EVALUATION OF VARIABILITY IN A FOG GRASS (*HOLCUS* spp.) GENE POOL

by

SIAN HOCK TEOW

A thesis

presented in partial fulfilment of the requirement for the

Master

of

Agricultural Science

Massey University

Palmerston North

New Zealand

1978

ABSTRACT

The utimate aim of a breeder's working collection is to utilize the genetic variation for breeding new cultivars. Before this variation can be utilised, it is necessary to obtain the description of attribute from the collection. These can either be obtained from the records of genetic resources (base collection or active collection) or obtained directly from the working collection itself. This will resulted in a huge amount of data. To be of any value, this information need to be classified systematically, and the classification need high degree of objectivity, especially for species of no a priori knowledge.

A Working collection of 160 Yorkshire Fog seed populations, from all over New Zealand, were planted out in Massey University. This formed the gene pool. During Summer 1975, 11 agronomic and morpho-logic characters were scored in a semi-quantitative scale. This yield approximately 42,000 data records. These were then systematically reduced to 550 by a series of multivariate analysis techniques. The procedures of Multivariate Analysis of Variance, Multiple Discriminant Analysis and Cluster Analysis were reviewed and their computer programmes were developed.

The clustering behaviours of seven agglomerative polythetic strategies were studied and compared, using the full set of characters. Most of the results concurred with studies carried out by other workers The Minimum Increment Sum of Squares strategy was found to be most suitable for this analysis. A probabilistic decision method was devised to decide objectively, the truncating point for clustering.

For all set of data, the studies did not reveal any ecotypes and hence did not agree wirh the ecoclinal trends hypotheses (of Yorkshire Fog in New Zealand) of Jacquee.The approaches of both studies (of Yorkshire Fog in New Zealand) of Jacques. The approaches of both studies (that of Jacques and the present one) were reviewed critically and a more appropriate approach was suggested for future ecological study.

Preliminary results revealed that there were a few promising groups showing agronomic desirable characters. They were promising breeding materials for future lines selection.

Of all the characters studied, flowering date and clump erectness were found to be the most discriminating characters amongst groups, and the most dominant characters in clustering. These implied that selection should be beneficial, if they had moderate high predictive heretability.

ACKNOWLEDGEMENTS

I am deeply indebted to my supervisor, Dr. I. L. Gordon for his guidance and assistance during the course of the work and the preparation of this thesis. Without him this thesis would not have been completed.

Special gratitude must go to my parents, Mr. and Mrs. Teow Lee-Seng, my brothers and sisters for providing me the opportunity to study in New Zealand.

Thanks to Professor B. R. Watkin and all the staff members of Agronomy Department for their assistance and encouragement.

I wish to thank my sisters, Bee Ling, Bee Geok and Bee Ing for their considerable time, effort, patience in typing this thesis. Their help is gratefully acknowledged.

Special thanks are also due to : Dr. D. A. Ratkowsky of CSIRO Hobart, Australia and Dr. R. Pringle of Mathematics Department, Massey University, for their advices on statistical field; Mr. Terry Lynch and the field technicians of Agronomy Department, for maintaining the working collection of Yorkshire Fog; staff members of computer center and library for their assistance; and all those who help in one way or others.

The awards of D. J. McGowan and Helen E. Akers scholarships to partially finance this study are gratefully acknowledged.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

FIGURES Page

TABLES

TABLES Page

TABLES		Page

INTRODUCTION

Hill country pasture is important in New Zealand, as it is a highland country, with 28 million acres (64 percent) of the farming land being steep hill country (Scott 1956).

Hill country pasture production can only be maintained if suitable species are grown. In particular, attention must be given to the species' soil fertility requirements. It is common for high producing pasture species to lose their producing ability where soil fertility falls below their requirements, or where practices to suit their best growth are not followed. In the hill country, soil fertility is often poor (White 1973), and species suit to this are preferred.

Yorkshire Fog grass (Holcus lanatus L.), is noted for a virtual absence of edaphic specialization, wide climatic tolerance, low soil fertility requirement, and good adaptation to extensive agricultural (pastoral) system. This grass has established well in New Zealand in humid hill country, unploughable steep hills, acidic peat soils and even swampy lands (Basnyat 1957). It is the major constituent of some 8 million acres of North Island marginal pasture land, of which 5 million acres are in the wetter hill country of the west. It's contribution toward farm productivity has been judged as significant (Munro 1961, Basnyat 1957).

In 1953, Jacques started to investigate Yorkshire Fog as a useful pasture species. The investigation commenced with collecting a wide range of local seed populations from most major areas of New Zealand. After a series of progeny tests and selections, a synthetic cultivar "Massey Basyn" was bred in 1960 (Jacques 1962). A synthetic cultivar is made up of genotypes which have previously been tested for their ability to produce superior progeny when crossed in all combinations (i.e. they have good combining abilities) (Allard 1960).

In a sward productivity trial, Munro (1961) found that "Massey Basyn" compared favourably with perennial ryegrass and performed much better than a commercial Fog line. Riveros(1963) found that dry matter yields were always significantly higher in Yorkshire Fog swards, than in ryegrass swards. In a trial from 1961 to 1964, Watkin and Robinson(1974) Found that "Massey Basyn" had a similar dry matter yield to the ryegrasses (Ariki, Manawa and Ruanui); and the seasonal production of "Massey Basny" well distributed, with relatively good production in winter and summer.

To enable further detailed investigation, and to improve further the agronomic value of Yorkshire Fog grass, Dr. R. G. Clements organized another collection of Fog accessions in 1972. A total of 201 seed populations were collected. These included 108 from the North Island, 89 from the South Island, 3 from Australia and 1 from Spain. This collection was sampled from most parts of New Zealand, even the Westland region (Which Jacques had missed out), and the Northland region (which Munro(1961) and Jacques(1962, 1974) considered as having the most potential for highly productive genes).

Out of the 201 seed populations collected, 160 were planted out as a working collection at Massey University. According to Vavilov, "collection and classification" is the first phase of scientific plant breeding (Frankel 1947); and the aim of the presant study was to examine the phenotypic variability in, and to reveal the relationship amongst, major characters within this collection. In this study, populations have been grouped into phenotypically-similar clusters. The clusters so obtained could be used as sources for breeding material. Several multivariate analysis techniques have been used to achieve this. Firstly, multivariate analysis of variance has been used to investigate whether there were any differences amongst populations. Secondly, multiple discriminate analysis has been used to estimate the relationships between the phenotypic characters, and to ordinate the original scores into uncorrelated discriminant scores. Finally, cluster analysis has been used to group together populations with similar phenotypic patterns of dispersion.

CHAPTER 1. LITERATURE REVIEW

1.1. Yorkshire Fog

Yorkshire Fog (Holcus lanatus L.) probably has it origin in the Iberian Peninsular (Vinal and Hein 1937). However, it is found throughout Europe and North West Africa, and in the most recently developed temperate farming areas of Americas, South Africa and Australasia (Hubbard 1945, Munro 1961).

Yorkshire Fog was introduced into New Zealand either as a seed impurity or as a hay grass during the 1860's. It has spread throughout the mainland and even to the remote Chatham and Auckland Islands (Chesseman 1923). It is one of the most widespread and adaptable grasses introduced from Western Europe. It has established in pasture of diverse type over a wide altitude range (Basnyat 1957). Despite its relatively recent introduction to New Zealand, natural selection has probably taken place, and a considerable number of ecotypes appear to have developed. The variability in New Zealand is believed to be similar to that in the centre of origin. New Zealand is believed to be a secondary centre of diversity of this species (Munro 1961, Jacques 1962, 1974).

Yorkshire Fog has shown an almost complete absence of edaphic specialization. It is capable of growing in a wide range of soil types, from heavy loams to sands (Hubbard 1945). Though the optimum soil pH is considered to be within 5.0 to 7.5, it is found growing on areas with much hgiher acidity (Davies 1944). It also grows on areas of extreme soil moisture content (Hubbard 1945). However, its growth will become aggressive only where the soil moisture content is "adequate" (Jacques 1962). Levy(1970) suggested that Fog will tolerate tidal salt mud flats, swamp and flood areas, very wet or water-logged areas and soils with average moisture; but it will not tolerate moderately dry, dry and extremely dry soil. Yorkshire Fog is classified as "lower fertility" grass (Suckling 1960). It will dominate on soils with fertility which is high, moderately high, average, moderately low, or low (Levy 1970). But on soils with extremely high fertility, very low fertility, or extremely low fertility, it will not dominate (Levy 1970). Though the exact physiological basis of these wide tolerances is not known, Munro(1961) suggested that the anatomy, competative absorbing ability,and the endrotrophic mycrorhiza of the roots are important factors.

Yorkshire Fog exhibits a wide tolerance of temperature and light regimes. A high rate of growth is maintained at temperatures ranging from 12.8° to 29.4°C (Mitchell and Lucanus 1960); and at light ranging from dense shade to open and sunny (Levy 1970). Even during the winter season, where temperature and light could be limiting to other species, growth and new tiller formation continue (Munro 1961). It has always been regard as a good winter grower (Hubbard 1945, Watkin and Robinson 1974).

Crown rust (Puccini coronata) is the major disease, which will not kill the infested plant but reduces its palatability and yield. Yorkshire Fog is regarded as being relatively less palatable than other pasture grasses. Pubescence and extremely prostrate growth forms are believed to be the factors (Jacques 1962).

In the sward, growth of fog is centred on leaf expansion on a moderate number of large tillers, whereas in ryegrass and brown top, it is centred on large number of smaller tillers (Munro 1961). In growth form classification, Levy(1970) classified fog as heaving its crown at or above ground level. the growing points are elevated above ground (Jacques 1962). The leaves are soft, being comparatively low in strengthening tissue (Jacques 1962). The general growth habit, and mode of vegetative reproduction of Fog, are most suited to a lenient system of grazing. It's grazing tolerance lies between perennial ryegrass and cocksfoot (Mitchell 1956). It is probable that its feeding value is high, as a result of its low proportion of both strengthening tissue and collateral vascular bundles (Jacques 1962).

In most New Zealand pastures, the flush of growth is in spring, with more variable production in Summer and Autumn (depending on district rainfall), while winter production is low. The winter (June, July, August) dry matter yield of New Zealand pastures range from 0 to 13% of the total annual yield (Radcliffe 1974a, Radcliffe 1974b, Radcliffe and Cossens 1974, Radcliffe 1975a, Baars et.al.1975, Baars 1976b, Radcliffe 1976, Round-Turner et. al.1976, Rickard and Radcliffe 1976). However, in areas where either summer growth is restricted by moisture strees, such as Gisborne plains (Radcliffe and sinclair 1975), and Wairarapa and Hawke's Bay (Radcliffe 1975b), or where winter growth is encouraged by high temperatures, such as in Northland (Baars 1976a, the winter yield may reach 15-17% of the total annual yield.

Similar patterns occurred in unimproved hill country where winter low production is always the limiting factor determining stocking rate, and spring flush is poorly utilized (White 1973). The good winter growth of Fog might be able to ease this limitation. Watkin and Robinson(1974) have show that "Massey Basyn" not only has higher total yield, but also a more even seasonal distribution of yield. It's winter yield was 16% of total, as compared to 11.5% for Ruanui, 14.7% for Manawa and 12.5% for Ariki ryegrasses.

In summary, these findings indicate that Fog is well suited to less intensive farming systems, typical of many dairy pastures and upland sheep farms. It would seem especially suited to the humid, low fertility North Island hill country.

1.2. Gene Pool Concept And Maintenance

Response to selection is based on genetic variation in the original population (Allard 1960, Bennett 1970). Therefore, it's true that plant breeding's success is dependant on this variation. In the progress from Neolithic to scientific plant breeding, not only the method of selection, but also the nature and range of variation has changed. Intense and directional selection for modern "improved" cultivars has reduced the genetic variability generally utilized in agriculture (Bennett 1970). This is especially so when the "improved" cultivars are either selected for uniformity (as in purelines or multilines), or selected for closely defined objectives.

Modern scientific farming, which enable widespread cultivation of relatively few "improved" cultivars, not only intensified the tendency, but also threatened to wipe out the broad genetic variation of primitive cultivars by encroaching on their habitats. The primitive cultivars of wheat, coffee and barley in Ethiopia are under such a threat (Mengesha 1975). Also the introduction of wheat from CIMMYT, and of "miracle" rice from IRRI, to Latin Americas ans Asia provide the same threat to their primitive cultivars (Frankel and Bennett 1970).

The adaptability of a population will decrease as the genetic variability decreased. Most of the "improved" cultivars are well adapted to a restricted range of environments. These they can productively exploit but at the expenses of their adaptability. Simmonds(1962) noted, critically, that such a sacrifice of adaptability is unwise. Finlay and Wilkinson(1963) also suggested that selection should aim for "general" rather than "special" adaptation. However, the amount of adaptability retained should depend on the degree of environmental fluctuation. On the other hand primitive populations have greater genetic variability and adaptability. These features should be preserved for future exploration and exploitation by plant breeders (Frankel and Bennett 1970).

