Character Selection and Data Structure in geographic Variation in Pinus contorta

By K. W. Newman and R. C. Jancey

Dept. of Plant Science, University of Western Ontario,
London, Ontario N6A 5B7, Canada*)

(Received 21st June 1982)

Summary

An algorithm which ranks characters on the basis of a comparison of their variation between locations to their variation within locations was used to select cone characters for a multivariate study of geographic variation in *Pinus contorta*. A resemblance function was developed which not only was appropriate for a geometrical investigation of data structure, but also was standardized in a manner which took advantage of the replicated sampling at each location so that trends between locations were emphasized rather than obscured. Principal components analyses applied to cones and to locations was followed by the plotting of stereograms of the first three principal components to provide three-dimensional displays of the data structure. By correlating resemblance matrices, it was shown that there was very little distortion in these three-dimensional displays. The stereograms revealed a clear group structure corresponding to the four subspecies defined by Critchfield (1957), with the addition that the population having serotinous cones from the serpentine soil of northwest California was grouped with the populations having serotinous cones from the highly acid soil near Mendocino.

Key words: geographic variation, *Pinus contorta*, pine cone, descriptor ranking, multivariate analysis.

Zusammenfassung

Um bestimmte Zapfenmerkmale für eine multivariable Studie der geographischen Variation von *Pinus contorta* zu selektieren, wurde ein Algorithmus angewandt, der — auf der Grundlage eines Vergleichs der Variation zwischen Orten mit der Variation innerhalb von Orten — die Merkmale in eine Rangfolge bringt. Ferner wurde eine Ähnlichkeitsfunktion entwickelt, die nicht nur für eine geometrische Untersuchung der Datenstruktur geeignet, sondern auch derart standardisiert war, daß sie die wiederholte Stichprobennahme an einem Ort ausnutzte. Daher wurden Trends zwischen Orten eher betont als verschleiert. Auf die Anwendung der Hauptkomponenten-Analyse auf Zapfen und Orte folgte für die ersten drei Hauptkomponenten die Ausgabe von Stereogrammen, um eine dreidimensionale Darstellung der Datenstruktur zu bekommen. Durch Inbeziehungsetzen der Ähnlichkeitsmatrizen wurde gezeigt, daß diese dreidimensionalen Darstellungen nur sehr wenig verzerrt sind. Die Stereogramme offenbarten eine klare Gruppenstruktur, die mit den von Critchfield (1957) definierten vier Unterarten übereinstimmt. Es ergab sich zusätzlich nur, daß die Population mit sich langsam öffnenden Zapfen von den Serpentinböden Nordwest-Kaliforniens in die Gruppe mit sich langsam öffnenden Zapfen von den stark sauren Böden nahe bei Mendocino kam.

Introduction

In selecting characters for a multivariate analysis of geographic variation in a species, there are two constraints which must be considered. An increase in the number of characters measured must result in a decrease in the number of individuals that can be included in the study for a given amount of time available. In addition, many characters may obscure the underlying patterns by increasing random variation. Therefore, prior to an investigation of geographic variation in the cones of *Pinus contorta*, a study was carried out to determine objectively those characteristics of the cones that would be most useful in a study of geographic variation. Several of the characters considered have definitions which are original to this study, and may be of use in other studies involving variation in gymnosperm cones.

This paper also describes the selection of a resemblance function particularly suited to studies of geographic variation, and then demonstrates the application of principal components analysis and stereograms to elucidate the basic structure of the data.

The selection of Cone characters

The material used for the pilot study of cone characters consisted of forty cones of *Pinus contorta;* that is, ten cones randomly selected from each of four locations, with each location chosen from a different one of the four geographic regions for which Critchfield (1957) defined subspecies. Three of these locations were selected from the stands studied by Rall (1979) (i.e. Mendocino and Lake Tahoe, California, and the Cypress Hills, Alberta), the fourth location was represented by a collection of cones provided by Illingworth (from Hauser Dunes on the Oregon coast).

