most. However, there were large changes for seedlots represented in all 10 blocks. Overall, the changes in seedlot means seemed well worth the extra effort involved in the adjustment calculations.

The other important consequences concerned the analysis of variance (*Table* 1). In the red pine experirnent, the adjustments had little effect on the seedlot sums of squares but reduced the error sum of squares by almost 50%. As a result, the **F** value for seedlot rose from a barely significant 1.7 to a respectable 3.2, and the coefficient of variation was reduced to 10% of the mean.

General Applicability

The moving average method has been used with data from four other plantations having pronounced site differences within and between blocks. In each case, the improvement in statistical precision was considerable although not as marked as in the red pine provenance test. The method has also been tried with data from plantations without obvious large differences in site quality within blocks. In such cases, the adjustments amounted to random changes in plot means and had little effect on seedlot means, sums of squares or mean squares. In other words, unwarranted use of the proposed new method has done neither harm nor good.

Effectively, the moving average method causes an experiment such as the red pine provenance test to be considered as consisting of a very large number of small, incomplete blocks. Statistical efficiency becomes just as great as if a very efficient lattice or incomplete block design had been used when the experiment was installed.

In Michigan there are approximately 250 replicated genetic test plantations, of which 5 need to be analyzed by the rather laborious moving average method and the remainder can be analyzad by the simpler methads applicable to the randomized complete block design.

It would have been possible to achieve high statistical precision in all plantations by routine use of balanced incomplete block designs. However, the extra effort would have reaped commensurate benefits anly 2% of the time. Availability of the running average method makes it possible to use simple designs but attain the high precision inherent in the more complex designs if there is a need.

I requested views as to the legitimacy of the moving average method from three trained statisticians. One had reservations but the other two considered it valid. The one with reservations considered it preferable to base the height adjustments on some measure of site quality rather than on height itself. In other words, he favored analysis of covariance.

Multivariate Classification in Provenance Research

A comparison of two statistical techniques[*]

By E. R. Falkenhagen and St. W. Nash

Forest Research Institute, (P.O. Box 727)
Pretoria, Transvaal, South Africa
and
Department of Mathematics
University of British Colurnbia
Vancouver, British Columbia, Canada
respectively.

(Received June 1976 / January 1978)

Summary

The relationships of two kinds of multivariate statistical analysis are discussed and illustrated with data from a forest trae provenance study. Canonical correlation analysis of the biological data, taking into account geographical data, separated provenances and ecological regions more effectively than did canonical variate analysis (discriminant analysis). In each case Mahalanobis distances between provenances could be calculated. Distances under the first analysis were greater than under the second, but the pattern was fairly consistent from one analysis to the other.

Key words: Picea sitchensis (Bong.) Carr., provenances, canonical correlation analysis, discriminant analysis, Mahalanobis distance.

Résumé

Titre: Classification multivariable en recherches sur les provenances d'arbres forestiers: comparaison de deux techniques statistiques.

Les relations entre deux analyses statistiques multivariables sont discutées et illustrées a l'aide de données accumulées lors d'un test de provenances d'arbre forestier. L'analyse des corrélations canoniques des donnees biologiques, tenant compte des ooordonnees géographiques, a separé les provenances et les régions écologiques de façon plus effectice que l'analyse de variables canoniques (appellée aussi, analyse discriminante). Dans tous les cas, les distances de Mahalanobis entre les provenances ont pu être calculées. Lors de la premiere analyse, les distances s'avérèrent plus grandes que lors de la seconde, mais leur relations resterent passablement constantes d'une analyse a l'autre.

Zusammenfassung

Titel: Multivariable Klassifikation in der Herkunftsun-

[*] This paper grew out of research for a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Faculty of Graduate Studies, University of British Columbia, by the first author.

tersuchung: ein Vergleich zweier statistischer Verfahren.

An Hand von Beobachtungswerten aus einem Herkunftsversuch mit *Picea sitchensis* wurden zwei Verfahren der mehrdimensionalen statistischen Analyse auf deren Wirksamkeit hin miteinander verglichen. Danach hat die kanonische Korrelationsanalyse der biologischen Beobachtungswerte die Herkünfte und die ökologischen Gebiete wirksamer getrennt als die Diskriminanzanalyse (Analyse kanonischer Zufallsvariablen). In beiden Fällen konnten die verallgemeinerten Abstände von Mahalanobis zwischen den Herkünften berechnet werden. Während das Muster ziemlich übereinstimmte, waren die Abstände bei der kanonischen Korrelationsanalyse größer.

1. Introduction

The study of the biological parameters of the natural populations of a tree species is often considered to be a useful step in the study of the genetic variability of that species and in the breeding program for that species. It is also of interest in studies of the evolutionary differentiation and ecological adaptation of the species. Certain multivariate techniques can be used to give a clear, economical description of such biological variation.

The material considered here came from the cones of Sitka spruce (*Picea sitchensis* (Bong.) Carr.) collected during the fall of 1970 from British Columbia and Alaska by the International Union of Forestry Research Organizations (I.U.F.R.O.), Section 22, "Working group on the procurement of seed for provenance research". (A provenance is, strictly speaking, the location of a population of trees growing in close proximity and from which seed has been collected; by extension it is this population of trees itself.) The collections were made in each of 39 locations (provenances), which range from 48.38° to 58.37° north latitude and from 121.93° to 134.58° west longitude. Elevations range from 0 to 2200 feet above sea level. In 33 of the provenances the collections were made from 15 trees. (The collection from each tree was kept separate.) This study is based only on these 33 provenances.

Five seeds were randomly sampled from each tree and mounted on a sheet. Then seed-wing length and width, seed length and width were each measured to the nearest 0.01 millimeter. Ten cones were randomly selected from each tree and their length measured to the nearest millimeter. The basic figures of this study are tree averages of these five characters.

The provenances were provisionally grouped into five geographical regions on the basis of the bioclimatic and physiographic data available:

I. Eastern Vancouver Island and the Lower Mainland,
II. Western Vancouver Island,
III. the Queen Charlotte Islands,
IV. Souheastern Alaska, and
V. the Skeena and Nass River drainages and adjacent areas.

In region I, provenance 2 at Squamish was the only one from the Lower Mainland for which data from 15 trees were available.

The locations of the provenances are shown in *figure 1*. The basic data upon which this paper is based are summarized in *tables 1 and 2*.

2. Methods

This section describes the types of multivariate statistical analysis used in analyzing the data, and their rationales. There are no mathematical derivations nor any detailed instructions for computation. To avoid lengthy formulas, we have used vector and matrix notation. Lower-case letters will denote vectors; capital letters will denote matrices.