The goal of preserving genetic diversity of plants, in genetic resource collection or gene pools was originated in 1920 by N. I. Vavilov (Frankel 1975). To preserve these variations, the initial collecting has to be extensive, and, also, the resulting collection has to be maintained carefully. The problems and procedures of exploring the centres of diversity, the optimum statistical sampling techniques and sample sizes, and the procedures of collecting have been reviewed by Frankel and Bennett(1970), and Hawkes(1975).

The maintenance of collections has two distint but interrelated aims: firstly, to conserve the maximum amount variability for future generations; secondly, to allow plant breeders easy access to utilize this variation (Marshall and Brown 1975). Two types of collection are proposed generally to suit these aims: (i) base collections, for long term conservation; (ii) active collections for: (a) medium term conservations: (b) regeneration; (c) multiplication and distribution; (d) evaluation; and (e) documentation. The detail requirements for maintaining this two collections are outlined in Frankel(1975). Breeder's working collections are different from the above two, and are actually derived from active collections. However, they may generate valuable information which should be incorporated into the genetic resources records (Frankel 1975).

Other problems associated with maintenance are whether the collection should be maintained as seed, or as living plants, and whether it should be maintained as separated or bulk populations. Simmonds(1962) proposed that a bulk living collection, or "Mass gene reservoir", was best for long term conservation, as he regard seed collection, or "museum collection", as a wasting resource with high rate of losses. However, in simulated "mass reservoir" study of barley, Jain(1961), Allard and Jain(1962) and Clegy <u>et</u>. <u>al</u>.(1972)

found that under a common environment, "mass reservoirs" not only retain a small portion of the genetic variation, but also tended to retain the same spetrum of the variation. They concluded that "mass reservoirs" have little value in preserving genetic variation. Moreover, the technology and facility of seed storage has been improved such that seed of many species can be stored for a longer time before regeneration is neccessary. Nevertheless, completely "static" preservation is impossible, and loss of genetic variation can occur through differential survival of genotypes in storage, and through selection, hybridization and genetic drift during the seed rejuvenation process (Allard 1970).

Collections maintained as separated populations provide more flexible usage. This allows any subset of the whole collection to be used, whereas collection maintain as bulk population had to be used as a whole. Thus seed stored as a separated populations is most widely practiced (Marshall and Brown 1975).

1.3 Multivariate Analysis

Multivariate analysis is the simultaneous analysis of data from several correlated random variables, orginating from independent individuals. The use of a series of univariate analysis on each variable separately is often inadequate. It can overestimate the true dimensionality of divergence, as it does not separate covariance among the variables from their apparant variances. This may result in declaring too many significant differences. In general, multivariate analysis can reveal the relationships, interdependence and relative importance of the characters examined (Bryant and Atchley 1975, Kshirsagar 1972).

The main "obstacle" to the application of multivariate analysis has been the computational work involved. However, with the modern availability of high speed computers, this obstacle hardly exists today. Various multivariate techniques based on different statistical models are being used more frequently. Researchers are often faced with the problem of selecting the appropriate technique for their particular hypothesis and types of data. To aid this, a brief key to the major multivariate techniques is provided in appendix A-1.

Multivariate analysis techniques have been used in almost every field of biological research. To name a few of the recent ones: physical anthropology (Howells 1970, 1971), behavioural science (Cooley and Lohnes 1971), medical science (Kshirsagar 1972), taxonomy (Clifford and Stephenson 1975) and ecology (Crovello 1970, Pritchard and Anderson 1971). These techniques have also been applied in Agronomy, such as in pasture grazing trials (Williams and Edye 1974). They have been used also in Plant Breeding, such as in selection of parents from a gene pool (Bhatt 1970, 1976), in genotype-environment interaction studies (Mungomery et. al.1974, Shorter et. al.1977), and in screening gene pools (Burt et. al.1971, Edye et. al.1973).

1.4 Assumptions for Multivariate Analysis

The results from statistical analysis are strictly valid only, when the data conform to the basic assumptions underlying the analysis. If the assumptions are not fulfilled, the validity, efficiency (i.e. the accuracy of estimating the population parameter from the sample statistic), and sensitivity (i.e. the fineness of actual differences detected) will be affected. The essential assumptions for multivariate analysis are derived from those of univariate analysis, as considered by Cochran(1947) and Eisenhart(1947). There are some less common methods that either require no assumptions (such as nonparametric methods), or need only some of the assumptions (such as multivariate time series and stochastic processes) (Kshirsagar 1972). However, these methods can only be used to summarize properties of the data in hand, and could not be used to infer properties of the population from which the data were drawn (Andrew et. al.1971, Eisenhart 1947).

Multivariate analysis can be expressed in symbolic forms. If there are p variables, $x_1, x_2, x_3, \ldots\ldots\ x_p$ observed from an individual x, then this can be expressed in vector form as

$$\underset{(p\times 1)}{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_p \end{bmatrix}.$$ If there are n individuals observed, then the whole

data matrix can be expressed as

$$\underset{(p\times n)}{X} = \begin{bmatrix} x_{11} \cdots\cdots\cdots\cdots x_{1n} \\ x_{21} \qquad\qquad x_{2n} \\ \vdots \qquad\qquad\quad \vdots \\ x_{p1} \cdots\cdots\cdots\cdots x_{pn} \end{bmatrix} \text{ with mean, } \underset{(p\times 1)}{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

Then the variance matrix of X is defined as

$$\underset{(p\times p)}{\Sigma} = \begin{bmatrix} var(x_1) & cov(x_1\ x_2) \ldots\ldots & cov(x_1\ x_p) \\ \vdots & var(x_2) & \vdots \\ cov(x_p\ x_1) & \ldots\ldots\ldots\ldots\ldots & var(x_p) \end{bmatrix}$$

It is usually assumed that $\underset{(p\times 1)}{X} \sim N\left(\underset{(p\times 1)}{\mu}, \underset{(p\times p)}{\Sigma}\right)$. If X's are obtained from $g(g \geq 2)$ populations, then it is further assume that all Σ matrices are equal i.e. $\Sigma_1 = \Sigma_2 = \Sigma_3 = \ldots\ldots\ldots\ldots = \Sigma_g$.

1.4.1. Multivariate Normal Distribution of Variables

Multivariate analysis assumes that the variables are drawn from a population with a multinormal distribution. The multivariate normality test is difficult. A series of tests for goodness of fit (Rohlf 1971), or for probability curve fitting (Press 1971), may be used to examine the univariate normality (marginal normality) separately. However, conformity of all variates to marginal normality does not automatically imply multivariate normality, because of interactions amongst characters(Andrews et. al.1971, Press 1971, Rohlf 1971).

In general, directly observed biological measurements and qualitative characters do not diverge too far from multivariate normal, and can be assumed to fit this distribution (Seal 1968).

Under moderate non-normality, F-test and two tail t-test of univariate data will still be valid, unless large departures occured in the outlying regions (Bartlett 1947). However, their efficiency and sensitivity will be affected more. In general, univariate non-normality tends to lead to underestimation of the significance levels. This results in declaring too many significant differences and increasing type I error. Type I error is the probability of rejecting the true null hypothesis (Freund 1972). Similary, in multivariate cases, Ito (after Press 1971) has shown that when sample sizes are very large, violation of the multivariate normality assumption has only slight effect on testing hypothesis by MANOVA (multivariate analysis of variance).

The choice of directions of the canonical axes and discriminant axes have nothing to do with normality and so are less sensitive to non-normality. However the subsequent significance tests still depend on normality (Seal 1968).

1.4.2 Independence of Observations

Multivariate analysis assumes observations from different individuals to be independent, though the variables (attributes) may be correlated (Kshirsagar 1972). If the observations are not independent, the estimates of variance will be biased by this covariance, and this will affect the analysis. Observations from random individuals are usually considered to be independent.

1.4.3 Additivity of Effects

As in univariate case, some multivariate analysis (such as MANOVA) assumes the main and environmental effects are additive. Non-additivity (i.e. interaction) will cause overestimation of error and increase the significance level, i.e. increase type II error. Type II error is the probability of accepting the false null hypothesis. Unless the non-additivity is very severe, the effect can be neglected, for most of the tests are still valid.

1.4.4 Equality of Variance-covariance Matrices

The dimension and orientation of the ellipsoid-shaped multivariate normal population are characterized by the variance-covariance matrix. The variances represent its dimension (size) whereas the covariances determine its orientation (direction) (Seal 1968, Cooley and Lohnes 1971.) Inequality in variance-covariance matrices is caused when the dimension and/or the orientation of the corresponding ellipsoids are different. Standardization will only standardize the dimension and turn the variance-covariance matrices into correlation matrices. Their orientations will not be the same, unless the correlation matrices are equal (Seal 1968).

If the observations are sampled from more than one population, most of the multivariate analyses assume the variance-covariance matrices to be equal (i.e. equal in the dimension and orientation of the ellipsoids). However, some methods may need only the equality of correlation matrices (i.e. equality in orientation of the ellipsoids), e.g. canonical variate analysis (Seal 1968), and multiple discriminant analysis (Cooley and Lohnes 1971).

At present there is no criterion to test the equality of correlation matrices. However the equality of variance-covariance matrices can be tested by Bartletts' Chi-square (Seal 1968) or Box's M criterion (Cooley and Lohnes 1971). These tests are reasonably powerful in detecting inequality (Cooley and Lohnes 1971), but very sensitive to non-normality (Seal 1968). Non-normality is the main cause of inequality in variance-covariance matrices. Frequently, in large samples, non-normality leads to the rejection of the null hypothesis that the variance-covariance matrices are equal (Press 1971). If the null hypothesis is rejected mainly because of differences in dimensions (i.e. variances), but one believes that their orientations (i.e. the correlations) are approximately equal, canonical variate analysis and multiple discriminant analysis are still valid (Seal 1968).

When variance-covariance matrices are unequal, Anderson(1963) and Press(1961, 1971) have proposed other methods than MANOVA. (These are the multivariate extensions of Behrens-Fisher problem). They are seldom used, because they are more complex and subject to other restictions. Also, it is generally believed that MANOVA is still robust, even under this situation (Cooley and Lohnes 1971, Press 1971).

In univariate analysis, heterogenity of variance affects efficiency and sensitivity most, and validity of the F-test is least affected (Cochran 1947).

1.4.5 Transformations

Bartlett(1947), Corchran(1947) and Eisenhart(1947) have summarized the consequences and remedial methods when the assumptions are not fulfilled in univariate analysis. Based on an extension of the Box and Cox (1964) likelihood method, Andrews et.al.(1971) have proposed methods for obtaining data-based transformations of multivariate observations. The fact that the characters may be correlated means the value of marginal transformation (transforming one variable at a time) is doubtful. Though functions for joint transformation (transforming all the variables simultaneously) are available (Andrews et. al.(1971) they are very complex (Press 1971), and their validity becomes doubtful as the dimensionality of data increases (> 3) (Andrews et. al. 1971). Also, complex transformations will reduce the flexibility and interpretability of the original data (Andrews et. al.1971).

For most types of biological data, the extent of violation of the assumptions may not seriously invalidate the statistical techniques, as most of them are considered sufficiently robust (Steel and Torrie 1960, Seal 1968).

1.5 Multivariate Analysis of Variance (MANOVA)

Fisher(1948) defined the analysis of variance (ANOVA) as "the separation of the variance ascribable to one group of cause, from the variance ascribable to the other groups".

In simple (one way classification) ANOVA, sums of squares are denoted by scalars. The total sums of squares can be separated into two components: the among-group and the within-group sums of squares. Each of these two partitions, divided by it degree of freedom, is a separate independent estimator of the within group variance when the null hypothesis holds, The ratio of these two estimators is the F-value, the probability of which is used to test the equality of the two estimates of the variance. If the two estimates are not equal, it implies that the among group component is non-zero (i.e. the null hypothesis is rejected). This implies further that the group means are not equal.

In simple MANOVA, sums of squares and cross-products are contained in a square matrix of p order (p= No. of variables). As with ANOVA, the "total sums of squares and cross-products" matrix (T-SSCP) can be partitioned into "among group SSCP" (A-SSCP) and "within group SSCP" (W-SSCP). Both W-SSCP and A-SSCP have independent characteristic distributions (Wishart distributions), with n-g and g-1 degrees of freedom, when the null hypothesis is true (Kshirsagar 1972) (When n= total no. of individuals, g= no. of groups). Since W-SSCP and A-SSCP are matrices, their ratio does not exist. The determinant of such a matrix is it's generalized variance (Anderson 1958, Press 1972, Searle 1966). Wilks(1932) therefore introduced the determinant ratio statistic, Lambda (λ), to test the variance-equality null-hypothesis for matrices. Wilk's Lambda plays the same role in MANOVA as F plays in ANOVA. It is defined as:

$$\lambda_{(n-1,\ p,\ g-1)} = \frac{|W\text{-}SSCP|}{|A\text{-}SSCP + W\text{-}SSCP|} = \frac{|W\text{-}SSCP|}{|T\text{-}SSCP|}$$

In order to use this criterion for testing the null hypothesis, it is necessary to know its distribution and its percentage points. However Lambda is a family of curves each with three parameter (n-1, p, g-1), the percentage points of which are tedious to tabulate (Cooley and Lohnes 1971). The percentage points of Lambda distribution and its percentage points have been tabulated for certain restricted values of its parameters (Schatzoff 1966, Pillai and Gupta 1969). Pearson and Hartley(1972) have improved the tables and listed the 5% and 1% points for P=3 to 10, g-1 = 2 to 22, and some selected values of n. However, general utility still depends on the transformation of λ to χ^2 of F approximations (Cooley and Lohnes 1971).