The procedure used evaluates characters based on the ratio of their variance among locations to their variance within locations (Jancey, 1979). One advantage of this ranking procedure as compared to other ranking procedures is that it utilizes the information obtained by the replication of sampling within each location. The philosophy behind the use of the procedure is that the criterion for selecting a useful character for studying variation among locations is not the variation of that character in a pooled data set which ignores the source of the variation, but rather a consideration of whether the variation of the character among locations is large compared to its variation within locations.

The details of the basic method are given in Jancey (1979) where it is suggested that all pairwise comparisons between groups be considered. This suggestion was used since it was desired that each pairwise comparison of sites should contribute its most effective characters to the set selected, rather than ranking on "overall effectiveness."

Description of Characters

The characters which were considered are given in the following list. For some characters, an abbreviation is provided which will be used to identify that character in the tabular summary of results.

*) This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

1. Cone weight: measured after the cones had been oven-dried at 63° C for 24 hours.
2. Cone volume: measured using the volume-equivalent method. This is the method which was judged most accurate by CRITCHFIELD (1957), after his experimental evaluation of various methods of determining the volume of pine cones.
3. Cone length: the greatest linear dimension of the cone, see *Figure 1* for a comparison of this to the next character.

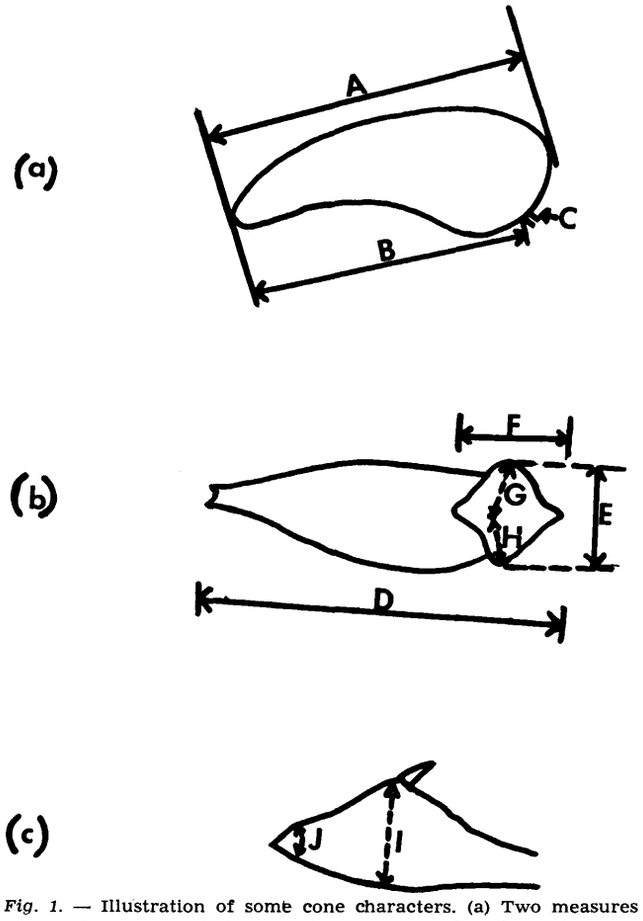

Fig. 1. — Illustration of some cone characters. (a) Two measures of cone length: greatest linear dimension (A); length (B) from tip to point of attachment of the peduncle (C). (b) Abaxial view of cone scale illustrating: cone scale length (D); apophysis width (E); apophysis length (F); apophysis longest lateral slope (G); apophysis shortest lateral slope (H). (c) Longitudinal section through apophysis of cone scale illustrating: apophysis height (I); apophysis tip angle (J).

4. Cone length from tip to point of attachment of the peduncle.
5. Cone width: cones were soaked in warm water before measurement, so that they were closed.
6. Cone serotiny: measured on a scale of 1 to 3, using the method of RALL (1979) in which 1 represents fully open cones, 3 represents fully closed cones, and 2 represents cones which were partially open.