Readers well versed in multivariate statistical methods

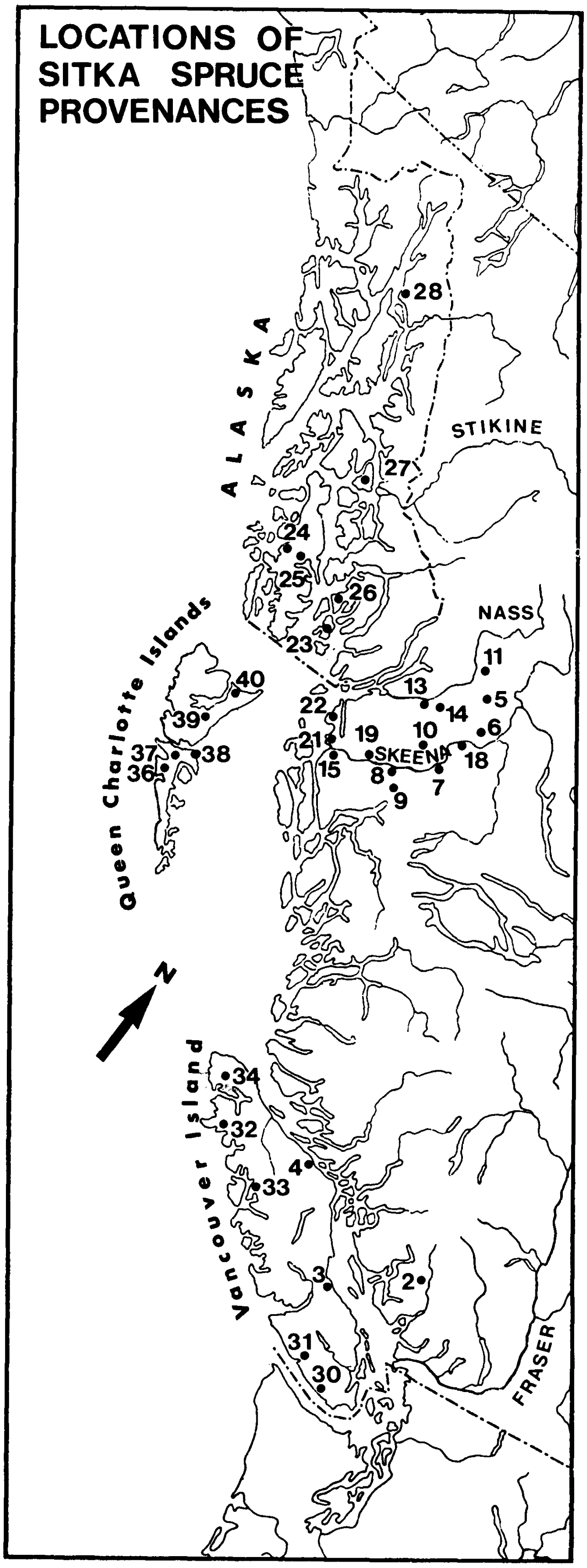

Figure 1.

Table 1. — Basic data: provenance means

Region	Provenance	Latitude	Longitude	Elevation	Wing	Wing	Seed	Seed	Cone
		($^0\frac{1}{100}$'s)	($^0\frac{1}{100}$'s)	above sea level	length	width	length	width	length
		North	West	(feet)	(mm.)	(mm.)	(mm.)	(mm.)	(mm.)
		y_1	y_2	y_3	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$
I	2	49.92	123.25	100	8.062	3.583	3.077	1.838	66.807
	3	49.38	124.62	0	6.965	3.415	2.949	1.771	64.900
	4	50.38	125.95	0	6.819	3.408	2.824	1.825	60.193
II	30	48.38	123.87	0	7.036	3.405	3.040	1.754	59.313
	31	48.58	124.40	25	7.254	3.535	2.915	1.806	63.627
	32	50.08	127.50	100	6.353	3.216	2.577	1.754	52.513
	33	49.83	126.67	10	7.059	3.444	2.849	1.863	57.840
	34	50.62	128.12	100	6.703	3.311	2.789	1.836	56.280
III	36	52.87	132.08	50	6.509	3.507	2.855	1.859	49.200
	37	53.05	132.08	200	7.937	4.067	3.110	1.955	68.987
	38	53.13	131.80	250	6.695	3.415	2.751	1.765	49.833
	39	53.50	132.17	300	6.949	3.685	2.827	1.897	53.833
	40	53.92	132.08	0	6.959	3.607	2.699	1.769	66.260
IV	23	55.03	131.55	0	6.343	3.448	2.503	1.757	44.520
	24	55.50	133.13	0	7.776	4.010	3.038	1.991	64.507
	25	55.47	132.67	0	7.131	3.813	2.813	2.011	62.620
	26	55.42	131.70	50	7.561	3.685	3.005	1.963	58.833
	27	56.58	132.73	25	6.837	3.596	2.863	1.898	55.573
	28	58.37	134.58	100	7.735	3.607	2.967	1.927	58.580
V	5	55.47	128.23	1700	7.503	3.495	2.726	1.681	40.647
	6	55.17	127.87	2200	7.529	3.414	2.775	1.695	42.600
	7	54.63	128.40	450	8.006	3.618	2.889	1.807	59.587
	8	54.40	128.95	100	7.996	3.911	3.149	1.965	66.080
	9	54.13	128.62	550	7.179	3.581	2.743	1.766	54.713
	10	54.72	128.77	450	7.529	3.691	2.995	1.815	56.460
	11	55.68	128.68	800	6.544	3.293	2.481	1.741	38.820
	13	55.15	129.22	50	7.085	3.392	2.551	1.885	51.300
	14	55.15	128.97	1300	7.314	3.621	2.745	1.875	53.067
	15	54.20	129.92	0	7.295	3.714	2.831	1.989	62.120
	18	55.02	128.32	800	7.819	3.721	2.829	1.788	44.500
	19	54.28	129.42	100	7.375	3.721	2.856	1.859	60.393
	21	54.20	130.25	50	7.450	3.952	2.973	2.090	59.967
	22	54.27	130.32	2100	6.466	3.652	2.721	1.837	49.733

Table 2. — Dispersion matrices (mean squares and products)
(Individual observations are tree means.)

	x_1	x_2	x_3	x_4	x_5
B, between provenances			32 degrees of freedom		
x_1	3.7715 00				
x_2	.9724 275	.6271 925			
x_3	.8689 403	.3169 219	.4158 769		
x_4	.2416 392	.2173 505+	.1166 947	.1410 951	
x_5	25.1485 1	12.3603 6	13.8937 9	6.3409 44	948.9303
W, within provenances			462 degreees of freedom		
x1	.5599 613				
x_2	.1258 203	.1118 063			
x_3	.0544 4052	.0188 8568	.0399 4569		
x_4	.0193 1643	.0159 2320	.0091 1842 6	.0139 5238	
x_5	1.6982 10	.5930 816	.4867 753	.1635 873	53.6379 9
F	6.7353	5.6096	10.4111	10.1126	17.6914
		$F_{\cdot 005}$ (32,462) = 1.8101			

and those interested only in botanical results may go directly to section 3 and refer back to sections 2.1 and 2.2 as need arises. However, one should have some idea of what a Mahalanobis distance is and how it can arise, not only as one aspect of canonical analysis, but also as a by-product of canonical correlation analysis. In the latter case the standardization needed for the new variates of canonical correlation analysis is different from the standardization commonly used.

2.1 *Discriminant analysis and canonical variate analysis.* The concept of a linear discriminant function was originally introduced by R. A. Fisher in 1936 as a solution to the problem of classifying an observation into one of two predetermined groups. Linear discrimination between two groups is widely used and its calculation procedures can be simplified to be a slight variation of those of multiple regression. We therefore go directly to the general problem of linear discrimination between q groups.