Bartlett introduced the χ^2 approximation for Lambda. He derived that ($-m \log_e \lambda$) is approximately distributed as a χ^2, with p(g-1) degree of freedom, where $m = n-1-\frac{1}{2}(p+g)$, is Bartlett's correction factor. This approximation is sufficiently accurate, only if n is comparatively large. In fact, it is accurate to three decimal places, if $p^2 + (g-1)^2 \leq (1/3)m$. (Kshirsagar 1972), i.e. approximately, $n > 3(p^2+g^2)$. Rao(1952) derived an F-approximation for Lambda, which is superior to the χ^2 method in that it gives a more accurate approximation even for very small degrees of freedom (Cooley and Lohnes 1971). Rao's F approximation is:

$$F_{(n_1,\ n_2)} = \frac{n_2}{n_1}\ \frac{1-\lambda^{\frac{1}{s}}}{\lambda^{\frac{1}{s}}} \quad \text{Where } s = ((p^2\ (g-1)^2 - 4)/(p^2 + (g-1)^2-5))^{\frac{1}{2}}$$

$$n_1 = p(g-1)$$

$$n_2 = ms- \tfrac{1}{2}(p(g-1)-2)$$

m = Bartletts' correction factor.

When either p or (g-1) is less than 3, a special F-ratio is used (Kshirsagar 1972, Cooley and Lohnes 1971) (See Appendix A-2). This special F-ratio will reduce to univariate ANOVA F in cases of g groups and one variate (i.e. p=1), and to student's t in cases of two groups (g=2) and one variate (Cooley and Lohnes 1971).

1.6 Multiple Discriminant Analysis

Discriminant analysis estimates a set of linear coefficients vector(V) which will transform the original data vector (x) to a new vector (Y), such that the differences between the new vectors are maximized. This is achieved by minimization of residuals orthogonal to the function Y (Cooley and Lohnes 1973). Linear multiple regression also obtains such a linear function, but its minimization residuals is with respect to Y (Draper and Smith 1966). There are three types of discriminant analysis.

When there are two univariate populations (Π_1, Π_2) with means μ_1 and μ_2, and a common variance σ^2. The standardized distance between the means is $(\mu_1 - \mu_2)/\sigma$, if $\mu_1 > \mu_2$. An observation x will be classified to belong to Π_1, if it is nearer to μ_1 and to Π_2 if it is nearer to μ_2. The risk of misclassification is smaller if $(\mu_1 - \mu_2)/\sigma$ is larger (i.e. the two populations are further apart). For this reason, Fisher suggested that in multvariate case, the p variates be combined linearly in such a way that the $(\mu_1 - \mu_2)/\sigma$ for the linear combination is a maximum. The classification rule is then based on this optimum linear combination. The Fisher's discriminant function is: $\underset{(1\times 2)}{Y} = \underset{(1\times p)}{V'} \; \underset{(p\times 2)}{X}$. Where V' is the vector of linear coefficient which will maximize the standardized distance between Y_{11} and Y_{12}. The original p-dimensional classification procedure is reduced to a one dimensional one. The discriminant function (Y) obtain from original grand centroid (i.e. the grand mean vector of these two populations) will be the boundary point for classification. The discriminating ability of the function is measured by Mahalanobis D^2. This is the original <u>discriminant analysis</u> (Anderson 1958, Kshirsagar 1972, Press 1972, Rao 1952).

Discriminant function had been used by Fisher (1936) to discriminate two *Iris* species; by Rao to allocate Highdown skull to Bronze Age or Iron Age Populations; by Salia and Flowers(1969) to allocate American lobster into one of the two groups; and by Mather and Philip (after Mather 1949) to discriminate Barley into two groups. Mather and Philip have used the discriminant function as a "super character", which is undefinable in ordinary scales and is impossible to measure directly. Their "super-character" was concerned with ear-conformation in Barley (Mather 1949, Mather and Jinks 1971). Some other examples had been cited by Kshirsagar(1972).

When there are $g(g > 2)$ groups, the problem becomes more-complex and "generalized discriminant analysis" is used. The main purpose is to find g discriminant functions which will serve as indices for allocating a new individual into one of the g groups. For each group the generalized discriminant function is defined as:

$$\underset{(1\mathrm{x}1)}{Y} = \underset{(1\mathrm{x}p)}{V'} \underset{(p\mathrm{x}1)}{X} + \underset{(1\mathrm{x}1)}{K}$$

where $$\underset{(p\mathrm{x}1)}{V} = \underset{(p\mathrm{x}p)}{(W\text{-}MSCP_X)^{-1}} \underset{(p\mathrm{x}1)}{\overline{X}} ,$$

$$\underset{(1\mathrm{x}1)}{K} = -\tfrac{1}{2} \underset{(1\mathrm{x}p)}{\overline{X}'} \underset{(p\mathrm{x}p)}{(W\text{-}MSCP_X)^{-1}} \underset{(p\mathrm{x}1)}{\overline{X}} .$$

An observation is fitted into all the g discriminant functions and each associated probability is obtained. The observation is classified into the group for which it has the highest associate probability (Anderson 1958, Anonymous 1975, Anonymous 1968). Some authors, such as Anonymous(1975), Gould and Johnstone(1972), and Anonymous(1968) refer to this method as "multiple discriminant analysis". However, to avoid confusion with the next method, the term "generalized discriminant analysis" is preferred.

Rao(1952) used this method to allocate army recruits into different neurotic groups. It's been widely used in the study of geographic variation (Gould and Johnstone(1972) have cited more than 20 examples).

Multiple discriminant analysis is an ordination method, and is different from the last two allocation methods. Ordination is a group of techniques which is used to reduce the original P-dimensional space to a new q-dimensional space, with minimum loss of information. In this usage, q is the rank of the model, where $q \leq g-1$ (if $g-1 < p$), or $q \leq p$ (if $p \leq g-1$).

The purpose of multiple discriminant analysis is to find a set of coefficient vectors (V), the application of which to the original data vector maximized the observed differences amongst the groups (Clifford and Stephenson 1975).

The multiple discriminant functions are defined as:

$$\underset{(q\mathrm{x}g)}{Y} = \underset{(q\mathrm{x}p)}{V'} \underset{(p\mathrm{x}g)}{X} ,$$

such that the ratio of $(A\text{-}SSCP_Y)$ to $(W\text{-}SSCP_Y)$ is maximized, subject to the orthogonal constraint that $V'V = I_{(q)}$. This constraint is necessary, for otherwise the ratio can be indefinitely maximized.

It is similar to canonical variate analysis in that the canonical variate is defined in the same way. However, in this case it is ($A-MSCP_Y$) which is maximized, subject to the new constraint that $W\text{-}MSCP_Y = V' (W\text{-}MSCP_X)V = I_{(q)}$.

The q discriminant functions are obtained sequentially according to their discriminating ability. This is measured by their corresponding eigenvalues, or roots. The first discriminant function provides the maximum "separation" of the group means. The second discriminant function provides the second largest"separation" of the group means in an orthogonal direction to the first one, and so on. By testing the discriminating ability of successive functions, one can retain only the first, most significant functions, and ignore the rest without sacrificing too much information (Kshirsagar 1972). (The detailed procedure will be discussed in Chapter 2). However, the insignificant functions must not be disregarded unreservedly, as they may reveal some small, subtle and highly interesting variation (Gould and Johnston 1972). There are some cases of misuse in this sense. e.g. without any test, Glenday and Fejer(1956) regarded the first multiple discriminant function as the only useful function and ignored the others. They further used the coefficients of the first function in a selection index in the selection of Lolium Species.

In some cases of ordination, the first few axes may be adequate to explain most of the variation in the original data. Examples of this follow. (a) Project Talent (a project that studied the relationships among the abilities, interest and other characteristics of American Youth) of Cooley and Lohnes(1971). The first function explained 99.83% of the variation in the three original variates. (b) Soy bean study of Mungomery et. al.(1974), in which the first two axes explained 78.3% and 82.6% of the original variation of seed yield and protein content, respectively. However, in more complex cases, more axes may be required For example: (a) Lavarack's(1972) taxonomic study of orchids, (b) Noymeir's(1970, 1971) study of semi-arid Australian vegetation. In each of these cases more than ten axes were needed.

Multiple discriminant analysis, and canonical variate analysis, can be used as descriptive and exploratory tools. The former summarizes the complex relationships amongst variables, and provides a useful method of reducing the dimensionality of a problem by considering only the first few important axes. In short, it is a systematic technique for analysing an interacting complex system (Kshirsagar 1972). However, the whole system is by no means simplified, for while the space has been reduced, the complexity of the axes has been increased. The whole system has been changed from a complex space defined by many simple variables to a simple space defined by several complex variables (i.e. a "super character", as defined by Mather and Phillip) (Clifford and Stephenson 1975). These complex variables are "artificial" combinations of the original variables, and have no meaning in the original scales (Kshirsagar 1972).

1.7 Clustering Analysis

The terms "Cluster Analysis" and "Classification" have been used loosely to refer to a wide variety of fundamentally different numerical techniques (Cormack 1971, Williams 1971, Williams and Clifford 1971, Lance and Williams 1967a, Anderberg 1973). In this study, "Classification" will be used in a very broad and general sense to mean allocation of individuals into groups, so that individuals within groups are in some sense more similar to one another than to individuals in other groups. This includes both "pattern recognition" and "pattern extraction". Pattern recognition (Williams 1972), "identification" or "assignment" (Cormack 1971), aims at identifying given individuals and fitting them into *a priori* defined patterns. This includes methods of simplification or ordination, such as principal component analysis, generalized discriminant analysis, and multiple discriminant analysis (Refer to section 1.3, 1.6 and Appendix A-2). Pattern extraction, "pattern analysis" (Williams 1972), or "cluster analysis", sorts a given set of individuals into meaningful patterns undefined *a priori*.

Cluster analysis implies a numerical model, plus a strategy (or algorithm) whereby the model is implemented (Williams 1971, Cormack 1971). They are interdependent. The numerical model will translate the concept of "similarity" into some measure, which the strategy will work upon. An example is the increment sum of squares strategy which utilises only the Squared Euclidean Distance measure of similarity.

The basic data for clustering normally comprises a set of individuals (elements, entities or OTU---Operational Taxonomic Units) described by a set of attributes (characters). Attributes are any form of numerically coded descriptions (Lance and Williams 1967). There are four main types of attributes (or measurement scales), graded from "weakest" to "strongest" with respect to information content (Conover 1971). They are binary, disordered multistate, ordered multistate, and continuous (Clifford and Stephenson 1975). Different terminology has been used by other authors. (Refer to Appendix A-3).

1.7.1. Similarity Measures

A wide range of numerical definitions for interindividual "similarity" or "dissimilarity" have been devised. The extensive reviews of these measures by Goodman and Kruskal, Dagneli, Sokal and Sneath (all cited in Williams 1971), Cormack(1971), Anderberg(1973), and Clifford and Stephenson(1975) may be consulted, but any single one of them is far from complete. Of all these measures, relatively few are in current use. Most of the neglected measures either are variants of others and have some undesirable properties, or are highly specialized for certain types of data only. The important properties of similarity measures have been discussed by some authors. For example, Boyce(1969) and Williams(1971) have discussed the choice between similarity or dissimilarity measures. (A similarity measure has similar properties to a probability or a correlation coefficient, its maximum positive value represents identity, and differences cause reduction in the measure. A dissimilarity measure has similar properties to a linear distance, it is zero for identity and increases positively for increased extent of difference). These authors also considered choices between size and shape measures, whether to have double zero matches or not, and the properties of spatial or probabilistic models. With size measures, such as Euclidean Distances, two individuals are identical if corresponding attributes have equal absolute value. With shape measures, such as correlation coefficient, they are identical if attributes occur in equal proportion. Lance and William(1976a), and Clifford and Stephenson(1975) have discussed the metric and additivity nature of the measures. There are four fundamental requirements for a metric measure: (i) symmetry, (ii) triangular inequality, (iii) distinguishability of non-identicals, and (iv) indistinguishability of identicals (Williams and Dales 1965). These requirements clearly indicate the geometrical advantages of a metric measure. Additivity of measures will be important only when the attributes are dimensionless (Lance and Williams 1967).