The following four characters were measured for a cone scale approximately one-quarter of the distance from the tip, on the abaxial side of the cone. This is the location used by RALL (1979) for measurements of characters of cone scales. *Figure 1* illustrates the measurement of thes characters, as well as the other apophysis measurements.

7. Cone scale length:
8. Apophysis width
9. Apophysis height
10. Apophysis tip angle

All of the following apophysis measurements were made on a cone scale approximately one-quarter of the distance from the base, on the abaxial side of the cone. The reason for choosing this location was that, based on visual examination, it appeared that the differences in apophysis between cones from different locations were more pronounced in this region of the cone than closer to the tip.

11. Apophysis width
12. Apophysis length
13. Apophysis height
14. Apophysis longest lateral slope: referred to by the abbreviation "slope 1."
15. Apophysis shortest lateral slope: referred to by the abbreviation "slope 2."

In addition to these basic variables, various combinations of them were also evaluated. The reason for evaluating these combinations is that while the variation within a location for "size" characters, such as, cone length or cone width, may be too great compared to their variation between locations for them to be of much diagnostic value, nevertheless, some combination of these characters such as cone width divided by cone length, which is a "shape" character, may be relatively constant for any particular location while varying between locations.

There has been considerable controversy concerning the use of ratio characters (ATCHLEY, GASKINS, and ANDERSON, 1976; CORRUCCINI, 1977; HILLS, 1978; DODSON, 1978; ALBRECHT 1978; ATCHLEY and ANDERSON, 1978; ATCHLEY, 1978). The main argument against the use of this type of character is that if the original characters are linearly related, then non-linearity is introduced into the data by the use of ratios of these characters. However, this problem is easily avoided if logarithms are used, since log (x/y) = log (x) — log (y), which is a linear relationship. Furthermore, it may be argued that the use of logarithms provides a more reasonable expression of the resemblances among the cones than do the untransformed measurements. For example, if cone B is twice as long as cone A, and cone C is twice as long as cone B, it is reasonable to consider that, in terms of the character "cone length", cone B resembles cone A to the same extent as cone C resembles cone B, this is the resemblance implied in the use of the logarithm of cone length, whereas, the untransformed length implies that cone C differs form cone B by twice the amount by which cone B differs from cone A. Therefore, in this study, ratio characters were allowed and the logarithmic transformation was used.

Combined characters submitted to ranking were:

1. Specific gravity
2. Cone width/cone length
3. Cone length from the tip to the point of attachment of the peduncle, divided by the maximum length of the cone. This character may be regarded as a measure of cone symmetry, since it has a maximum value of I for symmetric cones, and lesser values for cones that are asymmetrically curved to one side. Thus, it is referred to by the abbreviation: "symmetry".
4. Cone length multiplied by the square of cone width multiplied by pi/4. This is the volume of a cylinder with length equal to the length of the cone, and diameter equal to the width of the cone. This character is refer-

red to by the abbreviation: "cylvol" or "cylindrical volume".

5. Cone volume divided by the preceding character: This is a measure of cone shape which is larger for cones with little taper, which approximate the shape of a cylinder, and smaller for cones which taper more. It is referred to by the abbreviation: "vol./cylvol".
6. Apophysis length divided by apophysis width
7. Apophysis height divided by apophysis width
8. Apophysis height divided by apophysis length
9. Sum of apophysis lateral slopes, that is slope 1 plus slope 2.