Consider then q populations, each sampled randomly for p characters. We assume that the populations have a common covariance matrix, which is estimated by a pooled within-group sample covariance matrix W. Let B denote the sample between-groups covariance matrix. The observations from any population may be pictured as a swarm of points of more or less ellipsoidal shape. Since there is a common covariance matrix, the ellipsoids have the same orientation and size, which for the samples are measured by W. The sample means too form a swarm, whose configuration may also be ellipsoidal, but usually of different orientation and size than the other ellipsoids. This orientation and size are measured by B.

Discriminant analysis, also called canonical analysis, can be divided conceptually into two steps. First, the p variates are transformed in such a way that the swarm of points from each sample becomes spherical in shape. This spherical shape means that the within-group random error variation has been standardized so as to be the same in every direction. The swarm of sample means too has been changed in shape and size, but will still be ellipsoidal. Second, the new coordinate axes are rotated so as to be parallel to the principal axes of the new ellipsoid representing the swarm of sample means. The two steps together give a set of canonical variates, which not only are uncorrelated but also are unaffected by what kinds of units of measurement were used in recording the original observations. The first canonical variate gives variation along the longest principal axis of the ellipsoid representing the swarm of sample means, the second canonical variate gives variation along the second longest principal axis, and so on. As much of the between-groups variation as possible is concentrated in the first few canonical variates.

Canonical analysis must not be confused with principal component analysis. The latter omits step one above entirely and simply rotates the coordinate axes to be parallel to the principal axes of, say, the original ellipsoid of the swarm of sample means. The principal component variates will depend on what kinds of units of measurements were used in recording the original observations. The purpose of step one will not be achieved, even if the original variates are standardized by dividing each by its standard deviation. The swarm of points from any sample does not become spherical under such standardization: it remains ellipsoidal, but will have its projections on the several standardized coordinate axes all equal. Unlike principal component analysis, canonical analysis is not simply a rotation of the coordinate system.

The idea of canonical analysis is to replace the p-component observational vector x by a vector z of as low a dimension as possible that will most clearly and economically bring out the differences between the populations. One seeks a vector a_1 of coefficients such that the variate $z_1 = a'_1x$ has the maximum possible F-ratio $(a'_1Ba_1)/(a'_1Wa_1)$. Here and hereafter we adopt the standard convention that $a'_jWa_j = 1$, so that the jth F-ratio is simply a'_jBa_j. After z_1 has been found, one seeks a vector of coefficients a_2 such that a'_2Ba_2 is maximized, subject to the condition that $z_2 = a'_2x$ should be uncorrelated with z_1. And so one continues. Let $\Theta_1 \geqq \Theta_2 \geqq \ldots \geqq \Theta_m > 0$ denote the non-zero roots of the determinantal equation $|B - \Theta W| = 0$. Here m is the lesser of p and $(q - 1)$. Then a_j is the solution of linear equations $(B - \Theta_jW)a_j = 0$, with

$$a'_iWa_j = \begin{cases} \mathrm{Var}(z_j) & = 1 \text{ if } i = j \\ \mathrm{Cov}(z_i,z_j) & = 0 \text{ if } i \neq j, \end{cases} \quad \text{and } a'_iBa_j = \begin{cases} \Theta_j & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

The canonical variates $z_1, \ldots, z_m$ remain invariant under all non-singular linear transformations of the vector x.

Graphing observations, using canonical variates, results in less overlap and greater separation of samples from different populations than does graphing using the original x-variates. Canonical variates and their graphs are also useful in picking out clusters of related populations.

In cluster analysis, and in discriminant analysis as well, it is useful to use the Mahalanobis squared distance D^2. The squared distance between points x_1 and x_2 is

$$D^2\,(x_1,x_2) = (x_1 - x_2)'W^{-1}\,(x_1 - x_2) = (z_1 - z_2)'(z_1 - z_2) = \sum_{j=1}^{m} (z_{j1} - z_{j2})^2,$$

where $z_{jk} = a'_jx_k$ and $z'_k = (z_{1k}, \ldots, z_{mk})$, $k = 1,2$. Thus D^2, like z, is invariant under linear transformations of x, and, in the space of the canonical z-variates, becomes the square of an ordinary Euclidean distance. In clustering, the smaller $D^2\,(\bar{x}_h,\bar{x}_k)$ is, the stronger may the affinity between the hth and kth populations be considered to be. In this paper we do not go into formal cluster analysis, but use Mahalanobis distances more informally to learn the degrees of affinity of the various provenances.

It needs to be pointed out that the actual distance is D, not D^2. D has all of the usual properties of a distance, such as the triangular inequality, namely that $D\,(x_1,x_2) \leqq D\,(x_0,x_1) + D\,(x_0,x_2)$.

In discriminant analysis let x_0 be a new p-variate observation coming from one of the q populations, but from which being unknown. Then one can set up the following rule for deciding from which of the q populations x_0 comes. The rule is: Assign x_0 to that population for which $D^2\,(x_0,\bar{x}_k)$ is smallest.

Blackith and Reyment (1971) have objected that this rule does not work when the relative abundance of the several populations varies from population to population. However, the rule can be modified in a simple way to take care of this situation. Let π_k denote the relative abundance of the kth population, $\pi_1 + \ldots + \pi_q = 1$. Then π_k is the probability that x_0 comes from the kth population. The larger π_k is, the wider should be the region in which x_0 is assigned to the kth population. Often there is a great deal of empirical information available about the π_k. When the π_k can be well estimated from such information, their use in discriminant scores increases the proportion of new individuals assigned to the right population. The usual linear discriminant score is

$$\begin{aligned} L\,(x_0,\bar{x}_k) &= \bar{x}'_kW^{-1}x_0 - \tfrac{1}{2}\,\bar{x}'_kW^{-1}\bar{x}_k + \log_e\pi_k \\ &= \bar{z}'_kz_0 - \tfrac{1}{2}\,\bar{z}'_k\bar{z}_k + \log_e\pi_k. \end{aligned}$$

The rule then is: Assign x_0 to that population for which $L\,(x_0,\bar{x}_k)$ is largest. But this is equivalent to the rule: Assign x_0 to that population for which $D^2\,(x_0,x_k) - 2\,\log_e\pi_k \geqq 0$ is smallest. When all of the π_k are equal, one can simply omit the $\log_e\pi_k$ term.

We have followed Seal (1964) in using the term, "canonical analysis", for the analysis described in this section. However, this usage is far from universal, and the same term is often applied to the canonical correlation analysis described in the next section. Since the analysis of this section is formally a special case of the analysis of the next section, the similarity of terms does not matter too much. Another terminology for the analysis of this section is Anderson's (1958) "characteristic roots and vectors of B in the metric of W", which suggests the term, "principal component analysis of B in the metric of W".