Three classes of measures seem to dominate, and are being used widely. All of them are dissimilarity and size measures, and have been proved applicable to mixed types of attributes (Williams 1971, 1972). They are as follow:-

(1) Measures based on first Minkowski metric (L_1)

$$L_{1ij} = \sum_{k=1}^{p} W_k \left| X_{ik} - X_{jk} \right| \quad \text{(Lance and Williams 1967b).}$$

When $W_k=1$, it is the "City Block" or "Manhattan" metric (Lance and Williams 1967b, Cormack 1971).

When $W_k = \left| \dfrac{1}{(x_{ik} + x_{jk})} \right|$, it is the "Canberra" metric (Lance and Williams 1967b, Cormack 1971).

Here, x_{ik} and x_{jk} denote the value taken by two individuals or clusters (i) and (j), for the k th of p attributes; and L_{1ij} denotes the measure of dissimilarity between individuals of clusters (i) and (j). These measures are metric and additive over attributes.

(2) Measures based on the second Minkowski metric (L_2), or Euclidean Distance (ED).

$$L_{2ij} = \left(\sum_{k=1}^{p} W_k \; (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} \quad \text{(Lance and Williams 1967b).}$$

The square of ED is known as Squared Euclidean Distance (SED). W_k is the standardizing factor: $W_k=1$ for unstandardized ED, $W_k = 1/\sigma_k^2$ for standardized by standard error (Gower 1966), $W_k = 1/\max\,(x_{ik} - x_{jk})^2$ for standardized by range (Cormack 1971), $W_k = \sum_{h \neq k}^{p} X^2_{hk}$ for standardized by measure of importance of the attribute (Williams and Dale 1964). The "importance" is measured in the following way. X^2_{hk} is calculated between every pair of attributes h and k, and the sum of all the X^2 which involve a particular attribute h (i.e. $\sum_{h \neq k} X^2_{hk}$) is the measure of importance of that particular attribute h.

SED is not a metric, it is additive over attributes, and it possesses combinatorial properties. ED is a metric (when there is no missing data), but it is not additive over attributes (Clifford and Stephenson 1975, Lance and Williams 1967a, 1967b).

SED has an important (and sometimes undesirable) property of giving extra weight to outlying values, so that a single large difference will dominate over several small differences. A prior square root or log transformation of attributes will correct this (Clifford and Stephenson 1975).

When attributes are measured in different scales, ED and SED have nondefinable physical dimension. To avoid this,standardizing factors, such as standard error, mean, cube root of the sample third moment (Gower 1966), the range (Cormack 1971) and the measure of importance (i.e. ΣX_{hk}^2 mentioned previously) (Williams and Dale 1964) may be used.

Sokal(1961) pointed out that, if the attributes are correlated then $\Sigma(x_{ik} - x_{jk})^2$ is not SED. In this case, Maholanobis D^2 or "restricted" SED is used. Restricted SED is evaluated from the least correlated attributes only (Corkmack 1971). However, this method is being criticized on two bases. Firstly, if the correlation matrices vary from group to group, the pooled matrix is inappropriate. Secondly, much of the correlation present may be an intrinsic property of the true clusters which are being sought. This correlation must be retained (Cormack 1971). Gower(1966) proposed that principal components should replace correlated attributes for evaluating SED to overcome these difficulties. This is practicable, provided the attribute set does not contain "too many" ordered multistate attributes or "too many" missing or inapplicable entities (Williams 1972).

(3) Information or Diversity Measures.

The taxonomist prefers the term "information", whereas the ecologist prefers the term "diversity". There are three major types. The detailed derivation and explanation of these are outlined in Clifford and Stephenson(1975).

(a) Shannon diversity index.

$$H = N \log N - \sum_{k=1}^{p} N_k \log N_k;$$

(b) Brillouin diversity index.

$$H_{(B)} = \text{long } (N!) - \sum_{k=1}^{p} \log (N_k!);$$

(c) Shannon Information Gain.

$$SIG = P(N \log N) - \sum_{k=1}^{p} (a_k \log a_k + (N-a_k) \log (N-a_k));$$

When N = No. of individuals involved,

P = No. of attributes,

N_k = No. of individuals in k th attribute,

a_k = No. of individuals at state 0 in k th attribute.

When dealing with a sample from the total population, H is prefered. H will be maximum when all N_k are equal. When dealing with the total population, $H_{(B)}$ is more appropriate. However, the ratio of H and $H_{(B)}$ is almost constant over a large range of N_k, so they are not very different.

Both H and $H_{(B)}$ are measures of a particular group. They may be used to measure similarity between members of a pair of groups ("information gain", ΔI), ΔI = H of group (1 + 2) - H of group 1 - H of group 2. Whereas, SIG is in itself a ΔI measure.

Apart from special interest of users, the choice of inter-individual measure will largely depend on the nature of data. For highly skewed binary data, such as the presence-and-absence records of species in plant ecology, an information or diversity index is preferred. For data defined by a small number of continuous attributes, with no strong outliers the Euclidean Distance is preferred. For positive data with few zeroes, but with occasional extreme outliers, (which should not dominate), the Canberra metric is indicated (Williams 1971, 1972). When data have no striking pecularities, the choice of clustering strategy is much more important than the choice of similarity measure (Williams 1971).

1.7.2. Clustering Procedures

The major decisions in selecting a clustering procedure can be represented as a series of dichotomous choices (Appendix A-4) (Williams 1971). These are considered in more detail in the following.

An exclusive clustering procedure is one in which a given individual occurs in one cluster and one cluster only; the population is divided into a set of mutually exclusive clusters, which nowhere overlap in their membership. This type of clustering is usually seen in the Taxonomy of living organisms. Conversely, a non-exclusive clustering procedure is one in which any given individual may appear simultaneously in more than one cluster (e.g. disease classification, medical diagnosis, and forest survey) (Williams 1971).

In intrinsic clustering, all attributes used are regarded as equivalent. There are two types: (a) intrinsically intrinsic, and (b) extrinsically intrinsic. In the first type, resultant clusters are of interest in their own right, as in pure taxonomy. In the second type, the boundaries between clusters are examined to find out if they reflect discontinuities in some external attributes (e.g. environmental attributes, which will affect the individual, such as altitude, temperature, and soil fertility). The nature of the external discontinuities is unknown in advance (e.g. in land survey problems) (Williams 1971). On the other hand, with extrinsic clustering, the external attribute is known in advance, together with the internal attributes. The resultant clusters, though based on the internal attributes (i.e. the attributes measured from the individual itself, such as height, size, and weight) are required to reflect discontinuities in the external attribute as closely as possible. Reallocation based on the external attribute may be required (Williams and Lance 1968, Williams 1971). It should be noted that extrinsic clustering is different from clustering based solely on external attributes, followed by examination of internal attribute of the resultant pattern of clusters. In such a clustering (which is based initially on external attributes), the resulting configuration of internal attributes may lack any pattern of their own, so it has no predictive value (Williams 1971). If there are more than one external attributes, canonical correlation analysis is recommended (Williams and Lance 1968).

A hierarchical clustering always optimizes a "route" between the entire population and the set of individuals of which it is composed.

The best "route" may be obtained at the expenses of having a slight reduction in homogeneity of individual clusters (Lance and Williams 1967a). Conversely, a nonhierarchical clustering always optimizes the structure of the individual clusters themselves, which are made as homogeneous as possible. The infrastructure of such a cluster cannot be examined, because no route is defined either between cluster and constituent individuals, or between a cluster and the complete population. When homogeneity of clusters is of prime importance, non-hierarchical clustering is preferred (Williams 1971). Hierarchical clustering has higher organizing ability. It's more traditional; and it parallels evolutionary theory approaches. It is much preferred by the taxonomist. Other advantages it has are: Flexibility in the final number of clusters formed, and availability of highly developed computer programmes (Clifford and Stephenson 1975).

The basic principle of serial optimization is: a cluster is defined and removed from the total population, and further clusters are serially formed by sequential definition and removal. This process continues until all the population is accounted for. Final reallocation may be needed to end the process. However, the general methodological principles of such straegies are being criticized on numerical and computational grounds (Williams 1971). In simultaneous optimization, the total population is partitioned, and the clusters are simultaneously optimized by an iterative process. It is strictly based on a Eucliden model. However, this method is believed to lack power, in that it produces types of cluster, not usually wanted (Williams 1971).

An agglomerative strategy is one that proceeds by progressive fusion, beginning with the individuals and ending with one complete population. Conversely a divisive strategy progressively splits the population into smaller and smaller clusters, begining with the complete population and ending with the collection of individuals (Williams 1971, Anderberg 1973). A polythetic system is one based on a measure of similarity or dissimilarity, applied over all attributes simultaneously. This results in an individual being grouped with those individuals which, on the average, it most resembles. A monothetic system is one based on a single attribute at a time. The first "division" attribute must be optimized in some sense, dividing the population into two parts as unlike as possible. The selection of the attribute depends on the properties of population. All agglomerative strategies are polythetic, and most commonly used divisive strategies are monothetic (Williams 1971).

As agglomerative strategies begin at the individual level, they suffer from two disadvantages. Firstly, there is comparatively longer computation time. Secondly, theoretically they are prone to comparatively greater amounts of misclassification, because of the greater possibility of error at the interindividual level. Monothetic divisive strategies produce relatively stable clustering structure when new individuals are added, provided the priority in the attributes remains the same (Clifford and Stephenon 1975). Monothetic cluster definitions are simple and clear. However, monothetic clustering tends to produce an unduly large number of fragmentary clusters at later stages (Williams 1971).

1.7.3. Hierachical Clustering Strategies

There are three main types of hierachical clustering strategies, namely polythetic divisive, polythetic agglomerative and monothetic divisive (Williams 1971). Strategies are based on different algebraic models, so that each of them will exhibit different properties (Clustering behaviours). These properties also depend on the similarity measures used. The understanding of these properties will help in deciding on which strategy to use (Clifford and Stephenson 1975). The main properties are discussed in the following.

(a) Combinatorial or Noncombinatorial

a combinatorial strategy is one in which the original interindividual similarity measures can be discarded immediately a cluster is formed. The similarity measure of this newly formed cluster is sufficient for later processes. In noncombinatorial strategies, the original interindividual similarity measures must be retained for later calculations, even though the individuals are already in a cluster. The combinatorial strategy has conceptual and computational advantages (Lance and Williams 1967a).

(b) Compatible and incompatible

A compatible strategy is one in which measure calculated later in the process are of exactly the same kind as the initial measures; they have the same dimension (if any), are subject to same constraints, and can be illustrated by an exactly comparable model. Whereas in an incompatible strategy, at least some of the properties of initial interindividual measures are lost later. This causes difficulties in interpretation (Lance and Williams 1967a).

(c) Space-conserving or Space-distorting

The original interindividual measures are regarded as occuring in a given space with known properties. If the properties of this space remain unaltered when clusters form, the strategy is "space conserving". If the opposite occurs, the strategy is "space-distorting" (Lance and Williams 1967a). If a cluster, on formation, appears to move nearer to some or all the remaining entities, the method is "space-contracting". The chance that a remaing "unclustered" individual will add to a preexisting cluster rather than act as the nucleus of a new cluster is thereby increased and the system is said to "chain" (Williams 1971). If clusters appear to recede from other entities, on formation and growth, the method is "space-dilating". Individuals not yet in cluster are now more likely to form "non-conformist" clusters, in which members are quite dissimilar. This tendency is cluster size dependent: the larger the existing cluster, the greater the tendency to form a second cluster. The tendency of cluster size dependency may be "asymptotic", so that once the cluster has attained a modest size, further accretion makes little difference; or it may be "indefinite" so that every accretion makes the cluster substantially more remote and therefore more difficult to join (Williams 1971).

(d) Monotonic and Non-monotonic

A monotonic strategy is one which will not cause reversal in the dendrograms. As clustering proceeds and clusters grow, the similarity measure is non-decreasing. However, in non-monotonic strategies, the similarity measure of the new cluster may be less than that of the two before they are merged. The conceptual illogic of reversals makes non-monotonic strategies, obsolete. They should be avoided (Clifford and Stephenson 1975).

1.7.3.1 Hirearchical Divisive

Although the polythetic divisive method is comparatively promising, the method has seldom been used, because the development of the computer programmes is still rather primitive (Williams 1971). There are two common approaches. One subdivides the initial population on the basis of a single attribute and then reallocates apparently misclassified entities on the basis of a maximum likelihood procedure. The other undertakes a principal component analysis, and then subdivides the initial population on the basis of the principal component scores on successive axes (Clifford and Stephenson 1975).

In the monothetic divisive method, the main feature is the careful choice of first and successive attributes on which the entities are divided. Two main methods are available for the determination of division attribute---those depending on information theory and those depending on χ^2.

The χ^2 method is often referred to as "Association Analysis" (Williams and Lambert 1959, 1960, 1961). It looks for an attribute which will divide the entities into two most-dissimilar clusters. χ^2 are calculated for every pair of attributes over all entities. These are then summed over all attributes and that with the largest $\sum_{h=k}^{P} \chi^2_{hk}$ is used as the basis for dividing the set of entities into two subsets. Each subset is futher subdivided in the same manner as the first one, until the required number of subsets is obtained. (Note: $\sum_{h\neq k}^{P} \chi^2_{hk}$ is the measure of importance used for standardized ED is section 1.7.1) (Clifford and Stephenson 1975). A computer programe is available for this method (Lance and Williams 1968b).

