The Use of Multiple Comparison Tests in Forest Genetics Research

By A. W. Douglas[1]), M. Alvo[1]) and M. A. K. Khalil[2])

(Received 28th October 1981)

Summary

This paper presents the results of a study comprising 30 provenances of red spruce *(Picea rubens* Sarg.) from the Maritime Provinces of Canada and the northeastern United States. The seed was sown in the Canadian Forestry Service Research Nursery near Fredericton, New Brunswick in the spring of 1960. The seedlings were planted in a ten-replicated randomized complete block design experiment with four-tree square plots in the North Pond Experimental Area in east-central Newfoundland in the spring of 1964 at the age of four years. Measurements were made on five response variables in September 1979 at the age of 20 years from seed. An analysis of variance indicated very highly significant differences between provenance means for all variables. On the basis of height a discussion is presented of five commonly used multiple comparison procedures, *i. e.* those of Scheffé, Bonferroni-t, Tukey, Student-Newman-Keuls and Duncan, used to determine the location of the significant differences. The interpretation of these tests with respect to their probability error rates is presented together with a discussion of their advantages and drawbacks in terms of this experiment, and recommendations for their use are made.

Key words: *Picea rubens* (Sarg), multiple comparison tests, Scheffé, Tukey, Bonferroni, Student-Newman-Keuls, Duncan.

Sommaire

Cet article présente les résultats d'une étude fondée sur 30 provenances d'épinette rouge dans les provinces maritime du Canada et dans le nord des Etats Unis. Les grains ont été planté au printemps 1960 dans une pépinière du Service Forestier du Canada près de Fredericton au Nouveau Brunswick. Le plan experimental du bloc complet avec répartition au hasard à 10 répétitions fut employé avec placeaux de 4 arbres âgé de 4 ans dans la région experimentale du North Pond dans la partie est du centre de Terre Neuve au printemps 1964. En septembre 1979, on a collectionné des données sur 5 variables pour les arbres âgés de 20 ans.

L'analyse de variance a devoilé qu'il existe des différences significatives parmi les moyennes des provenances. Dans cet article, on présente une discussion sur 5 mèthodes utilisées pour situer plus précisement ces différences. En particulier ce sont les méthodes de Tukey, Scheffé, Student-Newman-Keuls, Duncan et Bonferroni-t. De même, on interprète ces méthodes vis-à-vis les taux d'erreur, on présente les avantages et désavantages en vue de cette expérience et l'on propose quelques recommendations pour leur usage en général.

Zusammenfassung

Die vorliegende Arbeit ist das Ergebnis einer Studie an 30 Herkünften der amerikanischen Rotfichte *(Picea rubens* Sarg.) aus den Küstenprovinzen von Kanada und dem Nordosten der Vereinigten Staaten von Amerika. Die Aussaat erfolgte im Frühjahr 1960 in einer Baumschule des kanadischen Forstdienstes in der Nähe von Fredericton, New Brunswick. Die vierjährigen Sämlinge wurden im Frühjahr 1964 in einer zehnfach-wiederholten vollständig randomisierten Blockanordnung in Vier-Baum-Quadratparzellen im ost-zentralen Versuchsgebiet „North Pond" auf Neufundland gepflanzt. Im September 1979, im Alter von 20 Jahren wurden fünf Merkmale aufgenommen. Eine Varianzanalyse zeigte hoch signifikante Differenzen zwischen den Herkunftsmitteln für alle Variabeln. Anhand des Merkmals Höhe wurden fünf gebräuchliche multiple Vergleichsverfahren diskutiert und zwar: Scheffé, Bonferoni, Tukey, Student-Neuman-Keuls und Duncan, die i.a. zum Bestimmen signifikanter Differenzen dienen. Sie werden hinsichtlich ihrer Fehlerraten sowie ihrer Vor- und Nachteile diskutiert. Es wurden einige Vorschläge für die Anwendung der Tests gemacht.

Introduction

Forestry experiments usually have a large number of treatments under study, whose number may exceed 100 in forest genetics experiments. In the case where the F-statistic in an analysis of variance is statistically significant, interest centers on identifying which treatments are different. This generally involves multiple comparisons between pairs of treatments to identify the significantly different pairs. The number of such pairwise comparisons is I(I-1)/2, where I is the number of treatments. The number of comparisons further increases if some combinations of three or more treatments also have to be compared.