2.2 *Canonical correlation analysis.*

Suppose there are two sets of variables measured on the same individuals or on the same sampling units. In our example the first set of variables consists of the cone length and seed measurements, and the second set consists of the geographical coordinates (latitude, longitude, elevation) of the provenances. Let the p_1-vector x denote the first set of variables, the p_2-vector y denote the second set of variables. A random sample of size n gives the sums of products matrix

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

for total variation of the observations about their means: $(n - 1)^{-1}S_{11}$ is the sample covariance matrix of the x-variates, $(n - 1)^{-1}S_{22}$ is that of the y-variates, and $(n - 1)^{-1}S_{12}$ is the matrix of sample covariances of the x-variates with the y-variates, $(S_{21} = S'_{12})$. In canonical correlation analysis one seeks vectors a_1,b_1 of coefficients such that the variates $u_1 = a'_1x$, $v_1 = b'_1y$ have the maximum possible correlation. After u_1, v_1 have been found, one seeks vectors of coefficients a_2,b_2 such that $u_2 = a'_2x$, $v_2 = b'_2y$ have maximum correlation, subject to the condition that u_2, v_2 are uncorrelated with u_1, v_1. And so one continues. Let $\varrho^2_1 \geqq \varrho^2_2 \geqq \ldots \geqq \varrho^2_m > 0$ denote the non-zero roots of the determinantal equation $|S_{12}S^{-1}_{22}S_{21} - \varrho^2S_{11}| = 0$, which are always the same as the non-zero roots of $|S_{21}S^{-1}_{11}S_{12} - \varrho^2S_{22}| = 0$. Here m is the lesser of p_1 and p_2. Then a_j is the solution of linear equations

$$(S_{12}S^{-1}_{22}S_{21} - \varrho^2_jS_{11})a_j = 0, \text{ and}$$

Furthermore, ϱ_j is the correlation in the sample between $u_j = a'_jx$ and $v_j = b'_jy$, and is called the jth canonical correlation between x and y. Geometrically the canonical correlations are a measure of the extent to which individuals occupy the same relative positions in the p_1-dimensional x-space and the p_2-dimensional y-space.

The vectors of coefficients are usually standardized so that

$$\alpha'_jS_{11}\alpha_j = \beta'_jS_{22}\beta_j = (n - 1),$$
$$\alpha'_jS_{12}S^{-1}_{22}S_{21}\alpha_j = \beta'_jS_{21}S^{-1}_{11}S_{12}\beta_j = (n - 1)\varrho^2_j$$

However, in the present study, we are primarily interested in working out the affinities of the various provenances of Sitka spruce, and will use another standardization which relates the variates of canonical correlation directly with the canonical variates of the previous section. It is easy algebraically to show that the equations given above to solve for the ϱ_j, a_j and b_j are equivalent to the following.

$$\text{Let } B_1 = p^{-1}_2S_{12}S^{-1}_{22}S_{21}, \quad W_1 = (n - p_2 - 1)^{-1}(S_{11} - S_{12}S^{-1}_{22}S_{21}),$$
$$B_2 = p^{-1}_1S_{21}S^{-1}_{11}S_{12}, \quad W_2 = (n - p_1 - 1)^{-1}(S_{22} - S_{21}S^{-1}_{11}S_{12}),$$

$$\text{and } \frac{\varrho^2}{(1 - \varrho^2)} = \frac{p_2\Theta}{(n - p_2 - 1)} = \frac{p_1\Phi}{(n - p_1 - 1)}\,.$$

Solve $(B_1 - \Theta_jW_1)a_j = 0$ and $(B_2 - \Phi_jW_2)b_j = 0$ for a_j and b_j respectively. Standardize the vectors of coefficients so that

$$a'_jW_1a_j = b'_jW_2b_j = 1,$$
$$\text{giving} \quad a'_jB_1a_j = \Theta_j \text{ and } b'_jB_2b_j = \Phi_j.$$

B_1 is the sample covariance matrix of x due to regressions of the x's on y; W_1 is the sample covariance matrix of x due to residual error after the removal of the effects of the regressions of the x's on y. B_2, W_2 have analogous interpretations with the roles of x and y reversed.

In terms of our example, even if the sample covariance matrix W (of section 2.1) of trees within provenances had not been available to estimate purely random variability, the covariance matrix W_1 above, which represents residual random variability of the botanical variables about their regressions on the geographical variables, can be used to estimate such purely random variablity. Thus the first conceptual step of canonical analysis (described in section 2.1) can use W_1 to so transform the botanical variables that the purely random variability becomes spherical, that is, is the same in every direction. B_1 is the sample covariance matrix representing that part of variation between provenances which is systematically related to the geographical factors. Since between and within covariance matrices, B_1 and W_1, are available, the canonical analysis of section 2.1 applies, and one can also get Mahalanobis distances. But the results are entirely equivalent to the canonical correlation analysis described earlier in this section. There is a similar analysis for geographical variables using B_2 and W_2.

There are now two sets of Mahalanobis distances, one in the x-space, one in the y-space. The squared distances between points $P_1 : (x_1,y_1)$ and $P_2 : (x_2,y_2)$ are

$$D^2_1(x_1,x_2) = (x_1 - x_2)'W^{-1}_1(x_1 - x_2) = (u_1 - u_2)'(u_1 - u_2) = \sum_{j=1}^{m}(u_{j1} - u_{j2})^2,$$
$$D^2_2(y_1,y_2) = (y_1 - y_2)'W^{-1}_2(y_1 - y_2) = (v_1 - v_2)'(v_1 - v_2) = \sum_{j=1}^{m}(v_{j1} - v_{j2})^2,$$

where $u_{jk} = a'_jx_k$ and $u'_k = (u_{1k}, \ldots, u_{mk})$,
$v_{jk} = b'_jy_k$ and $v'_k = (v_{1k}, \ldots, v_{mk})$, $k = 1,2$.

The two standardized forms of the vectors of coefficients b_j of $(S_{21}S^{-1}_{11}S_{12} - \varrho^2_jS_{22})b_j = 0.$ are related by the equations

$$a_j = \left\{\frac{(n - p_2 - 1)}{(n - 1)(1 - \varrho_j)}\right\}^{1/2}\alpha_j, \qquad b_j = \left\{\frac{(n - p_1 - 1)}{(n - 1)(1 - \varrho^2_j)}\right\}^{1/2}\beta_j.$$

3. Application of the analyses to the Sitka spruce data

3.1 *Application of canonical correlation analysis*

Forest genecology suggests that the phenotypic variation of the Sitka spruce parent trees is closely related to geographical location. Accordingly we first consider a canonical correlation analysis of the parental traits and the geographical coordinates of the provenances studied.

Let seed and cone traits (seed-wing length and width, seed length and width, cone length) be the set x, and geographical coordinates (latitude, longitude, elevation) the set y. The results of the canonical correlation analysis are given in *table 3*. The coefficient vectors are given in their second standardized forms as described in section 2.2. To gauge the relative importance of the contribution of a variable to a canonical correlation variate, the coefficient corresponding to that variable should be multiplied by its standard deviation between provenances. Sample variances and covariances involving the x's used provenance means as the basic observations, rather than observations of individual trees.