The information theory method looks for an attribute which will divide the entities into two clusters that have the greatest internal homogeneity. This is based on information measure. The attributes which give maximal "information gain", ΔI (Refer to section 1.7.1), is used as the basis for dividing the set of entities into two subsets (Clifford and Stephenson 1975). Computer programmes are available (Lance and Williams 1971).

1.7.3.2 Agglomerative Polythetic Strategies

There are two main groups of such strategies, discussed in the following.

1.7.3.2.1 Strategies Based on Successive Information Gain

These strategies are based on minimum information gain (ΔI) at each fusion, or on a minimum value of the decision function $2\sqrt{\Delta I} - \sqrt{2n+1}$ at each fusion (Clifford and Stephenson 1975). They are used when most of the attributes are binary or disordered multistate. Programmes are available for completely binary attributes (William et. al.1966), and for mixed attributes (with or without missing data) (Lance and Williams 1967b). Edye et. al.(1970) and Burt et. al.(1971) have used these programmes for clustering the legumes.

They are space-dilating and the cluster-size dependence is indefinite for both individual/cluster merges and cluster/cluster merges. They are non-combinatorial (Williams 1971). Their advantage is strongest when dealing with binary attributes and is weakest when dealing with continuous or metrical attributes (Clifford and Stephenson 1975).

1.7.3.2.2 Strategies Not Based on Successive Information Gain

These are the most widely used and most studied strategies (Burr 1968, 1970, Cormack 1971, Lance and Williams 1967b, Williams 1971, 1972, Clifford and Stephenson 1975).

There are nine main strategies. In all of them fusion begins with the most similar pair of individuals, as established by whatever similarity measure is employed. The differences amongst strategies appear in subsequent fusions.

(1) Single Linkage or Nearest Neighbour

This is the simplest and oldest strategy. Two clusters are fused if the similarity between their closest elements, one in each cluster is maximum. It's combinatorial, compatible and monotonic. The space-contracting and consequential chaining tendencies are notorious (Lance and Williams 1967). Due to this, Pritchard and Anderson(1971) decribe it as "least useful" for ecological study. From a practical point of view, this strategy should be regarded as obsolete. However, its simplicity and comparative stability are preferred by mathematicians, thus ensuring its continual popularity (Williams 1972, Clifford and Stephenson 1975, Cormack 1971).

(2) Complete Linkage or Furthest Neighbour

This is the exact antithesis of single linkage, in that two clusters are fused if the similarity is maximum between the most remote pair of elements, one in each cluster. It's combinatorial, compatible, monotonic and markedly space dilating (Lance and Williams 1967). The intense clustering results in meaningless relationships amongst clusters. Only relationships within individual clusters are interpretable (Anderberg 1973). This strategy is particularly suited for weakly structured data, and so Pritchard and Anderson(1971) regard it as one of the most useful for ecology.

(3) Centroid

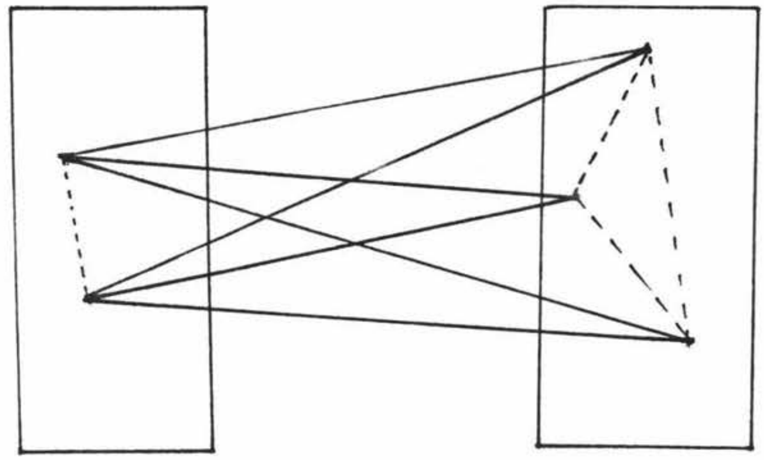
In this strategy, every cluster is regarded as a single point at its centroid in multidimensional space. Clusters with minimum distance between centroids (i.e. most similar) are fused. The new centroid is the weighted average (by cluster size) of the two original centroids. It's compatible, nonmonotonic and space-conserving (Lance and Williams 1967b). However, under a "simplex test" (a simulated clustering analysis with the similarity measures between all entities being made equal at the beginning of clustering), Burr(1970) found it produced chain reversal (i.e. nonmonotonic and space contracting). It is combinatorial, if the similarity measure is metric or SED. Otherwise, for correlation coefficient and other non-metric measures, it is non-combinatioral (Lance and Williams 1967), and difficult to interpret (Anderberg 1973). Due to the weighted average, small clusters tend to lose their identity by being merged with large clusters.

(4) Median

The median method was proposed to retain the identity of smaller clusters, as discussed in the previous method (Gower 1966). The general idea is that, after fusion, centroids are weighted equally regardless of their sizes. The new centroid will lie at the midpoint between the two "old" centroids. It is space conserving, nonmonotonic and combinatorial. It's compatible for SED and non-metric measures, but non-compatible for correlation coefficients (Lance and Williams 1967).

(5) Group Average Between Merged Clusters
(often referred to as "Group Average")

Two clusters are fused if the average interindividual similarity of every possible pair of entities is maximum (where each pair comprises



Cluster I   Cluster J

FIGURE 1.1 The difference between "Group Average Between Merged Clusters" method and "Group Average Within New Cluster" method.

— = relationships examined in the first method, (see text).

-- = additional relationships examined in the second method, (see text).

of one entity from each cluster). See Figure 1.1. Assuming cluster I contains $N_I = 2$ entities, and cluster J contains $N_J = 3$ entities. The group average is defined as the average of the $N_I N_J$ (=6) possible similarity measures amongst these entities. That is, the average of 6 relations expressed by solid lines in Figure 1.1. Only a similarity measure which has a meaningful average can be used. Some, such as the correlation coefficient, should not be used (unless correlation are transformed to covariance first) (Cormack 1971). The method is combinatorial, monotonic, and compatible. Since it has no marked tendencies to space distortion, it may be regarded as a space-conserving strategy (Lance and Williams 1967). Pritchard and Anderson(1971) considered it as one of the more useful strategies.

(6) Group Average Within New Cluster

Two clusters are fused if the average interindividual similarity of every possible pair of entities within the cluster to be formed, is maximal (Anderberg 1973). In this case, the group average is defined as the average of the $\frac{1}{2}(N_I + N_J)(N_I + N_J - 1)$ similarity measures. That is,the average of the 10 relations shown by all the (solid and dotted) lines in Figure 1.1. It's properties are believed to lie between those of the Single Linkage and Complete Linkage. However, detail is not known, and, in practice, it usually gives similar results to Complete Linkage (Anderberg 1973).

(7) Minimum Increment Sums of Squares (Ward's Method)

Two clusters are fused if the fusion results in a minimal increase of the pooled within-cluster sum of squares. This strategy was first described by Ward(1963), and later by Burr(1968, 1970). Both of them recommended the minimal increase, rather than the minimal sum, of within-cluster sums of squares as the criterion of merging. The latter gave absurd results (Ward 1963, Burr 1970). Since the total sum of squares is constant, this strategy is equivalent to maximizing the among-cluster sums of squares.

If measures other than SED are used, the properties of this method are not known (Anderberg 1973, Clifford and Stephenson 1975). Anderberg(1973) has shown that the within-cluster sum of squares is equal to half the sum of SED. The method is combinatorial (Anderberg 1973, Wishart 1969), Monotonic(Burr 1970), non-compatible and space-dilating (Clifford and Stephenson 1975). Cluster size dependence is asymptotic for indidual/cluster merges, and indefinite for cluster/cluster merges. A new individual will tend to join the smallest available cluster (Williams 1971). Pritchard and Anderson(1971) felt that this strategy was useful for ecology.

(8) Minimum Variance

Two clusters are fused if the fusion results in a minimal increase of pooled within-cluster variance. The strategy was introduced by Burr(1970) and it's properties are believed to be similar to those of Ward's Method, although the detail is not fully known. For SED, it is combinational (Williams 1971), and monotonic (Burr 1970).

(9) Flexible

Lance and Williams(1966, 1967) suggested that the Nearest Neighbour, Furthest Neighbour, Group Average Between Merged Clusters, Centroid, and Median strategies are special cases of a general system, at least when the similarity measure is SED. Ward's Method was included in this system later (Wishart 1969, Burr 1970, Cormack 1971). They defined a system with three clusters (h), (i) and (j), containing n_h, n_i and n_j individuals, respectively, and with inter-cluster distances defined as d_{hi}, d_{hj} and d_{ij}. Also, it was assumed that d_{ij} was smallest, so that i and j fused to form a new cluster K, with n_k (= $n_i + n_j$) individuals. Their general linear model was:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma \left|(d_{hi} - d_{hj})\right| ,$$

where the parameters α_i, α_j, β and γ determine the nautre of the clustering strategy, and are defined in the following;

	α_i	α_j	β	γ
Nearest Neighbour	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Furtherst Neighbour	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Group Average Between Merged Clusters	$\frac{n_i}{n_k}$	$\frac{n_j}{n_k}$	0	0
Centroid	$\frac{n_i}{n_k}$	$\frac{n_j}{n_k}$	$-\alpha_i \alpha_j$	0

Ward's Method	$\frac{n_i + n_h}{n_k + n_h}$	$\frac{n_j + n_h}{n_k + n_h}$	$\frac{-n_h}{n_k + n_h}$	0
Flexible ($x = \beta$)	$\frac{1}{2}(1-x)$	$\frac{1}{2}(1-x)$	$x(<1)$	0

Where $\gamma = o$, the strategies will be monotonic if $(\alpha_i + \alpha_j + \beta) \geq 1$ (Lance and Williams 1966, 1967, Williams 1971). Only Median and Centroid strategies fail in this requirement.

Applying the quadruple constraint ($\alpha_i + \alpha_j + \beta = 1$, $\alpha_i = \alpha_j$, $\beta = x < 1$, $\gamma = 0$) to the linear model, a monotonic "Flexible" strategy is derived (Lance and Williams 1966). This strategy is then completely defined by β (or X), the clustering intersity coefficient. As β decreases from < 1 to a negative value, the clustering intensity will increase from weak to intense, and its space distorting properties change from space-contracting to space-dilating.

These strategies are combinatorial and compatible for SED. They are meaningless and non-compatible for correlation coefficients (Lance and Williams 1967). With negative β, cluster-size-dependence is asymtoptic for both individual/cluster and cluster/cluster merges.

1.7.3.2.3 Number of Clusters

A practical problem in performing a cluster analysis is deciding on the number of clusters to obtain. The result of hierarchical clustering can be represented in a dendrogram (tree diagram). The number of clusters which may be obtained from the dendragram varies from one to the number of entities, depending on the level at which the hierarchy is "cut-off" (Anderberg 1973). The choice of "cut-off" point involves subjective decisions (Clifford and Stephenson 1975). Some authors seek the largest proportionate changes in the clustering criterion (inter-cluster distance) at successive stages of clustering (Pritchard and Anderson 1971). Most seem to use a subjective "optimal" number of clusters (Anderberg 1973). Burt et. al.(1971) arbitrarily chose 20 clusters as their "optimal" for 154 stylosanthes plants. Mungomery et. al.(1974) arbitrarily chose 10 clusters for their 58 lines of soy bean. These arbitrarily choices seem to be too subjective, especially, if the nature of the original population (i.e. population to be clustered) is unknown. A more objective, probabilistic decision method will be discussed later.

CHAPTER 2. MATERIALS AND METHODS

2.1 Field Design

This study uses a working collection of a Yorkshire Fog gene pool aggregated by Dr. R. J. Clements (Unpublised). The geographical location, altitude and habitat of the 201 accessions are given in Appendix B-1.

The working collection consists of seed progengy of 160 accessions, each seed lot having been open-pollinated at its original site. Each accession (seed population, group, entity) was replicated in three randomized complete blocks. Experimental units consisted of a single row of eight seedling plants. Plant spacing was 60cm in both directions.

2.2 Field Measurements

In summer 1975, 11 characters were scored on every plant of the working collection. These scores provided semi-quantitative defined scales of measurement. This system had also been used previously in 1974 (Gordon, unpublised). All characters, except flowering head colour and flowering date, were scored at about two to three weeks prior to elongation for flowering. The characters and their scoring systems are as follows:

(1) Clump Diameter (C.DIA)

Scores ranged from ½ to 5 with increment of ½. Each unit represented 15cm of length across the average diameter of the clump. Those exceeding 60cm were scored as 5. For computation, scores were doubled (scale: 1-10).

