years.

Similar power levels ($\alpha = 0.05$) were reported for longleaf pine (*Pinus palustris* Mill.) for these growth variables. At age 34 years, Haywood and others (1998) reported the greatest power for tree height (0.443) and the lowest power for volume per acre (0.119). Statistical power for basal area and d.b.h. were intermediate, 0.244 to 0.284. Therefore, the ranking of growth variables in terms of relative power will vary with species, genotype (Burr and Tinus 1996), study (VanderSchaaf and others 2003), and response variable (South and others 2003).

We recommend that, prior to installing a test, researchers decide what difference between means they wish to detect as statistically significant and use computer programs to aid in designing experiments (e.g., Zedaker and others 1993). Too often a study is installed without any idea of how powerful the test will be. For example, in many experiments ($\alpha = 0.05$), researchers can not detect a 12 percent decline in forest floor carbon (Yanai and others 2003), a 10 percent difference in control of woody competition (Zedaker and others 1993), a 9 percent increase in basal area (Miller and others 2001), or a 10 percent increase in seedling production (VanderSchaaf and others 2003).

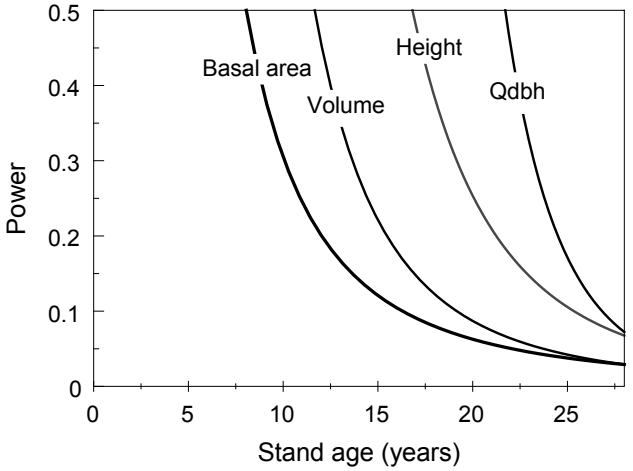

Figure 6—The effect of stand age on the statistical power for detecting a 1-year advance in stand development. A statistical power of 0.3 indicates there is a 70 percent chance of making a Type II error ($\alpha = 0.05$).

For experiments containing silvicultural treatments, we recommend researchers publish the power level ($1 - \beta$) that would detect a 10 percent difference from the control treatment (e.g., 5 percent α level and two-sided test). We suggest using 10 percent as a standard since a post-priori power test is of little value (unless the researchers have failed to report p-values). In post-priori power analyses, where the outcome determines the observed difference between treatment means, power level is inversely related to observed p-value (Haywood and others 1998, VanderSchaaf and others 2003).

Another suggestion would be for researchers to present LSD values. An LSD value would allow the reader to determine how much of an increase in growth would be required before the experimental design could detect a significant difference. Although some reviewers have frowned on reporting unprotected LSD values, this does not mean that an LSD has no value in providing the reader some idea of the statistical power of the test. We recommend LSD values be routinely listed, especially when there are no statistically significant treatment differences.

Another suggestion would be for researchers to switch to modeling the data. Statistical analyses can be conducted when the power of the test is high (during the early stages of stand development), while models can be made when the power is low (during the latter stages of stand development).

CONCLUSIONS

Researchers should expect statistical power to decline with stand age as the coefficient of variation increases over time. Eventually, a treatment that produces a "true" 1-year age-shift will eventually lose statistical significance (typically after age 10 years). However, just because an experiment has low power does not automatically mean the treatment effect has disappeared or has washed out. Forest researchers should remember that they can never "prove" a null hypothesis and therefore it would be unscientific to accept a null hypothesis (scientists can only fail to reject the null hypothesis).

LITERATURE CITED

Bennett, L.T.; Adams, M.A. 2004. Assessment of ecological effects due to forest harvesting: approaches and statistical issues. Journal of Applied Ecology. 41: 585-598.

Burr, K.E.; Tinus, R.W. 1996. Use of clones increases the power of physiological experiments on coastal Douglas-fir. Physiologia Plantarum. 96: 458-466.

Di Stefano, J. 2001. Power analysis and sustainable forest management. Forest Ecology and Management. 154: 141-153.

Foster, J.R. 2001. Statistical power in forest monitoring. Forest Ecology and Management. 151: 211-222.

Haywood, J.D.; Tiarks, A.E.; Elliott-Smith, M.L.; Pearson, H.A. 1998. Response of direct seeded *Pinus palustris* and herbaceous vegetation to fertilization burning, and pine straw harvesting. Biomass and Bioenergy. 14: 157-167.

Kirk, R.E. 1995. Experimental design: procedures for the behavioral sciences. New York, NY: Brooks/Cole Publishing Company. 921 p.

Miller, R.E.; Smith, J.; Anderson, H. 2001. Detecting response of Douglas-fir plantations to urea fertilizer at three locations in the Oregon coast range. Res. Pap. PNW-RP-533. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. 20 p.

Nemec, A.F.L. 1991. Power analysis handbook for the design and analysis of forestry trials. Biometrics Information Handbook 2. Victoria, BC: British Columbia Ministry of Forests. 26 p.

Peterman, R.M. 1990a. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences. 47: 1-15.

Peterman, R.M. 1990b. The importance of reporting statistical power: the forest decline and acidic deposition example. Ecology. 71: 2024-2027.

Richards, F.J. 1959. A flexible growth equation for empirical use. Journal of Experimental Botany. 10: 290-300.

South, D.B.; VanderSchaaf, C.L.; Smith, C.T. 2003. Number of trees per experimental unit is important when comparing transplant stress index values. New Zealand Journal of Forestry Science. 33: 126-132.

Tasissa, G.; Burkhart, H.E.; Amateis, R.L. 1997. Volume and taper equations for thinned and unthinned loblolly pine trees in cutover, site-prepared plantations. Southern Journal of Applied Forestry. 21: 146-152.

Thomas, L.; Krebs, C.J. 1997. A review of statistical power analysis software. Bulletin of the Ecological Society of America. 78: 126-139.

Toft, C.A.; Shea, P.J. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. American Naturalist. 122: 618-625.

VanderSchaaf, C.L.; South, D.B.; Doruska, P.F. 2003. The power of statistical tests of herbicide trials in forest nurseries. Proceedings of the Southern Weed Science Society. 56: 202-211.

VanderSchaaf, C.L.; South, D.B. 2004. Early growth response of slash pine to double-bedding on a flatwoods site in Georgia. In: Connor, K.F., ed. Proceedings of the 12th biennial southern silvicultural research conference. Gen. Tech. Rep. SRS-71. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station: 363-367.

Yanai, R.D.; Stehman, S.V.; Arthur, M.A. [and others]. 2003. Detecting change in forest floor carbon. Soil Science Society of America Journal. 67: 1583-1593.

Zedaker, S.M.; Gregoire, T.G.; Miller, J.H. 1993. Sample-size needs for forestry herbicide trials. Canadian Journal of Forest Research. 23: 2153-2158.

Zhang, S., Burkhart, H.E.; Amateis, R.L. 1996. Modeling individual tree growth for juvenile loblolly pine plantations. Forest Ecology and Management. 89: 157-172.