The canonical correlations are significantly greater than zero at probability levels .0001, .005, and .01 respectively. Thus variation in seed and cone traits is closely related to variation in geographical coordinates. *Figure 2* shows how the first pair of canonical correlation variates vary with

Table 3. — Results of the canonical correlation analysis

$u_j = a'_j x$ $v_j = b'_j y$		The canonical correlations of x with y ϱ_1 0.8476 72	 ϱ_2 0.7140 66	 ϱ_3 0.5895 15+
Vari-able	Standard deviation	Coefficients of the variables (2nd standard form)		
		a_1	a_2	a_3
x_1	0.5014 31	2.7922 5—	—2.4421 9	—2.3790 8
x_2	0.2044 82	3.3628 0	6.7757 0	4.6015 5+
x_3	0.1665 09	—5.5838 0	—0.6814 19	5.4093 7
x_4	0.0969 863	1.1864 8	4.5632 1	—12.5944 9
x_5	7.9537 43	—0.1874 86	—0.0339 646	0.0726 075+
		b_1	b_2	b_3
y_1	2.4794 90	0.7623 73	—0.3193 50—	—0.5970 80
y_2	2.8293 26	—0.3032 43	0.6481 81	0.3536 59
y_3	605.8763	0.0010 1051 7	0.0002 1426 4	0.0021 1151

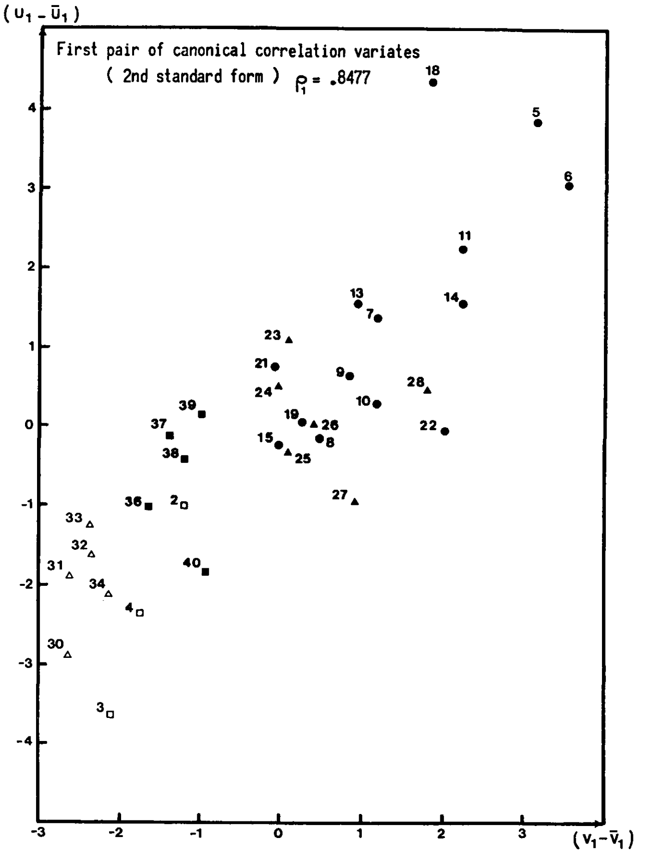

Figure 2. — Key to the provenances: Region I, □; II, △; III, ■; IV, ▲; V, ●.

each other. v_1 being the geographical variable, u_1 the botanical variable. The other two pairs of canonical correlation variables, (v_2, u_2) and (v_3, u_3), tell much the same story, though their correlations are lower.

First we consider what information the geographical coordinates can give us about seed and cone traits. *Figure 3* shows the locations of the provenances on the (v_1, v_2) graph, where $v_j = b'_j y$. This is a modified geographical representation of the location of the provenances, with some input from the botanical x-variates through the use of the covariance matrix W_2 in the calculation of the coefficients b_j. We point out once again that the transformation from the y-variables to the v-variables is not simply a rotation of coordinate axes. Furthermore the transformed points are scattered in three dimensions just as surely as were the original points in the rugged terrain of British Columbia and Alaska. Hence any two dimensional graph can only give a partial picture of the scatter of the transformed points. In *figure 3* the various regions are entirely distinct, and even within regions the relative positions of most provenances are what one would expect from the map. But coastal provenance 22 stands apart from other provenances of region V and deviates part way towards region IV because of its elevation.

The first variate v_1 represents mainly southwest to northeast variation from maritime to inland, mountainous conditions. The second variate v_2 represents mainly a southeast to northwest variation parallel to the coast. The third variate v_3 seems to bring out the contrast between low inland provenances such as 13 and high coastal provenances such as 22. The v-variates are uncorrelated and are the linear combinations of the geographical variables most highly correlated each with a corresponding appropriate linear combination of the botanical variables, which we consider next.

We consider now what information seed and cone traits can give us about geographical locations. *Figure 4* shows the locations of the provenances on the (u_1, u_2) graph, where $u_j = a'_j x$. This representation seems to give a much better picture of what the botanical variation is like. This is because the botanical variables x enter directly into the calculation of u, and not just indirectly through the calculation of coefficients a_j, b_j, as happens in the calculation of v. There is considerable overlap of regions III and IV, the Queen Charlotte Islands and Alaska, and of the lower part of the Skeena in region V. Nevertheless, there is a strong resemblance between *figures 3 and 4*. The description of the geographical meaning of the axes is about the same in both cases.

Further to study the analyses using u and v above, one can plot other pairs of coordinates of the provenances, such as (u_1, u_3), (u_2, u_3), or even make a three-dimensional model. However, in an analysis involving more than three canonical dimensions, there is no visual representation giving all dimensions in one figure. It is in this situation that Mahalanobis distances between pairs of points are especially useful. The distances give essentially all of the information needed to construct the configuration of the points, and in a form easy to grasp. *Table 4* gives the squared distances between pairs of provenances. To save space but still bring out the essential points, we give the squared distances just between the provenances in region V. The provenances are ranked according to increasing distances as far as possible. In the lower left half of the table, below the diagonal, are the D^2_1 ($\bar{x}_h$, $\bar{x}_k$) based on the botanical traits; in the upper right half of the table, above the diagonal, are the D^2_2 (y_h, y_k) based on geographical location. D^2_1 is more comparable to the squared distances found in the