Five methods are in common use for multiple comparisons, viz. Scheffé (Scheffé 1959; Steel and Torrie 1980); Bonferroni-t (Douglas 1979; Miller 1966; Timm 1975), Tukey, Student-Newman-Keuls and Duncan (Steel and Torrie 1980). This paper discusses the relative merits and limitations of the above methods, illustrating them with data from a red spruce *(Picea rubens* Sarg.) provenance experiment being conducted by one of the authors in Newfoundland.

Material and Methods

Bulked seed was obtained from 30 red spruce provenances, 16 from Nova Scotia, 10 from New Brunswick in Canada and two each from the states of Maine and West Virginia in the U.S.A. The 26 Canadian provenances lie in six forest sections of the Acadian Forest region (Rowe 1972) between latitudes 44°10'N and 47°39'N and longitude 62°27'W and 66°59'W. The four provenances from the U.S.A. lie within the region bounded by latitudes 38°47'N and 47°13'N and longitudes 68°38'W and 79°37'W.

The provenance experiment was established in the spring of 1964 with 2-2 seedlings in the Mint Brook Valley in the North Pond Experimental Area in east-central Newfoundland (lat. 48°0'N, long. 54°34'W) at an elevation of 107 m. A ten-replicated radomized complete block design with four-tree square plots was used in which a spacing of 1.8 × 1.8 m was adopted.

Statistical analysis

For the purposes of this paper, consideration will be limited to an analysis of the response variable Y = height

[1]) Director and Statistical Consultant respectively, Data Analysis and Systems Branch, Computing and Applied Statistics Directorate, Environment Canada, Ottawa, Canada.

[2]) Research Scientist, Newfoundland Forest Research Centre, Canadian Forestry Service, Environment Canada, St. John's, Newfoundland, Canada.

Table 1. — ANOVA on height at 20 years from seed.

Source	d.f.	Sum of squares	Mean square	F
Blocks	9	307 881.47	34 209.05	5.63 *
Provenance	29	591 952.64	20 412.16	3.36 *
Experimental Error	261	1 584 738.18	6 071.79	
Sampling Error	900	3 539 208.50	3 932.45	
Total (Corrected)	1 199	6 023 780.79		

* = significant at the 0.005 level.

Table 2. — Provenance means based on 40 observations.

1.	179.2	11.	142.5	21.	166.0
2.	169.7	12.	176.2	22.	151.3
3.	193.6	13.	182.4	23.	162.1
4.	174.5	14.	233.6	24.	157.7
5.	177.3	15.	192.4	25.	147.8
6.	172.1	16.	203.3	26.	166.5
7.	158.1	17.	215.0	27.	174.9
8.	143.9	18.	195.1	28.	192.2
9.	172.7	19.	173.5	29.	195.9
10.	146.7	20.	202.5	30.	213.4

Table 3. — Number of significant pairwise comparisons among provenance means.

Test	Significance level 0.05	0.01
Scheffé	0	0
Bonferroni-t	13	7
Tukey	18	7
Student-Newman-Keuls	22	8
Duncan	90	48

(cm) at 20 years from seed. An analysis of variance (ANOVA) was performed using a mixed randomized complete block model with subsampling where the block effect is random and the provenance effect is fixed. The results are given in *Table 1.* The provenance means are presented in *Table 2.*

Given the very highly significant F-value for provenances, interest now lies in determining the location of these differences. A common approach is to make all pairwise comparisons between the provenance means using one of a number of multiple comparison methods available. We will examine the results obtained by employing the five methods mentioned earlier. Subroutines to perform these calculations are available from the first author.

The number of significant pairwise comparisons among the 435 made is given for each method in *Table 3.*

As will be discussed in the next section, any difference declared significant using one of the methods listed in *Table 3* will also be declared significant by any method appearing below it in the list. For example, the specific 18 comparisons declared significant at the 0.05 level using Tukey's method will also be declared significant at that level using Student-Newman-Keuls and Duncan's tests.

Multiple Comparison Methods

The use of multiple comparison methods by a researcher requires that he identify at the outset the basic family or collection of statements for which it is necessary to guarantee a given probability of correctness. Since researchers invariably perform F-tests on all the treatments before doing any multiple comparison tests, the implication is that the experiment itself is the unit of interest. Below we identify the family of statements associated with each of the multiple comparison methods listed in *Table 3.* Before proceeding, however, it is important to note that the probability error rate for a given family of statements is defined to be

$$\frac{\text{Number of incorrect statements}}{\text{Number of statements in the family}}$$

With the exception of Duncan's multiple range test, all of the multiple comparison techniques keep this error rate fixed throughout the analysis.