(2) Clump Density (C.DEN)

Scores ranged from ½ to 5 with increment of ½. Each unit represented a green leaf coverage of approximately 1/8 (12.5%) of the total ground area covered by the Clump. A score of 1 indicated that more than ½ (50%) was non-green (included bare ground and dead material); and a score of 5 indicated that all ground area of the clump was covered by green tissue. For computation, scores were doubled (scale: 1-10).

(3) Clump Erectness (C.ERE)

Scores ranged from ½ to 5 with increment of ½. Each unit represented approximately 18^{0} of average angular elevation between the ground level (horizontal) and the tiller axes. Vertical tillers (90^{0}) were scored as 5. For computation, scores were doubled (scale: 1-10).

(4) Clump Height (C.HEI)

Scores ranged from ½ to 5 with increment of ½. Each unit represented an average height of 10cm from ground level. For computation, scores were doubled (scale: 1-10).

(5) Rust (RUST)

Scores ranged from 0 to 5, with increment of 1 unit. A score of 0 represented no rust, 1 represented $<$ 10% of leaf area with rust lesions, 2 represented 10-25%, 3 represented 26-50%, 4 represented 51-70% and 5 represented $>$ 70%. This scale (0 to 5) was transformed and centralized to a complementary scale of 6 to 1 by function TRANSF (see section 2-3) during MANOVA. In this form it represented putative resistance to rust.

(6) Overal Disease (O.DIS)

The system was the same as used for rust, but considered all leaf disease lesions present. The same transformation and centralization were performed also.

(7) Leaf Roll (L.ROL)

Scores ranged from 0 to 2, with increment of 1 unit. 0 represented flat leaf, 1 represented partially rolled (longitudinally) leaf and 2 represented very rolled leaf. This scale was centralized (One was added to all scores), to remove zeroes (function TRANSF) during MANOVA.

(8) Leaf Tip Colour (L.COL)

Scores ranged from 0 to 3, with increment of 1 unit. 0 represented green, 1 represented slight red, 2 represented red to light purple and 3 represented dark purple. This scale was centralized (one was added to all scores), to remove zeroes. The colour of leaf tip was believed to reflect the pigment content of the leaf tip (Gordon, unpublised). It was assumed that the deeper the colour, the greater was the concentration of the pigments. It was also tentatively assumed from the nature of the colour, that the pigments may have been flavonoid.

(9) Leaf Width (L.WID)

Scores ranged from 1 to 5, with increment of 1 unit. A score of 1 represented an average of 5mm across the widest part of leaf balde, and each additional unit represented an increase of 3mm. Those exceeding 14mm were scored as 5.

(10) Inflorescence Colour (F. COL)

Scores ranged from 1 to 5, with increment of 1 unit. They were scored immediately after the infloresence emerged. A score of 1 represented pale white, 2 represented green, 3 represented purplish green, 4 represented light purple, and 5 represented dark purple. Increase in score reflected putative increase in pigments, particularly those of the purple colour.

(11) Flowering Date (F.DAT)

Scores ranged from 1 to 9 with increment of 1 unit. They were scored, once, when the infloresence emerged. Each unit represented one week, starting from 11th November, 1975. Those flowering after 6th January, 1976 were scored as 9.

Under these scoring systems all characters were regarded as ordered multistate attributes (Clifford and Stephenson 1975). Also all these characters were intrinsic attributes (Williams 1971). Plants with any missing characters (attributes) were treated as missing plants.

Four sets of attributes are used in the computations of this study. They are: (1) all characters (ALLCHARA), which included all the eleven characters in the analysis: (2) Agronomic characters (AGROCHARA), which included eight of the characters namely, C.DIA, C.DEN, C.ERE, C.HEI, RUST, O.DIS, L.WID and F.DAT; (3) most discriminant characters (DISCCHARA), which included the five characters found from ALLCHARA to have greatest discriminating ability, namely, C.ERE, C.HEI, RUST, F.COL and F.DAT; and (4) Jacques' characters (JACQCHARA), which included four characters nominated by Jacques(1962) as being ecocline indicators, namely, C.ERE, RUST, L.WID and F.DAT. These four sets of attributes were used in separate complete analyses, in order to examine the phenotypic variation from these four points of view.

2.3 Multivariate Analysis of Variance (MANOVA) and Multiple Discriminant Analysis (DISCRIM)

A computer program MANDIS was used for this part of the analysis. The program was adapted and modified from MANOVA and DISCRIM of Cooley and Lohnes(1971). These modifications included the addition of: function TRANSF(Gordon unpublished) for transformation and manipulation of input data; functions PRBF and SIGNIF (Gordon unpublished) for obtaining the probabilities and significance symbols of χ^2 and F-ratio; subroutines SMPRIN and DMPRIN for printing of matrices and vectors; and addition of Bartlett's χ^2 test of homogeneity of W-MSCP (Seal 1968) in the main program. For greater compatibility with other programmes, other data handling options were added also (See listing of MANDIS in Appendix B-2).

2.3.1 MANOVA

The linear model of MANOVA is:

$x_{ik} = \mu + \alpha_k + \xi_{ik}$ (Cooley and Lohnes 1971), where x_{ik} = the observed values of i th individual in k th group, μ = grand centroid of the whole population, α_k = the deviation of centroid of k th group from μ , and ξ_{ik} = the deviation of i th individual from centroid of k th group.

MANOVA calculated the mean, standard deviation and coefficient of variation of each character, and the determinant of MSCP for each group. Homogeneity of pooled W-MSCP was tested by two approaches, namely, Bartlett's χ^2 test and F-test. This involved the following:

$$A_1 = \left(\sum_{k=1}^{g} \frac{1}{N_k - 1} - \frac{1}{N-g} \right) \frac{2p^2 + 3p-1}{6(g-1)(p+1)},$$

$$A_2 = \left(\sum_{k=1}^{g} \frac{1}{(N_k - 1)^2} - \frac{1}{(N-g)^2} \right) \frac{(p-1)\ (p+2)}{6(g-1)},$$

$$n_1 = \frac{(g-1)\ (P+1)P}{2},$$

$$n_2 = \left| \frac{n_1 + 2}{A_2 - A_1^2} \right|,$$

$$M = (N-g) \log_e D_w - \sum_{k=1}^{g} (N_k - 1) \log_e D_k,$$

where g = No. groups,

P = No. of characters,

D_k = Determinant of W-MSCP of k th group.

D_w = Determinant of pooled W-MSCP of all group,

N_k = No. of individual in k th group,

and N = total no. of individual over all groups.

Then $\chi^2_{n1} = (1-A_1)M$ (Derived from Seal (1968)).

If $A_2 - A_1^2$ was positive,

then $b = \frac{n_1}{1-A_1 - (n_1/n_2)}$, and

$F_{(n_1,\ n_2)} = \frac{M}{b}$ (Cooley and Lohnes 1971).

If $A_2 - A_1^2$ was negative,

then $b = \frac{n_2}{1-A_1 + (2/n_2)}$, and

$F_{(n_1,\ n_2)} = \frac{n_2}{n_1}\frac{M}{(b-M)}$ (Cooley and Lohnes 1971).

Next, Wilk's Lambda and it's complement (the square of multivariate generalization of Fisher's correlation ratio) were calculated. Rao's F approximation for Wilk's Lambda was used to test the equality of the two estimates of variance (i.e. significance of amongst-groups variance). The complement of Wilk's Lambda is often referred to as "MANOVA Eta-square" (Cooley and Lohnes 1971). It was a descriptive statistic that expresses the proportion of criterion variance (total generalized variance) explainable by the predictor variance (generalized variance due to grouping). It was similar to the square of the multiple correlation coefficient (the coefficient of multiple determination), which was defined as the ratio of sum of squares due to regression to the total sum of squares (Draper and Smith 1966). "Nonindependent" univariate F-tests were also carried out (Cooley and Lohnes 1971). The "Nonindependent" univariate F-ratio was the corresponding ratio of the diagonal element of A-MSCP (Among group mean squares of the character) to the respective diagonal element of W-MSCP (Within group mean squares of the same character). This should be interpreted only if the MANOVA null hypothesis has been rejected. When Wilk's Lambda test has produced a rejection, inspection of the "nonindependent" univariate F-ratio may suggest which of the characters was contributing most to the discrimination of the groups. Though the probability of this F-ratio was reported also it should be used only as a rough indicator and should not be used as an explicit inference of significance level. A descriptive statistic was more suitable under this circumstance. Thus, Eta-square for each character was reported also. Eta-square was defined as:

$$\eta^2 = \frac{\text{Among group sum of squares}}{\text{Total group sum of squares}} .$$

η^2 expressed the proportion of criterion variance (Total variance) explainable by the predictor variance (Among group variance). This was similar to the square of the coefficient of multiple correlation (the coefficient of multiple determination). Consequently, η was similar to the coefficient of multiple correlation also.

Though true independent univariate F-tests can be performed their computations are very complex and the resulting "Stepdown F-ratios" are hard to interpret (Cooley and Lohnes 1971). Therefore they have been omitted.

The T-SSCP, W-SSCP, group centroids and grand centroid were obtained for the Multiple Discriminant Analysis, which followed.

2.3.2 Multiple Discriminant Analysis

The model was: $\underset{(q\text{x}g)}{Y} = \underset{(q\text{x}p)}{V'} \; \underset{(p\text{x}g)}{X}$,

$$\text{Leading to } \phi = \left.\frac{(\text{A-SSCP}_y)}{(\text{W-SSCP}_y)}\right|_{\max}$$

$$= \left.\frac{V'(\text{A-SSCP}_x)\;V}{V'(\text{W-SSCP}_x)\;V}\right|_{\max},$$

subject to the constraint that $V'V = I_{(q)}$.

Here, x was the matrix of original scores, with g groups and p characters;

V' was the matrix of the coefficient of q discriminant functions;

Y was the matrix of the q discriminant scores for g groups;

ϕ was the vector of q eigenvalues for the q discriminant functions;

and $I_{(q)}$ was the identity matrix of order q.

It can be shown by differential calculus that ϕ and V are the eigenvalues and their eigenvectors of the equation:

$\left[(W\text{-}SSCP_x)^{-1} \ (A\text{-}SSCP_x)-\phi I_{(p)}\right] V = 0$ (Seal 1968, Cooley and Lohnes 1971). Subroutine NROOT (Cooley and Lohnes 1969) was used to solve this.

Wilk's Lambda criterion (λ) may also be computed as a function of the eigenvalue as follows:

$$\lambda = \prod_{j=1}^{q} \frac{1}{1+\phi_j}$$

Where ϕ_j = the j th element of the ϕ vector, or the j th eigenvalue.

This was a different approach for obtaining λ to the previous one (as in MANOVA). The λ obtained here was also used, as a cross-check, to estimate Rao's F-approximation to test the equality of the previous two estimates of variance.

The decision of how many of the eigevalues and their eigenvectors should be retained to significantly explain the original variation can be made with the help of Bartlett's χ^2 test (Seal 1968). The procedures was to test whether all the eigenvalues after the L th, say, could be given zero value.

$$\chi^2_{(p-L)\ (g-L-1)} = -(N-\tfrac{1}{2}(p+g)-1) \log_e \lambda',$$

Where N = the total no. of individuals in the whole population,

$$\lambda' = \prod_{j=L+1}^{q} \frac{1}{1+\phi_j} .$$

If L=0, the $\lambda' = \lambda$ and χ^2 was then testing whether all the eigenvalues could be given zero value (i.e. were non-significant). If χ^2 was significant then this implied that all the eigenvalues (as a whole) were significantly different from zero. Next, the first eigenvalue (also the largest) was removed and all the other eigenvalues were tested for significance. This time L equaled 1. If χ^2 was significant, this implied that from the second to the last eigenvalues (as a whole) were significantly different to zero. Next, the first two eigenvalues were removed, and the rest tested against zero as before. This process continued on until the χ^2 was non-significant, or until all the eivenvalues were removed. The first of these alternatives implied that all the rest of the eigenvalues were not significantly different from zero; whereas the latter implied that all the eigenvalues were significantly different from zero (Cooley and Lohnes 1971, Seal 1968, Kshirsagar 1972).

The proportion of the total discriminating power contained in the k th discriminant function was expressed as percent trace:

$$\text{percent trace} = \frac{\phi_k}{\sum_{j=1}^{q} \phi_j} \times 100.$$

This indicated the percentage of the total discriminating information, available in ($W\text{-}SSCP_x^{-1}$ $A\text{-}SSCP_x$), which was accounted for by the k th discriminant function.

Multiple discriminant analysis can be considered as a special case of canonical correlation analysis in which a set of binary dummy variables on one side of the canonical equation carry the information about group membership (Cooley and Lohnes 1971). The canonical correlation coefficient between the k th discriminant function and the group variables(coded as a set of binary dummy variables) can be obtained as:

$$R_{ck} = \sqrt{\frac{\phi_k}{1+\phi_k}}.$$

R_{ck}, the canonical correlation coefficient, gives the correlation between the p original characters and the g group membership variables (coded in binary dummy variables), and therefore shows the predictive potency of the k th discriminant function. It's square, R_{ck}^2, is the eta-square for the k th discriminant function. This indicates the proportion of variation in the k th discriminant function which is in common with the variation in the specific matching linear function of the group membership variable (Cooley and Lohnes 1971). These statistics also were obtained by the present program.