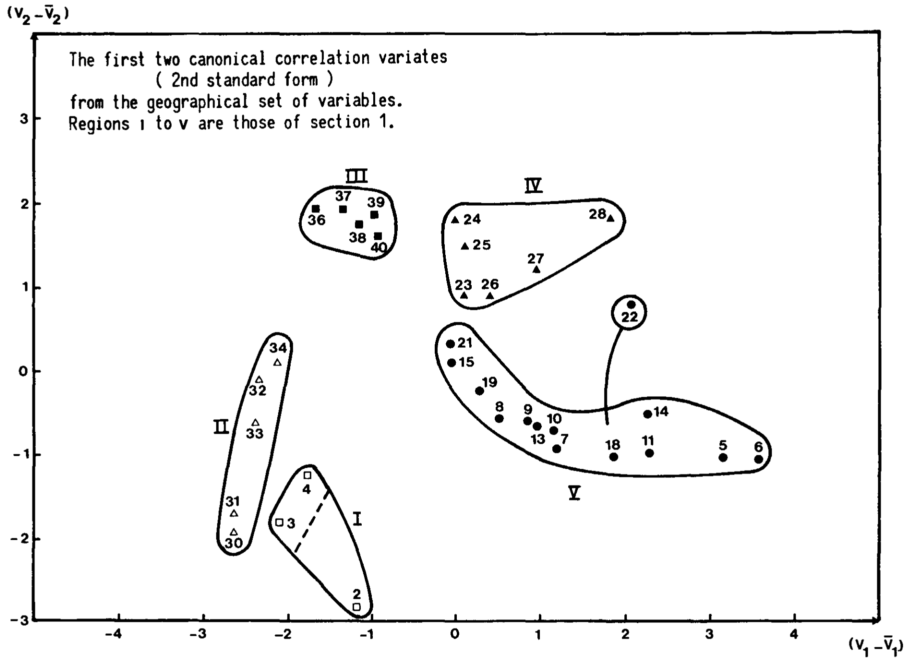

Figure 3

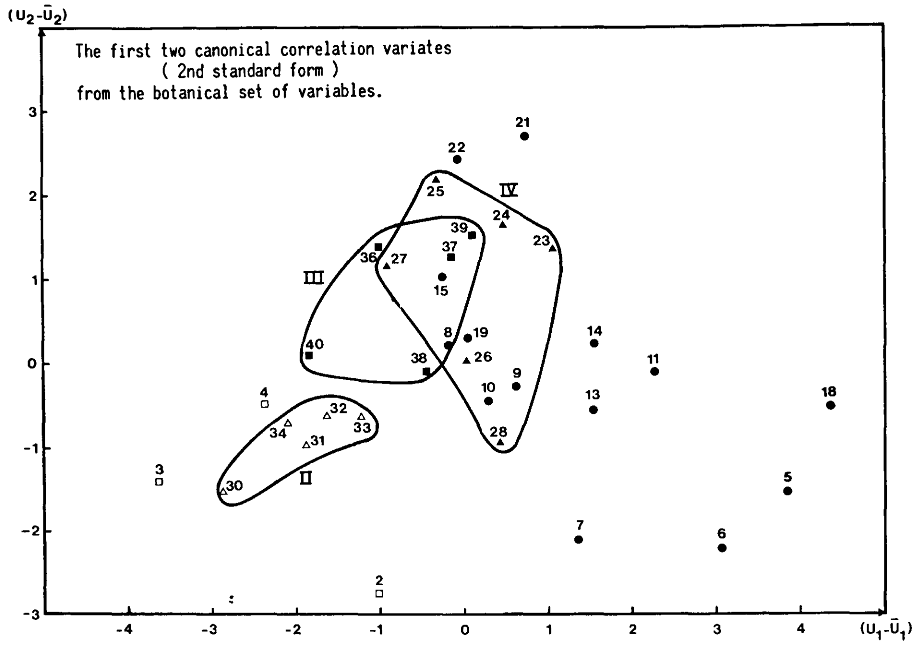

Figure 4

discussion of canonical analysis in the next section (3.2) than is D^2_2.

Squared distances, and perhaps even more so distances themselves, are good measures of the affinities of pairs of provenances. They are thus useful in clustering procedures. However, we do not pursue this matter further in this paper. In comparing *table 4* of squared distances with *figures 3 and 4,* the reader should remember that is is not squared distances but rather distances themselves which obey the triangular inequality, and also that the figures cannot show the contribution of a third dimension (v_3 or u_3) to the total distance between two provenances.

3.2 *Application of canonical variate analysis (discriminant analysis).*

We turn now to the canonical variate analysis of the seed and cone traits. This will give us the maximum discrimination between provenances, when we are working in the metric of W, the error covariance matrix. W will be the within-provenance, between-trees, sample covariance matrix; B the between-provenances sample covariance matrix. Since the basic observations are now means for individual trees rather than provenance means, B is 15 times the $(n-1)^{-1}S_{11}$ of the previous section (3.1), there being 15 trees observed per provenance.

The results of the canonical analysis involving all 33 provenances are given in *table 5.* To gauge the relative importance of the contribution of a variable to a canonical variate, the coefficient corresponding to that variable should be multiplied by its standard deviation between

Table 4. — Mahalanobis squared distances between the provenances of region V, calculated for the two spaces of the canonical correlation analysis. The lower left half gives D^2_1 $(\bar{x}_h, \bar{x}_k)$ below the diagonal. The upper right half gives D^2_2 (y_h, y_k) above the diagonal. Provenances are ranked according to increasing distances as far as possible.

Provenance	21	15	8	19	14	13	9	22	10	7	11	18	5	5	D^2_2
21		0.10	1.38	0.49	9.03	2.89	2.03	23.29	2.64	3.24	7.37	5.87	24.92	16.47	21
15	3.99		0.81	0.21	9.23	2.03	1.85+	25.32	2.16	2.62	6.71	5.36	25.43	16.54	15
8	7.32	1.54		0.23	7.56	0.41	1.12	27.08	0.68	0.78	3.69	2.88	22.43	13.45—	8
19	8.12	3.73	0.74		7.60	1.11	1.06	24.80	1.08	1.38	4.75—	3.64	22.70	14.05+	19
14	7.31	5.07	2.98	2.80		8.33	3.26	9.10	3.90	4.17	2.40	1.87	4.12	1.16	14
13	11.89	5.71	5.03	7.40	2.66		2.13	30.76	0.94	0.89	3.05+	2.82	23.39	13.74	13
9	13.42	8.73	3.30	1.25+	3.09	8.81		17.68	0.37	0.57	2.31	1.24	14.22	7.72	9
22	13.04	17.52	13.85—	9.11	15.25+	29.34	9.86		21.53	22.81	20.68	18.56	7.20	9.93	22
10	16.36	10.83	4.46	1.91	5.25—	11.81	0.29	9.45+		0.05+	1.34	0.78	15.51	8.12	10
7	23.74	12.78	7.83	8.97	5.76	3.34	7.31	33.32	8.97		1.19	0.70	15.69	8.17	7
11	16.14	15.44	9.44	6.24	3.52	10.47	2.79	12.94	3.95—	9.43		0.23	10.50—	4.28	11
18	28.10	29.85+	23.31	19.79	10.40	16.20	13.70	29.30	16.26	15.01	4.46		10.13	4.21	18
6	35.48	29.14	19.69	16.65+	11.14	14.71	9.72	32.43	10.76	7.43	5.07	4.57		1.38	6
5	37.11	35.15+	25.40	20.82	13.90	20.50+	12.80	30.90	14.04	14.31	4.95—	2.04	1.53		5
D^2_1	21	15	8	19	14	13	9	22	10	7	11	18	6	5	Provenance

Table 5. — Results of the canonical analysis for all 33 provenances.

		The characteristic roots of B in the metric of W				
	$z_j = a'_j x$	Θ_1 19.860	Θ_2 9.356	Θ_3 6.581	Θ_4 3.308	Θ_5 2.219
Variable	Standard deviation (from W)	Coefficients of the variables				
		a_1	a_2	a_3	a_4	a_5
x_1	0.7483 06	0.2262 83	—0.6095 73	0.6515 15—	—1.3392 5+	—0.1258 33
x_2	0.3343 74	—0.0767 183	0.3540 32	0.7308 12	1.8117 3	3.1187 0
x_3	0.1998 64	—1.0176 2	—3.7546 4	0.9024 48	3.7936 3	—1.9654 2
x_4	0.1181 20	—2.3418 7	6.8315 1	4.3573 7	—1.8169 1	—4.5582 9
x_5	7.3237 96	—0.1162 67	0.0121 653	—0.0803 238	—0.0372 569	0.0249 186