Scheffé's method is concerned with the family of all possible linear combinations among the theoretical means. Since this family is the largest possible, it is natural to expect that this method will be the least sensitive of those available. In our experiment, we note that even though the F-test for provenances is significant at the 0.005 level, there are no pairwise significant differences at both the 0.05 and 0.01 levels. The implication is that one or more other linear combinations of the treatment means are responsible for the rejection of the null hypothesis though the researcher may not be particularly interested in them. Finally, we mention that since the Scheffé method is based on the F-test itself, the error rate α, is the same as the level of significance of the test, namely 0.05.

The Bonferroni-t method is perhaps the most flexible of the available techniques in that it allows the researcher to define in advance his own particular family of statements. Based on the number of members in the family, one computes an approximate critical value at which to test each member of the family. In our example, there are $I = 30$ provenances. If the family of interest consists of all $I(I-1)/2$ pairwise differences, then according to the Bonferroni inequality each pairwise difference is tested for significance at the $2\alpha/(I(I-1))$ level and the overall error rate is set at α. The use of the Bonferroni-t method imposes no restrictions on sample sizes for the different treatments (as we will see is the case for Tukey's method) nor is it required that the sample means be statistically independent. For these reasons, the Bonferroni-t method can sometimes be more effective than any of the other methods. Its principal drawback stems from the fact that a correct utilization of the method requires extensive tables of Student's t-distribution since $2\alpha/(I(I-1))$ is generally an unconventional value.

Tukey's method, which is based on the studentized range statistic (i.e. the maximum difference between any two sample means divided by the standard error), is concerned with the family of all pairwise differences among the means. Since this family is smaller in size than Scheffé's, one can expect to detect more significant pairwise differences. Both the level of significance and the error rate of the test are set at α. Tables of critical values for Tukey's method are readily available whenever there are equal sample sizes for the treatments. Special extensive tables are required in the case where this condition is not satisfied.

Both the Student-Newman-Keuls procedure and the Duncan multiple range test are based on the studentized range statistic and both proceed by analyzing for significance progressively smaller subsets of ranked means. Given I means, a subset of p means contains at least one signifi-

cant pairwise difference only if for all larger subsets there is at least one significant pairwise difference. In other words, the question of significance for the subset will not be obscured by the inclusion of additional means. Where the two methods differ is in the choice of significance level by which to test the subsets. The Student-Newman-Keuls method retains the same error rate for each subset of pairwise differences whereas Duncan's treats a test that p means are equal as equivalent to (p-1) independent tests that pairwise differences among the meansare zero. Hence, for Duncan's method the error rate for a subset of p means is $1(1-\alpha)^{p-1}$, which increases as the number of treatments in the subset increases. The level of significance associated with a single pair of means is α. The controversy which now exists concerning the Duncan test is due to this changing error rate which, as a consequence, makes this test the most sensitive in detecting differences among the means. In our example with 30 provenance means, the error rate for the largest subset will be, $1-(1-0.05)^{30-1} = 0.77$ at $\alpha = 0.05$. With such a large allowable error rate one must not be too surprised if many significant pairwise differences are detected.

In order to illustrate more fully the differences among the procedures of Tukey, Student-Newman-Keuls and Duncan, we consider the example at hand in more detail. With 30 treatment means and an experimental mean square error of $s^2 = 6\,071.79$ with 261 degrees of freedom, the Tukey procedure would accept the hypothesis if the maximum of the absolute differences is less than $qs/\sqrt{n}$ where $q = q(30,261) = 5.301$ for $\alpha = 0.05$ and 5.911 for $\alpha = 0.01$ (using tables of the studentized range statistic with 30 and 261 degrees of freedom). Here, $n = 40$, the number of observations on which a mean is based, and consequently for $\alpha = 0.05$, 0.01, $qs/\sqrt{n} = 65.31$, 72.83 respectively. If the difference between the two largest means was very large, the null hypothesis that the theoretical means were equal would be rejected and the difference declared significant. In order for the second largest difference to be declared significant, it would also have to exceed 65.31 for $\alpha = 0.05$ or 72.83 if $\alpha = 0.01$. Student-Newman-Keuls procedure allows this second difference to be compared to $q(29,261)s/\sqrt{n} = (5.277)(12.32) = 65.02$ for $\alpha = 0.05$. As the number of means in the subsets decreases, the critical value decreases, yet the error rate for the subset is fixed. Hence, the Student-Newman-Keuls procedure permits an error rate of α for the overall set of 30 provenance means as well as for each subsequent subset. In comparison, the Tukey procedure is more stringent in that even though the error rate is α for the set of 30 provenance means, it may be much less for subsequent subsets.