For ease of ordination and understanding, it is desirable that Y be centralized and standardized. That is,

$$\underset{(qxg)}{Y_c} = \underset{(qxq)}{(T\text{-}MSCP_y)^{-\frac{1}{2}}} \underset{(qxg)}{(Y-\overline{Y})},$$

$$= (T\text{-}MSCP_y)^{-\frac{1}{2}} V' (X-\overline{X}).$$

Let $B = (T\text{-}MSCP_y)^{-\frac{1}{2}} V'$.

$$\therefore Y_c = \underset{(qxp)}{B} \underset{(pxg)}{(X-\overline{X})}$$

B is the matrix of coefficients for transforming deviation scores, $(X-\overline{X})$, to standardized discriminant functions, Y_c.

$$\underset{(qxg)}{Y_c} = \underset{(qxp)}{B} \underset{(pxp)}{SD_x} \underset{(pxp)}{SD_x^{-1}} \underset{(pxg)}{(X-\overline{X})},$$

Where SD_x is the diagonal matrix formed from the diagonal elements of $(T\text{-}MSCP_x)^{\frac{1}{2}}$. That is a diagonal matrix of standard deviations of X.

Let $Z = SD_x^{-1} (X-\overline{X})$

and $C = B \; SD_x$

$$\underset{(qxg)}{Y_c} = \underset{(qxp)}{C} \underset{(pxg)}{Z}$$

C is the matrix of coefficients for transforming standardized scores, (Z), to standardized discriminant function Y_c.

Let R be the correlation matrix based on $T\text{-}MSCP_x$, then the matrix of "factor structure coefficients", S, can be expressed as:

$S = R\,C$.

S is the pxq "structure" matrix of correlations between the p original variables and the q discriminant functions.

Two other interpretative aids can be extracted from the structure matrix, S. These are the table of communalities for each variable, and the percentage of trace of R for each function. The communalities for each variable represent the sums of squares of q elements in every p row of S. This shows the proportion of each original variable being accounted for in the full set of q discriminant functions. This will only be of interest when $(g-1) < p$, since otherwise the communalities are always sum to 1. The percentage of trace of R for each function is found by dividing the sums of squares of p elements in every q th column of S by the trace of R. This shows the proportion of the trace of R being accounted for by each discriminant function. These fraction will sum to 1 only when $(g-1) \geq p$. The present program obtained these statistics as well.

Finally, Y_c, the standaidized discriminant score for each group, was obtained for ordination and further analysis (via tape storage).

2.4 Similarity Measures

Program SIMMAT was used to obtain the similarity measures used in this study. This program was capable of calculating standardized or unstandardized Squard Euclidean Distance (SED) and Euclidean Distance (ED). Either SED or ED could be norminated to be stored as a vector for use in cluster analysis (See listing of SIMMAT in Appendix B-2).

The similarity measures used in this study were standardized SED, obtained from the standardized discriminant scores, Y_c. Since the discriminant scores obtained from MANDIS were already standardized, the unstandardized option of SIMMAT was used (to avoid double standardization).

The use of Y_c rather than X or Z for obtaining the standardized SED concurred with Gower's(1966) proposal of using principal components (See section 1.7.1). Y_c and principal components are similar in the sense that they are both uncorrelated scores. However, where principal components are for observations from one population, Y_c are for observations from more than one population (See Appendix A-2). The standardized SED obtained from Y_c was expected to provided a good estimate of the similarity between groups for subsequent use in cluster analysis (Ratkowsky 1977, pers. Com.).

2.5 Cluster Analysis

Program CLUSAN was adapted and modified from subroutines of Anderberg(1973). At present CLUSAN has seven optional clustering strategies. They are Centroid, Complete Linkage, Average Linkage Within New Cluster, Single Linkage, Average Linkage Between Merged Clusters, Median and Ward's Method. According to Williams(1971), these are all hierarchical, polythetic, agglomerative clustering procedures. During each execution of CLUSAN, subroutine TREE produced the dendrogram of the Clustering.

All seven strategis have been applied to the full set of attributes (ALLCHARA); and the properties of each clustering stategy and the dendrogram has been studied and compared briefly. Those strategies with undesirable properties were eliminated then, and a final one was chosen to analyse the other sets of attributes (i.e. AGROCHARA, DISCCHARA and JACQCHARA).

2.6 Post Clustering Analysis

After clustering, program SEFWIG was used to reveal the relationships between a given hierarchical agglomerative strategy and each of the characters (attributes). Program SEFWIG (selected error for within group) was adapted and modified from ERROR of Anderberg(1973). The modifications included the calculating of F-ratios (and their associate probabilities) for each selected characters, and for the clustering criterion of Ward's Method. The latter was used to decide the clustering "cut-off" point.

SEFWIG examined the growth in the "error" sum of squares (i.e. pooled within-cluster sum of squares) of each attribute as clustering progressed through increasing levels of aggregation. At the beginning of agglomerative clustering, each individual was considered as a cluster on its own (i.e. a cluster with only one individual). Then, there was no within-cluster sum of square (WSS), and the total sum of squares (TSS) was solely represented by the among-cluster sum of squares (ASS). As clustering proceeded, TSS remained constant, WSS increased and ASS decreased. At the last stage of clustering, there was only one cluster, and it contained every individual. Then, TSS was solely represented by WSS and there was no ASS. At any stage of clustering, the ratio of WSS to TSS was the portion of the total sums of squares not explained by the current set of clusters (Anderberg 1973). This ratio was the complement of Eta-Square (ASS/TSS) (Refer to section 2.3.1 for Eta-Square). The growth of this ratio for each character may be different. For some characters the ratio may become large at early stages of clustering; whereas, for others, the ratio may remain small even until the last few stages. The former characters were considered as dormant and the latter as dominant. Dormant characters contributed little to the clustering results, and their elimination has little effect. Conversely elimination of dominant characters will have a marked influence on the clustering results. The F-ratio (Among cluster mean square to within cluster mean square), and the associated probability, for each character were also obtained.

As Ward's Method of clustering was based on the minimum increment of WSS of all characters as a whole, the clustering criterion was WSS. At every stage of clustering, SEFWIG calculated the overall F-ratio and the associated probability for the clustering criterion. These associated probabilities helped in deciding the most "suitable" stage to "cutoff" the clustering, and so to obtain the set of clusters for further consideration. Some basic properties of this application to Ward's Method of clustering, were studied.

After the clusters have been obtained, their structure can be studied by program POSTCA. This was adapted and modified from POSTDU of Anderberg(1973). The major modification was the inclusion of subroutine DIFFS (Gordon unpublished). POSTCA listed the clusters' memberships, and the attributes for each cluster. Next, DIFFS ranked the means of each cluster (one attribute at a time) and performed least significant difference tests or Duncan's multiple range tests.

Program CONVER has been developed to study the realtionships between Centroid, Average Linkage Between Merged Clusters, and Ward's Methods. The program was based on Gower's(1970) conversion equations. It converted the clustering criterion of any one of these three methods to the clustering criterion of the other two. It also reported the increment of pooled within cluster sums of squares due to that merge, and the variance of the newly merged (formed) cluster. F-ratio and the associated probability was claculated for each stage of clustering. This program helped in deciding the "cutoff" points for any of the three methods.

CHAPTER 3 RESULTS AND ASSOCIATED DISCUSSION

3.1 MANDIS

The mean of each character for each group is summarized in Appendix C-1. The grand mean, standard deviations and coefficient of variation for 160 groups are lsited in Table 3.1. The coefficient of variation ranged from 17.11%(F.DAT) to 40.22%(L.COL). The latter was considerably high. However, it was not surprising, as the within-plot variance was based on open-pollinated plants, containing potential genotypic diversity even within the one population. As there were four sets of data analysed, discussion here will be concentrated on ALLCHARA and the others (AGROCHARA, DISCCHARA, and JACQCHARA) will be discussed only briefly, to minimise repetition.

3.1.1 ALLCHARA

Prior to analysis, the equality of the group MSCP matrices was tested by two approaches: χ^2 test and F-test of Box's M Criterion. The results of these tests have been presented in Table 3.2. Both of these tests rejected the hypothesis that the MSCP matrices were equal, with very high significance level ($P<0.0001$). The exact cause of this rejection was not known. It could be due to the actual difference in dimension and orientation or due to non-multivariate-normality of the distribution, or both. However, non-multivariate-normality has generally been believed to be the main cause. By examining the marginal distributions (Appendix C-2), it was found that some characters (such as L.ROL and L.COL) were highly skewed and did not fit the marginal normality. This implied that multivariate normality was not satisfactory. As both the above tests are highly sensitive to non-normality (Seal 1968, Press 1971), the rejection of the null hypothesis would be expected from this cause alone.

Wilk's λ =0.272866, and F-approximation=2.8670 with df_1=1749 and df_2=39915. These results rejected the hypothesis that group centroids were equal, with a very high significance level ($P< 0.0001$). The complement of Wilk's λ (eta-square)=0.727134. This is the generalized coefficient of multiple determination, which shows the proportion of generalized total sums of squares (determinant of T-SSCP) explainable by the generalized sums of squares due to grouping (determinant

Charac -ter	Mean	S.D.[1]	C.V.[2]	AMS[3]	WMS[4]	F-ratio	Eta Square	Eta
C.DIA	6.4452	1.4736	0.2193	4.8145	1.9983	2.4093	0.09515	0.30846
C.DEN	6.4875	1.3124	0.2023	3.4053	1.7224	1.9771	0.07944	0.28185
C.ERE	4.6476	1.8393	0.3957	18.5039	3.3829	5.4699	0.19273	0.43901
C.HEI	4.1588	1.5144	0.3641	7.2677	2.2933	3.1691	0.12151	0.34858
RUST	3.6137	0.9544	0.2641	2.3600	0.9109	2.5909	0.10159	0.31873
O.DIS	2.8307	0.9018	0.3186	2.0252	0.8133	2.4900	0.09803	0.31310
L.ROL	1.3526	0.4917	0.3635	0.4995	0.2418	2.0661	0.08272	0.28761
L.COL	1.3255	0.5331	0.4022	0.5009	0.2842	1.7625	0.07143	0.26726
L.WID	2.8501	0.6507	0.2283	1.0059	0.4235	2.3756	0.9394	0.30650
F.COL	2.7586	0.7187	0.2603	1.5270	0.5165	2.9565	0.11429	0.33807
F.DAT	5.3831	0.9211	0.1711	5.2935	0.8485	6.2386	0.21401	0.46261

1. Standard deviation based on pooled within group mean squares.
2. Coefficient of variation.
3. Among groups sum of squares.
4. Pooled within groups sum of squares.

TABLE 3.1 Analyses Summary Of All Characters For 160 groups
(For F-ratio $df_1 = 159$, $df_2 = 3643$)

Box's M	= 16187.742		
F-ratio	= 1.26173	df_1 =	10494
		df_2 =	760375
		Prob.	0.0001
χ^2	= 13463.974	df =	10494
		Prob. <	0.0001

TABLE 3.2 The Results Of Equality Test Of MSCP Matrices, For ALLCHARA

of A-SSCP). It is a generalized coefficient because it considers all the characters simultaneously (Refer to Drapper and Smith(1966) for the definition of coefficient of multiple determination). This shows the overall efficiency by which the groups (considering all the characters simultaneously) were represented by their own centroids. In this case, it was 72.71%.

The non-independent" univariate F-ratios (Cooley and Lohnes 1973) are listed in Table 3-1. Though all the F-ratios were highly significant ($P<0.0001$), it should not be inferred automatically that group differences in each character were significant as the characters were "not independent" (see Section 2.3.1). The eta-squares, which are similar to the coefficient of multiple determination, showed the efficiency by which the groups were represented by their means, on a univariate basis. The eta-values, which are similar to the coefficient of multiple correlation, suggested that F.DAT contributed most to the discrimination amongst groups. This was followed by C.ERE, C.HEI, F.COL, RUST, O.DIS, C.DIA, L.WID, L.ROL, C.DEN, and L.COL, in that order.