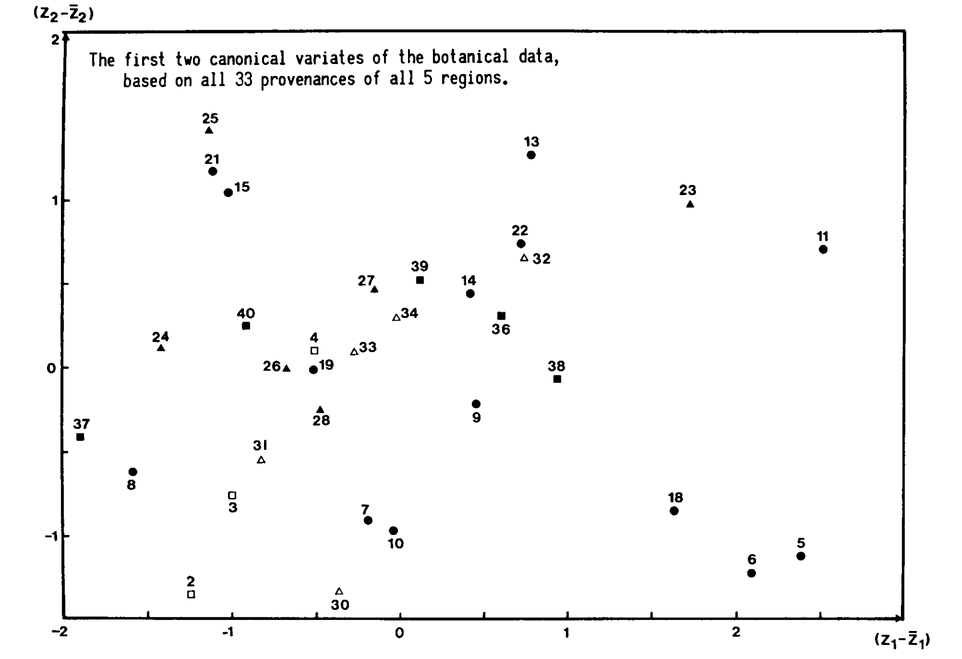

Figure 5

trees within provenances. The characteristic roots $\Theta_1, \ldots, \Theta_5$ are each significantly greater than zero at probability level .0001. This tells us that there are statistically significant differences between provenance means in each of the canonical variates.

Figure 5 shows the locations of the provenances on the (z_1, z_2) graph. There is much more overlap of the five regions here than in the canonical correlation analysis as depicted in *figures 3 and 4.* Nevertheless, if one inspects the regions separately, he finds here more or less the same configurations of provenances as occured in the previous figures. What is the explanation for so much overlap of regions? The canonical variate analysis takes no particular account of geography, but attempts to reveal all differences between provenances, whether local or regional.

Table 6 below gives the correlations of the canonical variates $z_1, \ldots, z_5$ with the canonical correlation variates v_1, v_2, v_3, representing geographical variation. One notes that z_1 and z_3 are most highly correlated with v_1, z_2 with v_2, z_4 with v_3. Since v_1, v_2, v_3 are uncorrelated and have com-

Table 6. — Correlation r_{ij} of v_i and z_j.

	z_1	z_2	z_3	z_4	z_5	$\varrho^2_i = \sum_j r^2_{ij}$
v_1	.5122	—.0473	.5814	—.2355—	.2445—	.7179
v_2	—.1400	.5276	.2150+	.3432	.2172	.5091
v_3	.3353	—.1514	—.1930	.4017	.1369	.3527
$\sum_i r^2_{ij}$	.3944	.3035+	.4215+	.3346+	.1257	1.5797

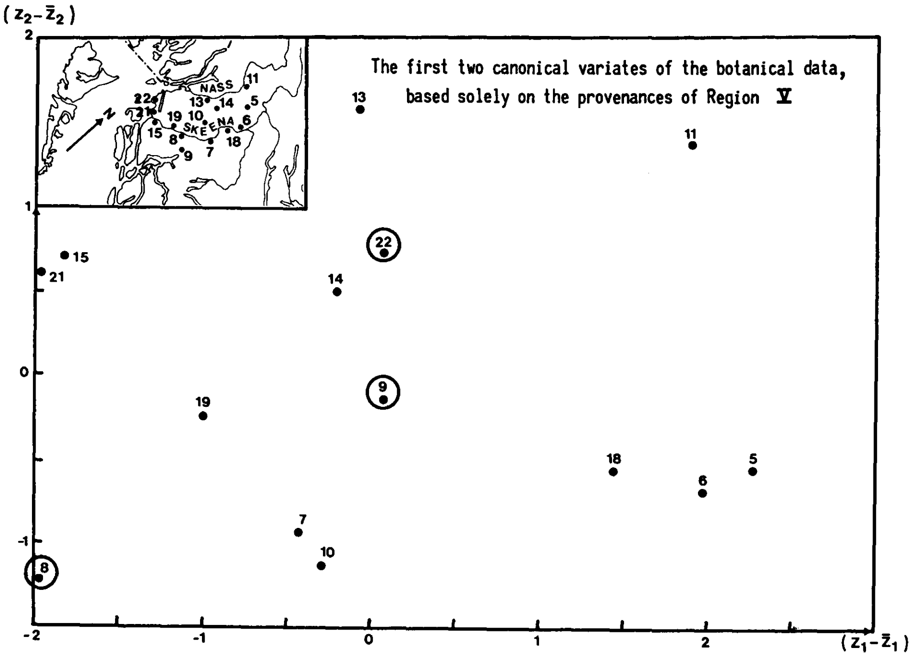

Figure 6

Table 7. — Mahalanobis squared distances between provenances of region V, calculated on the basis of canonical analyses of seed and cone traits. The lower left half gives D^2_{All} based on the analysis of all five regions. The upper right half gives D^2_V based on the analysis of region V alone. Note that on the average $D^2_V\ (\bar{x}_h,\bar{x}_k) = 1.11\ D^2_{All}\ (\bar{x}_h,\bar{x}_k)$.

Provenance	21	15	8	19	14	13	9	22	10	7	11	18	6	5	D^2_V
21		1.46	4.09	4.08	4.58	7.70	8.06	6.59	7.38	8.28	17.70	13.52	19.02	20.88	21
15	1.50		3.89	1.89	2.80	4.42	4.96	5.70	6.14	5.08	14.98	12.95	16.73	18.81	15
8	3.78	3.63		2.62	6.46	12.76	6.57	9.74	3.25	3.81	22.70	12.78	16.63	19.27	8
19	4.05	1.80	2.42		1.58	5.29	1.23	3.43	1.68	1.59	11.63	7.49	9.91	11.53	19
14	4.18	2.41	5.52	1.43		1.74	1.16	2.02	3.07	2.69	5.51	4.20	6.27	7.31	14
13	7.11	3.85	11.34	4.92	1.68		4.30	4.99	8.97	6.46	5.06	8.40	9.84	10.90	13
9	7.69	4.53	5.83	1.05	1.20	4.22		2.23	1.70	1.77	6.36	4.09	5.03	5.99	9
22	6.20	5.18	8.68	3.11	2.19	5.21	2.20		4.30	6.97	5.05	5.56	7.66	7.99	22
10	6.94	5.70	2.81	1.60	2.85	8.42	1.63	4.05		1.63	11.39	4.37	5.86	7.41	10
7	7.95	4.82	3.48	1.65	2.48	5.94	1.78	6.91	1.67		12.01	5.71	6.84	8.72	7
11	15.31	12.57	19.19	9.80	4.67	4.44	5.47	4.61	9.78	10.34		4.83	4.34	3.91	11
18	12.03	11.63	10.75	6.83	3.93	8.09	4.12	5.92	3.93	5.01	4.78		1.05	1.10	18
6	16.97	14.80	14.08	8.79	5.66	9.14	4.72	7.61	5.08	5.83	4.14	1.06		0.26	6
5	18.45	16.50	16.20	10.03	6.48	10.08	5.42	7.77	6.31	7.36	3.81	1.03	0.25		5
D^2_{All}	21	15	8	19	14	13	9	22	10	7	11	18	6		Provenance