Duncan's procedure, on the other hand, permits the error rate to change with increasing subset size. For example, the set of 30 provenance means would be tested with an error rate given as $1-(1-\alpha)^{29}$ which for $\alpha = 0.05$, 0.01 equals 0.77 and 0.25 respectively. The corresponding critical q value would be given by 3.550 and 4.514. Consequently, the value which the largest pairwise difference would have to exceed is, for $\alpha = 0.05$, given by $(3.550)(12.32) = 43.74$ and for $\alpha = 0.01$, it is $(4.514)(12.32) = 55.62$. These values are much smaller than the corresponding values for the Tukey and Student-Newman-Keuls procedures and consequently one would expect to detect more significant differences. We note that the value of α for Duncan's procedure is the error rate for a single pairwise comparison.

In conclusion, the reader should note that we have here reported on multiple testing, and the Student-Newman-Keuls and Duncan procedures, unlike the others, cannot be adopted for the construction of confidence intervals.

Conclusions and Recommendations

The preceding explanation of the probability bases for the multiple comparison procedures used may help to explain why we obtained the results we did in *Table 3*. It will not, however, necessarily make it easier for the experimenter to choose the appropriate test or tests. This choice is less a statistical than a philosophical one which depends primarily on the experimenter's view of what constitutes a natural family of statements from which he wishes to draw inferences.

It is the authors' impression, based on experience, that researchers conducting designed experiments generally consider, whether stated explicitly or not, the experiment as the unit of interest, i.e. it is in terms of the conceptual repetition of the experiment that the experimenter thinks when making inferences to future occurrences. This would tend to suggest that Scheffé's procedure should be the one universally adopted, but it is usual that the experimenter has a specific family of statements or contrasts in mind such as is the case in the experiment described in this paper. Therefore, it would seem that the Bonferroni-t method would be the best choice in the case where a general set of contrasts is specified, and Tukey's procedure would be most appropriate in the case where one wishes to make all pairwise contrasts between the provenance means. As is the case for the above tests, the Student-Newman-Keuls procedure preserves the error rate under the null hypothesis. However, it has different properties under the alternative. It is less conservative than Tukey's procedure, and it would be appropriate in the case where one perceives that each subset should have the same error rate as the overall set of provenance means. The conceptual experiment now consists of several sub-experiments corresponding to the subsets of the provenances means. The Duncan procedure does not preserve the error rate for any of the subsets of means except for the one consisting of two means. This fact seems to run counter to the basic philosophy of multiple comparison procedures which is to control the error rate for the entire null hypothesis at a specified level.

The above discussion assumes that the sample size (the number of blocks in this case) is fixed. If this is the case, and biologically meaningful contrasts are not declared to be significant the experimenter may wish to increase his level of significance to 0.10 or even 0.15 in order to detect these differences. In future experiments he may wish to decrease the number of provenances in order to increase the power of the tests in the presence of limited availability of experimental material.

If the experimenter has flexibility in the amount of experimental material available he should consider making sample size calculations (Scheffé 1959) in order that the initial F-test, and the subsequent multiple comparison tests, in the analysis of variance will detect differences of importance with the desired power.

Literature Cited

Douglas, A. W.: On levels of significance. Computing and Applied Statistics Directorate, Environment Canada, Ottawa, 14 pp.

(1979). — MILLER, R. G., Jr.: Simultaneous statistical inference. McGraw-Hill Book Co., New York. (1966). — ROWE, J. H.: Forest regions of Canada. Dep. of the Environ., Can. For. Serv. Pub. No. 1300, 172 pp. (1972). — SCHEFFÉ, H.: The analysis of variance. John Wiley and Sons, Inc., New York. (1959). — STEEL, R. G. D. and TORRIE,, J. H.: Principles and procedures of statistics, a biometrical approach, second edition. McGraw-Hill Book Co., New York. (1980). — TIMM, N. H.: Multivariate analysis, with applications in education and psychology. Brooks/Cole Publishing co., Monterey, California, pp. 368, 585, 660. (1975).

Genotype x Environment Interactions and Seed Movements for Loblolly Pine in the Western Gulf Region*)

By J. L. YEISER[1]), J. P. VAN BUIJTENEN[2]) and W. LOWE[3])

(Received 29th October 1981)

Summary

Fifteen plantations were established throughout the western gulf region to analyze genotype by environment (GXE) interactions in loblolly pine. Open pollinated families from five selected trees plus a checklot from each of four seed zones were planted at each location. The seed zones tested were southeastern Texas, southern Louisiana, northern Louisiana and southern Arkansas.