Choosing from the Ranked Characters

Table 1 summarizes the results of the ranking procedure by listing the characters described in the previous section in order of their highest ratio of variances based on comparisons between all pairs of locations, and also names the corresponding pair of locations. In addition to the characters listed in the table, the ranking procedure was also applied to various monotonic functions of these characters, for example, logarithms, the cube roots of volume and weight, and inverses of the ratio characters. The variance ratios corresponding to these monotonic transformations were approximately equal to those corresponding to the original characters, suggesting that the numerical values of the ratios are not very sensitive to the form of the underlying statistical distribution of the character values. A reasonable method for determining a "cut-off" point for selecting characters, could be based on a comparison of the magnitudes of the variance ratios to the F-statistic (with 1 and 18 degrees of freedom, since comparison is between 2 groups with 10 cones each). To be conservative in accepting the hypothesis that a character showed significant variation between a pair of locations, a "cut-off" of $F = 8.29$, corresponding to a significance level of 1%, could be used.

Four of the top five ranked characters (namely, cone width, cone volume, cone weight, and cone cylindrical volume) are all basically expressing information about cone size. To avoid redundancy, cone volume was selected as a good single measure of overall cone size. The other size measurements were then considered as ratios to express various aspects of cone "shape". In particular, the following three characters were selected to represent cone "shape": cone width divided by cone length, cone weight divided by cone volume (specific gravity), and cone volume divided by the volume of a cylinder of equal length and diameter. It is of interest to note that these three "shape" characters along with the character cone volume, each have their greatest diagnostic ability for a different location pair (i.e., cone volume receives its highest ranking based on the comparison between Tahoe and Cypress Hills, specific gravity receives its highest ranking based on the comparison between Mendocino and Hauser, etc.), whereas, the four original size variables all had their greatest diagnostic ability for the same location pair.

Table 1. — Highest ranking for each character, with the corresponding pair of locations.

Character	F Value	Locations
cone width	54.9	Tahoe, Cypress Hills
cone serotiny	52.4	Tahoe, Cypress Hills
cone volume	42.7	Tahoe, Cypress Hills
cone weight	41.9	Tahoe, Cypress Hills
cylindrical volume	36.6	Tahoe, Cypress Hills
cone width/length	29.7	Mendocino, Cypress Hills
apophysis height	24.9	Tahoe, Cypress Hills
apoph. ht./length	16.9	Cypress Hills, Hauser
cone length	16.3	Tahoe, Hauser
apoph. ht./width	15.0	Mendocino, Tahoe
symmetry	14.3	Mendocino, Tahoe
length to peduncle	12.9	Tahoe, Hauser
apoph. length	10.6	Tahoe, Hauser
vol./cylvol	9.1	Mendocino, Tahoe
specific gravity	9.0	Mendocino, Hauser
apoph. width/len.	6.8	Mendocino, Tahoe
apoph. width (Rall)	6.2	Tahoe, Cypress Hills
scale length	5.5	Tahoe, Cypress Hills
apoph. slopei	5.0	Tahoe, Cypress Hills
apoph. width	4.1	Tahoe, Hauser
apoph. sum slopes	3.6	Tahoe, Cypress Hills
apoph. tip angle	2.9	Tahoe, Cypress Hills
apoph. slope 2	2.5	Tahoe, Hauser
apoph. height (Rall)	2.5	Mendocino, Tahoe

Cone serotiny was selected as a character, but it was decided that an improved method of measurement would be used. Recall that serotiny had been measured on a scale of 1 to 3, where 1 indicated a "fully-open" cone, 3 indicated a fully-closed cone, and 2 indicated an intermediate cone. Fully-closed cones are quite obvious, but the distinction between a "fully-open" cone and an intermediate cone is rather vague, since some of the scales next to the peduncle are closed even for the most open cones. The new method of measuring cone serotiny is based on the observation that the cones open from the tip, and that a varying proportion of the length of the cone remains closed at the end adjacent to the peduncle. Thus, the proportion of the total length of the cone that is closed is used as a measure of cone serotiny, this measurement is made on the abaxial side of the cone. The new method of measuring cone serotiny not only converts the measure to a continuous scale rather than an interval scale, so that it is more compatible with the measurements of the other characters, but also is more informative of the degree of serotiny expressed by a cone. This new idea for measuring cone serotiny did not occur until after the cones in the pilot study had been treated in ways which made this measurement impossible for those cones, however, the high rank for the serotiny character based on the earlier method of measurement indicates the diagnostic ability of the characteristic.