mon variance one, $\sum_i r^2_{ij}$ in the last line of the table is the proportion of the variance of z_j between provenances due to regression of z_j on v, or equivalently, on the geographical locations of the provenances. z_3 shows the strongest influence of geographical location, followed in order by z_1, z_4, z_2, and far behind by z_5. Similarly $\varrho^2_i = \sum_j r^2_{ij}$ in the right hand column of the table is the proportion of the variance of v_i due to regression of v_i on z, or equivalently, on the seed and cone traits. It is in fact the square of the ith canonical correlation of the previous section. (The numerical discrepencies between the values of ϱ^2_i above and those obtainable from *table 3* are due entirely to round-off errors.)

When the canonical analysis is done on a regional basis, the configuration of provenances bears a stronger resemblance to the results of the previous section (3.1), unobscured by the overlapping of the several regions. We will illustrate this using the canonical analysis based just on the data from region V, the Skeena and Nass river drainages and adjacent areas. *Figure 6* gives a small map of the region and also the (z_1, z_2) graph of the provenances. Only three of the provenances seem to be out of place: 22, a high elevation coastal provenance; 9, above the head of Douglas Channel, the fiord to Kitamat; and 8, the nearest neighbor to 9, on the Skeena. Compare the locations of the 14 provenances of region V on this graph with their locations on *figure 5*. Similar results were obtained for the other regions.

The overall relationships of the provenances in the canonical analyses are best comprehended by studying a table of Mahalanobis distances. *Table 7* gives the squared distances between pairs of provenances in region V. In the lower left half of the table, below the diagonal, are the $D^2{}_{All}$ $(\bar{x}_h, \bar{x}_k)$ based on the canonical analysis of all 33 provenances; in the upper right half of the table, above the diagonal, are the $D^2{}_V$ $(\bar{x}_h, \bar{x}_k)$ based just on the data from region V. Note that on the average the entries in the upper right half of the table are about 1.11 times the corresponding entries in the lower left half of the table. Similarly the entries of $D^2{}_1$ in the lower left half of *table 4* are on the average about 1.97 times the corresponding entries of $D^2{}_{All}$ in the lower left half of *table 7*. The relationship between two kinds of distance is, however, more consistent in the first comparison than in the second.

In dealing with the same set of points for provenance means, the magnitude of distances depends on the error covariance matrix used. A covariance matrix representing small random variation gives a larger Mahalanobis distance than a covariance matrix representing large random variation. If two such covariance matrices have different correlation structures, the first may represent more variation in some directions, the second more variation in other directions. In the present case W_1 of the previous section (3.1) represents much smaller variation in most directions than the W covariance matrices of the present section. In the present section the W matrix based on region V alone represents slightly less variation in more directions than does the W based on all five regions. This explains the relationships of the three sets of Mahalanobis distances.

4. Conclusions

Extensive seed and cone measurements of Sitka spruce were analyzed by two multivariate statistical methods. The sample material was gathered in the fall of 1970 from 33 provenances in British Columbia and Alaska, scattered all the way from the Strait of Juan de Fuca to Lynn Canal in Alaska.

The two multivariate statistical methods, namely canonical variate analysis and canonical correlation analysis, are first described, then applied to the Sitka spruce data. In each case the differences between provenance mean vectors can be expressed in terms of distance functions, which are independent of the scales of measurement of the original measurements, and which take into account the correlations of the several variates. Both methods enable one to delineate biological zones, based on the cone and seed data collected, which in turn provide important information as far as future regeneration of the species is concerned. This delineation is more clearly seen in the canonical correlation analysis. The canonical variate analysis, with two extra dimensions at its disposal, also takes into account variation between provenances which is not of a geographical nature.

How Sitka spruce actually varies with geographical location is best seen in *figure 4* and the lower left half of *table 4*, which are based on the canonical correlation analysis of the botanical variables. There is a gradual but steady change in seed and cone characters as geographical location varies. The provisional regions with which we started may be revised somewhat. Regions I and II on Vancouver Island may be combined. Provenance 2 on the Lower Mainland deviates from the Vancouver Island provenances in the same direction that the inland provenances in region V deviate from the coastal provenances there. Provenances 8, 15, 19, and 21 on the lower Skeena and along the coast are more akin to the provenances of the Queen Charlotte Islands than to the more inland provenances of region V. The inland provenances of region V should definitely form a separate region. There may well be similar differentiation elsewhere between the provenances of Sitka spruce in inland valleys and those along the coast, but the data at hand for inland valleys only come from region V and provenance 2. Region IV in Alaska is not sharply divided from region III and the maritime part of region V, but rather shows the gradual botanical change from them related to the corresponding geographical change. The high coastal provenance 22 in region V and northerly provenance 28 in region IV stand somewhat apart from their fellows.

The two methods of analysis can prove useful in making as clear as possible the nature of variation among populations of a tree species. Canonical variate analysis has been much more extensively used in the past than the other method. But canonical correlation analysis deserves to be used much more extensively than heretofore, especially when one has data on a number of environmental variables as well as data on some biological variables of the species being studied.

A more detailed biological and silvicultural interpretation of the canonical analyses are presented in a paper which forms a second part of this one (Falkenhagen, 1978).

References

Anderson, T. W.: An introduction to multivariate statistical analysis. John Wiley & Sons, New York (1958). — Blackith, R. E. and Reyment, R. A.: Multivariate morphometrics. Academic Press, London and New York (1971). — Burley, J. and Burrows, P. M.: Multivariate analysis of variation in needles among provenances of *Pinus kesiya* Royle ex Gordon (syn. *P. khasya* Royle; *P. insularis* Endlicher). Silvae Genetica **21**: 69—77 (1972). — Falkenhagen, E. R.: Parent tree variation in Sitka spruce provenances: an example of fine geographic variation. Silvae Genetica **27**: 24—29 (1978). — Falkenhagen, E. R.: Genetic variation in 38 provenances of Sitka spruce. Silvae Genetica **26**: 67—75 (1977). — Fisher, R. A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7**: 179—188 (1936). Reprinted in Collected papers of R. A. Fisher, edited by J. H. Bennett, paper 138 in volume III: 465—475. The University of Adelaide (1973). — Jeffers, J. N. R. and Black, T. N.: An analysis of variability in *Pinus contorta*. Forestry, J. Soc. Foresters of Great Britain, **36**: 199—218 (1963). — Kshirsagar, A. W.: Multivariate analysis. Marcel Dekker, New York (1972). — Seal, H. L.: Multivariate statistical analysis for biologists. Methuen & Co., London; and John Wiley & Sons, New York (1964). — Van den Driesche, R.: La recherche des constellations de groupes à partir des distances généralisées D^2 de Mahalanobis. Biométrie-Praximétrie **6**: 36—47 1(965).