Significant heterogeneity effects indicated that the GXE interaction for height and volume could be reduced by stratifying environments. Regression estimates of slope and standard deviation indicated southeastern Texas and northern Louisiana sources were intermediate in stability. Southern Arkansas and southern Louisiana families were equally unstable. Southern Louisiana families were most responsive and southern Arkansas families least responsive to improved site quality. Ecovalences and coefficients of genetic prediction suggested that southern Louisiana families may be well adapted to high site index areas in northern Mississippi, but the same sources should not be moved more than 125 miles northwest to Southeastern Texas. Results also showed that southern Arkansas sources could be moved to northeastern Texas.

Key words: Pinus taeda, genotype × environment interaction, plant stability, seed movement.

Zusammenfassung

Es wurden fünfzehn Versuche mit *Pinus taeda* L. in der westlichen Golfregion angelegt, um die Genotyp-Umwelt-Interaktion zu analysieren. Frei abgeblühte Familien von fünf selektierten Bäumen sowie einer Kontrolle aus je vier Samenzonen wurden in allen Versuchen angepflanzt. Geprüft wurden die folgenden Samenzonen: Südost-Texas, Süd-Louisiana, Nord-Louisiana und Süd-Arkansas.

Signifikante Heterogenitätseffekte zeigten, daß die Genotyp × Umwelt-Interaktion für Höhe und Volumen vermindert werden kann, indem man eine Abstufung von Standort zu Standort vornimmt. Regressionsschätzungen von Gefälle und Standardabweichung zeigen, daß die Herkünfte Südost-Texas und Nord-Louisiana von mittlerer Stabilität sind, Familien aus Süd-Arkansas und Süd-Louisiana waren ähnlich unstabil. Familien aus Süd-Louisiana waren mehr an bessere Standorte angepaßt als Familien aus Süd-Arkansas. Ökovalenzen und Koeffizienten zur genetischen Vorhersage machten deutlich, daß Familien aus Süd-Louisiana wahrscheinlich gut an fruchtbare Böden in Nord-Mississippi angepaßt sind, aber dieselbe Herkunft sollte nicht weiter als 125 Meilen nordwestlich von Südost-Texas angepflanzt werden. Die Ergebnisse zeigen auch, daß Familien aus Süd-Arkansas mit gutem Erfolg in Nord-Texas angebaut werden können.

Introduction

Realization of the genetic potential of superior seed depends on its use on appropriate sites. Since state, federal, and industrial land holdings within the southern United States frequently span long distances and a variety of sites, guidelines governing seed movements are needed. Assessing plant stability and genotype by environment (G × E) interaction is the first step in developing a sound seed movement policy. This paper is a report of a study to determine:

1) The presence and magnitude of G × E interaction in selected families of loblolly pine *(Pinus taeda* L.) native to the Western Gulf Forest Tree Improvement Program (WGFTIP) region.

2) The stability of selected families of loblolly pine indigenous to the WGFTIP region.

Literature Review

Genotype by environment interaction may be defined as the inconsistent relative performance of two or more genotypes over two or more environments. A regression technique assessing G × E interaction and thus plant stability was introduced by YATES and COCHRAN (1938), popularized by FINLAY and WILKINSON (1963), and modified by EBERHART and RUSSELL (1966) and FREEMAN and PERKINS (1971). This technique produces estimates of slope (b_i) and deviations from regression (s^2), which may be jointly interpreted to explain G × E interaction. A family with $b_i > 1$ responds to improvements in site quality. As s^2 becomes smaller, family performance becomes more predictable. Stable families have estimates of slope near one and deviations from regression near zero.

Another stability parameter assessing G × interaction is Wricke's ecovalence (SHELBOURNE 1972). This statistic estimates the sums of squares contribution of each genotype to the overall G × E term. The smaller the contribution the more stable the genotype.

Both Wricke's ecovalence and the regression estimates for slope and variance have been used as practical indicators of plant stability. In a study of geographic seed sources, VAN BUIJTENEN (1978) used estimates of slope and

*) This study was completed in partial fulfillment of the requirements of a Ph. D. degree at Texas A&M University, 1980.

[1]) Assistant Professor of Forestry, University of Arkansas at Monticello and Arkansas Agricultural Experiment Station.

[2]) Principal Geneticist, Texas Forest Service; Professor, Texas Agricultural Experiment Station, College Station, Texas.

[3]) Associate Geneticist, Texas Forest Service; Assistant Professor, Texas Agricultural Experiment Station, College Station, Texas.