The highest ranked character, based on measurements of apophyses, was apophysis height. This character was selected to describe the apophyses along with the "shape" characters apophysis height divided by apophysis length, and apophysis height divided by apophysis width. Like the four characters already selected to describe cone size and shape, these apophysis characters each have their greatest diagnostic ability for a different pair of locations.

Finally, the ratio of the length from the point of attachment of the peduncle to the tip of the cone, compared to the total length of the cone was included in the character-set as a measure of cone symmetry.

The Data

The data for the main study of geographic variation consisted of measurements of nine characters for each of ten cones from each of 24 locations in California, Utah, and Colorado (Newman, 1982). One cone was randomly selected from each of ten trees at each location. These locations were selected to provide information on the variation of the cones of *Pinus contorta* among the four geographic subspecies recognized by Critchfield (1957), and to compare the magnitude of the variation between geographic regions to the variation with respect to changes in elevation, latitude, topography, climate and soil.

Selection of a Resemblance Function

For ease of interpretation of the data structure in terms of a spatial analogy, a Euclidean resemblance function is desirable. However, the standardization or weighting of variables must be carefully considered to avoid misrepresentation or distortion of the resemblance structure.

In its simplest form, the Euclidean distance between a pair of individuals, j and k, each described by the same p variables, is defined by

$$E(j,k) = \left[\sum_{h=1}^{P} (X_{hj} - X_{hk})^2 \right]^{1/2}$$

where X_{hj} and X_{kh} are the values of the hth variable for individuals j and k, respectively. However, this simple form has serious problems in that it requires commensurable variables, and even then, is dominated by those variables with the largest variation. Hence, standardization is usually carried out by dividing the variables by their standard deviation, so that a difference between a pair of individuals with respect to any variable is expressed in units of standard deviations of that variable. The problem with such standardization is that it can have the effect of obscuring trends, by forcing a spherical shape on the data structure.

In the present study, advantage was taken of the fact that there was replicated sampling of cones at each location, and thus for each variable a standard deviation could be computed based on the portion of variation which is within locations. When this set of standard deviations is used to standardize the Euclidean distance, a spherical shape is not forced on the data, but rather there is greater dispersion in the distribution of sample points in sample space in the direction of those variables for which the variation between locations is largest in relation to their variation within locations.

Thus, the resemblance function used for investigation of the data structure is given by

$$e(j,k) = \left[\sum_{h=1}^{P} (\bar{X}_{hj} - \bar{X}_{hk})^2 / Q^2_h \right]^{1/2}$$

$$\text{where } Q_h = \left[\sum_{m=1}^{a} \sum_{h=1}^{P} \frac{(X_{hlm} - \bar{X}_{hm})}{a(n-1)} \right]^{1/2}$$

is the standard deviation of variable h within locations,
a is the number of locations,
n is the number of replicates per location,
X_{hlm} is the value of variable h for replicate l at at location m,
$\bar{X}_{hm}$ is the average value of variable h at location m.

Three-dimensional Views of the Data Structure

Principal components analysis was used to reduce the data to three dimensions to enable the construction of stereograms. The presence of non-linear variation, and the reduction of dimensionality may both cause distortion in the representation of the data structure in three dimensions. The extent of this distortion was examined using correlation of resemblance matrices.

Principal components analysis was first applied to the set of 240 cones, and a stereogram of the first three principal components was plotted. In this case, the first three principal components accounted for a total of 82% of the variance. Except for cone volume, which had a correlation of only −.05, all of the cone characters had correlations with the first principal component of magnitudes greater than.6. These large correlations are not unexpected since this first component accounted for 59% of the variance. The general conclusion that could be drawn is that the main variation among the cones (relative to variation within locations) was a variation in the serotiny and shape of the cones rather than their size. The largest correlations with the first principal component were for the following characters: apophysis height/width (.89), serotiny (.86), and apophysis height/length (.84). Thus, the main trend in the data expressed in the direction of the first principal component was an increase in cone serotiny and "knobbiness" of the apophyses. Associated with this trend was an increase in specific gravity, and increase in the ratio of cone width to cone length, a decrease in the symmetry of the cones, and a decrease in the ratio of cone volume to the volume of a cylinder of the same length and width (i.e. an increase in the extent to which the cones depart from a cylindrical shape). The second principal component had a moderately large negative correlation with specific gravity (−.57), and a moderately large positive correlation with cone volume (.51), and so appeared to correspond to a trend along which cones become larger in volume although their weight remains approximately constant so that their specific gravity decreases. The third principal component had a moderately large negative correlation with volume (—.64), although the correlation with specific gravity was relatively small (−.11), hence, this component appeared to correspond to a trend along which cone volume decreased while specific gravity remained approximately constant.

Examination of a stereogram of the 240 cones plotted on their first three principal components, revealed what was apparently a single cloud of points with no evident group structure. However, when this same data was replotted with each cone represented by a symbol coresponding to its geographic region (with separate symbols being used for the populations with and without serotinous cones in the coastal region), this single cloud was seen to consist of four groups *(Figure 2)*.

A much clearer picture of group structure was obtained when the "noise" associated with variation within locations was reduced by averaging the character values for each location *(Figure 3)*. In the principal components analysis of the 24 locations, 94% of the total variance was accounted for by the first three principal components. The first principal component, which accounted for 75% of the variance, had essentially the same interpretation as the first principal component in the analysis of the indvidual cones, but with even higher correlation with those characters expressing cone shape and serotiny. For example, the largest correlations were a correlation of .96 with apophy-

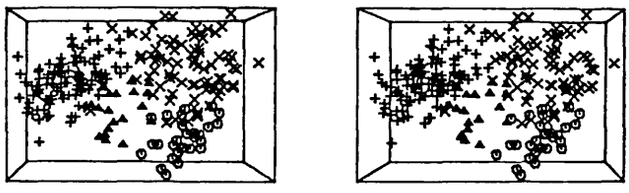

Fig. 2. — Stereogram of the 240 cones plotted on their first three principal components. (Use of a stereoscope is recommended to obtain a three dimensional view.) Circles identify the populations with serotinous cones of the coastal region. Triangles identify the populations with non-serotinous cones of the coastal region. Plus signs identify the populations of the Sierra Nevada region. X's identify the populations of the Rocky Mountain region.

sis height/width, and a correlation of .94 with cone serotiny. The second principal component had a moderately large negative correlation with specific gravity (−.52), and a moderately large positive correlation with cone volume (.51), a result which was similar to that for the second principal component for the individual cone data. However, there was a difference between the two cases in that the second principal component for the location means also had a moderately large positive correlation with the ratio of cone width to cone length (.51). This suggested that this character was of more importance for distinguishing variation among locations than it was for distinguishing variation among individual cones. In terms of geographic regions, this second component separated the Rocky Mountain locations from the closed-cone coastal populations. The third principal component for the location data had a moderately large positive correlation with cone volume (.64) and relatively small correlation with specific gravity (.03), thus, it had essentially the same interpretation as for the principal components analysis of the individual cones, only the sign of the axis was reversed with respect to cone volume.

The fact that the first three principal components accounted for 94% of the variance suggested that there was probably not much distortion in representation of the data structure for the locations by a three dimensional stereogram. However, it was considered desirable to investigate the extent of distortion due to reduction of dimensionality and possible non-linearity by an independent approach. The technique used was a correlation of matrices of resemblances between locations. One resemblance matrix being based on the full set of cone characters, while the other resemblance matrix was based only on the three first principal components. The correlation between the two sets of distances given by these two matrices was 0.993, indicating that the three dimensional stereogram provided a very accurate overall representation of the resemblances among locations.

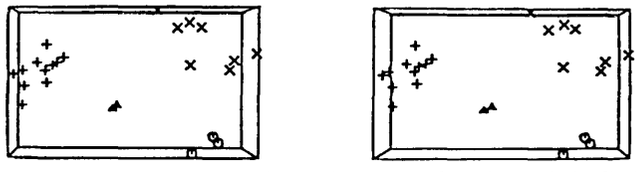

Fig. 3. — Stereogram of the 24 locations plotted on their first three principal components. Circles identify the populations with serotinous cones of the coastal region. Triangles identify the populations with non-serotinous cones of the coastal region. Plus signs identify the populations of the Sierra Nevada region. X's identify the populations of the Rocky Mountain region.

The conclusion was that there appeared to be a strong group structure in the data, conforming to the geographic regions recognized by Critchfield (1957), with the addition that the closed-cone population from the coastal mountains of northwestern California was grouped with the closed-cone population from Mendocino.

References

Albrecht, G. H.: Some comments on the use of ratios. Syst. Zool. 27 (1), 67—71 (1978). — Atchley, W. R.: Ratios, regression intercepts, and the scaling of data. Syst. Zool. 27 (1), 78—83 (1978). — Atchley, W. R. and D. Anderson: Ratios and the statistical analysis of biological data. Syst. Zool. 27 (1), 71—78 (1978). — Atchley, W. R., C. T. Gaskins, and D. Anderson: Statistical properties of ratios. I. Empirical results. Syst. Zool. 25, 137—148 (1976). — Corruccini, R. S.: Correlation properties of morphometric ratios. Syst. Zool. 26, 211—214 (1977). — Critchfield, W. B.: Geographic variation in *Pinus contorta*. Maria Moors Cabot Foundation. Publication No. 3, Harvard University, Cambridge, Mass. 118 pp. (1957). — Dodson, P.: On the use of ratios in growth studies. Syst. Zool. 27 (1), 62—67 (1978). — Hills, M.: On ratios -- a response to Atchley, Gaskins, and Anderson. Syst. Zool. 27 (1), 61—62 (1978). — Jancey, R. C.: Species ordering on a variance criterion. Vegetative 39, 1: 59—63 (1979). — Newman, K. W.: A multivariate study of geographic variation in the cones of *Pinus contorta* in relation to environmental variables. Ph. D. Thesis, University of Western Ontario (1982). — Orloci, L.: Multivariate analysis in vegetation research. Dr. W. Junk bv. The Hague. 451 pp. (1978). — Rall, L.: Geographic variation in *Pinus contorta*. M. S. thesis. University of Western Ontario (1979).

Genetic Structures and Expected Genetic Gains from Multitrait Selection in Wild Populations of Douglas fir and Sitka spruce

I. Genetic variation between and within populations

By Y. Birot and C. Christophe

Institut National de la Recherche Agronomique, Station d'Amélioration des Arbres Forestiers, Ardon, 45160 Olivet, France.

(Received 6th July 1982)

Summary

The paper reports results from nursery stage till age 12 on genetic variation between and within populations of two major conifer species: Douglas Fir and Sitka Spruce. The research is based on provenance-progeny test established with the IUFRO seed collection achieved in the late sixties, with two levels of sampling: population and single tree within population. For Douglas Fir, 371 open-pollinated progenies from 26 populations are under test whereas these numbers are respectively for Sitka Spruce of 292 and 21. Studied characteristics were mainly: growth phenology (bud set, flushing), form (branching, stem straightness).

Classical patterns of geographic variation were observed for both species. Genetic parameters (heritability, genetic correlations) varied from one population to another, especially for Douglas Fir, but also changed with the time. Additive effects were found surprisingly high for Sitka Spruce offering good prospects of future genetic gains. It is concluded that the genetic structure (between and wit-