The results from the multiple discriminant analysis are shown in Table 3.3. In this Table, R_{cj} is the canonical correlation coefficient between the j th discriminant function and the group variables. The value of R_{cj} for the first discriminant function was 0.5010. This showed that the 1st discriminant function had a predictive potency of 0.5010 between the 11 original variables and the 160 group membership variables. The second function had a canonical correlation of 0.40[illegible] and so on. The R^2_{cj} represented the eta-square of j th discr[illegible] function. The eigenvalue, ϕ_j indicated the "generalized variation" (available in $W\text{-}SSCP_x^{-1}$. $A\text{-}SSCP_x$) of the j th discriminant function. The decreasing trend in the eigenvalues (Table 3.3) was expected as the orthogonal functions were extracted according to their discriminating abilities. The % trace of ϕ_j showed the proportion of the total variation accounted for by each of the j functions. The first three functions collectively accounted for 51.1% of the total variation available in $W\text{-}SSCP_x^{-1}$. $A\text{-}SSCP_x$. $\chi^2_L (=\chi^2 j\text{-}1)$ is the χ^2 for testing whether all the eigenvalues after L th, can be regarded as zero value. The results showed that even the eleventh (last) eigenvalue (i.e. after removing the first ten largest eigenvalues), was highly significant ($P=0.0055$). These suggested that all the eleven discriminant

j	R_{cj}	R^2_{ij}	ϕ_j	% trace of ϕ_j	λ	χ^2_L	$df_{(x^2)}$	$Prob_{(x^2)}$
1	0.5010	0.2510	0.33510	28.82	0.2729	>1000	1749	0.0000
2	0.4037	0.1630	0.19472	13.84	0.3643	>1000	1580	0.0000
3	0.3989	0.1591	0.18920	14.45	0.4352	>1000	1413	0.0000
4	0.3220	0.1037	0.11568	8.22	0.5176	>1000	1248	0.0000
5	0.3205	0.1027	0.11447	8.13	0.5775	>1000	1085	0.0000
6	0.2990	0.0894	0.09815	6.98	0.6436	>1000	924	0.0000
7	0.2880	0.0829	0.09044	6.43	0.7067	>1000	765	0.0000
8	0.2690	0.0724	0.07800	5.54	0.7706	968.3	608	0.0000
9	0.2625	0.0689	0.07401	5.26	0.8307	689.2	453	0.0000
10	0.2409	0.0580	0.06162	4.38	0.8922	423.8	300	0.0001
11	0.2298	0.0528	0.05574	3.96	0.9472	201.6	149	0.0055

TABLE 3.3 Results From Multiple Discriminant Analysis, For ALLCHARA

functions were required, inorder to retain sufficient amount of the original variation.

Table 3.4 shows the S matrix (the correlations between the discriminant functions and the original characters). The first two discriminant functions were associated mainly with F.DAT, C.ERE, and C.HEI. This is seen from the following: the first discriminant function correlated -0.8094 with F.DAT, 0.6969 with C.ERE, -0.3841 with F.COL, -0.3085 with L.ROL and 0.2629 with C.HEI. The second discriminant function correlated 0.5952 with C.ERE, 0.5220 with F.DAT, 0.2840 with C.HEI and 0.2325 with C.DEN.

From the S matrix and the nature of the original characters, the 11 functions appeared to be measuring, respectively; (1) clump elevation measures(+) versus flowering characters(-); (2) clump compactness, elevation, and F.DAT measures; (3) C.HEI(+) versus C.DIA and disease measures(-); (4) overall size, and disease measures; (5) C.DEN and O.DIS(+) versus L.WID and F.COL(-), (6) clump measures(-) versus O.DIS and colour measures(+); (7) Horizontal size measures(-) versus disease and flowering measures(+); (8) F.COL and clump measures(-) versus L.WID and RUST(+); (9) L.WID and clump measures(-) versus leaf nature and disease measures(+); (10) L.ROL(+) versus L.COL(-), and (11) clump and leaf measures.

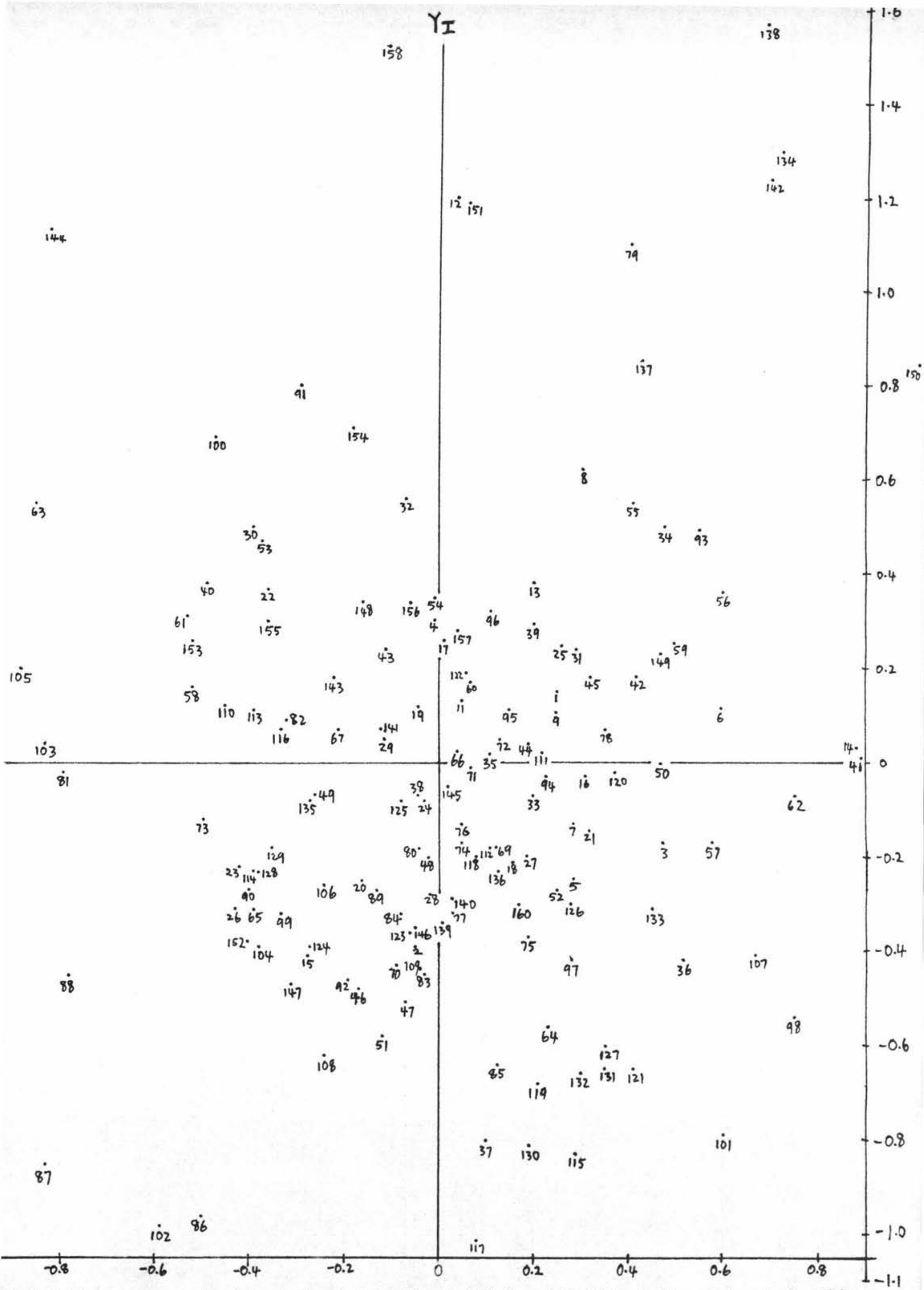
Table 3.5 shows the B matrix (the coefficients for producing standardized discriminant functions from the group deviation vectors). The location of the 160 group centroids for all eleven discriminant functions would fully describe the group differences from this data. The 160 group centroids for the first two discriminant functions have been displayed in Figure 3.1. This "ordination" (in Figure 3.1) was not very informative in the present case, and should be interpreted with care for the following reasons. The distances between group centroids (for these first two discriminant functions) did not represent the total centroid dissimilarity. This was due to the fact that the first two discriminant functions accounted for only 37.65% (refer to % trace of ϕ_j of Table 3.3) of the total discriminating information, and the other functions were significantly important, accounting for the other 62.35%. However, as the first two functions do have the largest individual discriminantory abilities, this "ordination" does give a rough indication of the group differences. It is impracticable

Character	Discriminant Function 1	2	3	4	5	6	7	8	9	10	11
C.DIA	-0.0657	-0.1451	-0.3302	0.4013	0.1457	-0.3714	-0.5643	-0.4208	-0.0748	-0.1671	0.1281
C.DEN	-0.0620	0.2325	-0.1628	-0.0380	0.4208	-0.2474	0.1482	-0.2572	-0.4857	-0.1856	0.5629
C.ERE	0.6969	0.5957	0.0768	0.1818	-0.0154	-0.1295	-0.0968	-0.2433	0.0525	0.1186	0.1324
C.HEI	0.2629	0.2840	0.4583	0.5616	-0.1372	-0.3546	-0.0419	-0.2586	-0.2238	0.0515	0.2416
RUST	0.0527	-0.0962	-0.4767	0.2855	-0.1777	-0.0844	0.7129	0.2154	0.2185	-0.1056	-0.1625
O.DIS	-0.0074	-0.1096	-0.2668	0.5075	0.3766	0.5333	0.4213	0.0724	0.1189	0.0817	-0.1699
L.ROL	-0.3085	-0.1732	-0.1431	-0.0600	0.1376	-0.0779	0.1853	-0.0910	0.2842	0.5717	0.6123
L.COL	-0.1436	-0.0881	0.2010	-0.0048	-0.0101	0.2392	-0.0383	-0.0341	0.5677	-0.6217	0.4024
L.WID	0.1468	0.1210	-0.2874	0.3014	-0.4167	0.1295	-0.4755	0.4728	-0.2894	-0.5470	0.2509
F.COL	-0.3841	0.0545	-0.2352	-0.0732	-0.4405	0.4860	0.2093	-0.5470	-0.1209	-0.0094	0.0176
F.DAT	-0.8094	0.5220	-0.0642	0.0650	0.0431	-0.0288	0.1087	0.0371	0.0341	0.1276	-0.1738

TABLE 3.4 The S matrix - The Correlations Between The Discriminant Functions And The Original Characters, for ALLCHARA

	Discriminant Function										
Character	1	2	3	4	5	6	7	8	9	10	11
C.DIA	-0.0739	-0.1571	-0.3718	0.2398	0.1034	-0.2630	-0.3662	-0.3370	0.1823	-0.0773	-0.1533
C.DEN	0.0011	0.1515	-0.1509	-0.1852	0.3371	-0.0359	0.2511	0.0214	-0.4869	-0.3172	0.3591
C.ERE	0.3037	0.4423	-0.2196	-0.1057	0.0475	0.0571	-0.0792	-0.1427	0.2839	0.0901	-0.0001
C.HEI	-0.1058	-0.1439	0.5567	0.5020	-0.2320	-0.1401	0.1982	-0.0595	-0.2088	0.0416	0.1088
RUST	0.0997	-0.0317	-0.5231	0.1406	-0.4552	0.6344	0.6467	0.1341	0.2371	-0.2934	-0.0574
O.DIS	0.0603	-0.0723	0.0336	0.6744	0.7033	0.8175	0.0316	0.0313	-0.0866	0.1649	-0.0889
L.ROL	-0.2185	-0.3753	-0.2110	-0.0035	-0.1043	-0.1495	0.0409	0.0175	0.7735	1.4817	1.2351
L.COL	-0.2053	0.2224	0.3850	0.1348	-0.0149	0.3533	-0.0748	0.0858	1.0842	-1.2061	0.7692
L.WID	-0.0062	0.2307	-0.4365	0.3457	-0.5558	0.4441	-0.4958	0.9300	-0.4021	0.0238	0.6281
F.COL	-0.1150	-0.0233	-0.2128	-0.1204	-0.7913	0.6309	0.1800	-0.9223	-0.3088	0.0183	0.0667
F.DAT	-0.6385	0.7645	-0.0407	0.1394	0.1135	-0.1588	-0.0991	0.1943	0.2314	0.0262	-0.2200

TABLE 3.5 The B Matrix - The Coefficients For Producing Standardized Discriminant Function (Y_c) from group deviation vectors ($X - \overline{X}$), for ALLCHARA



RE 3.1 The group centroids ordinated in the 1st(Y_I) and 2nd(Y_{II}) discriminant functions.

to diagram more than two functions at a time, and so this is the only choice available for Fig. 3.1.

3.1.2 AGROCHARA

The results of the equality test of the group MSCP matrices are given in Table 3.6. Both the χ^2 test and the F test rejected the null hypothesis that the MSCP matrices were equal, with a very high significance level ($P<0.0001$). Though the highly skewed characters (e.g. L.ROL and L.COL) were excluded in this set of data, non-multivariate normality was still believed to be the main cause for rejection. By visual examination, the distribution of C.ERE was found to be too "flat" for marginal normal distribution (refer to Appendix C-2). Though the marginal distribution of other characters may have resembled binomial distributions,their joint distribution may not necessarily be multivariate normal when the interactions were encountered (Andrew et. al.1971, Press 1971, Rohlf 1971).

The value obtained for Wilk's Lambda was 0.344865, and F approximation = 3.25746 with df_1= 1272, df_2 = 29074. These results rejected the hypothesis that the group centroids were equal, with a very high significance level ($p< 0.0001$). The complement of Wilk's λ = 0.655135. This eta-square showed that the overall efficiency of the groups (considered all the 8 characters simultaneously) being represented by their own centroids was 65.51%

The detail univariate investigation revealed the same results as for ALLCHARA (as listed in Table 3.1). These results were expected, as these univariate studies did not take the covariance between characters into consideration. The order of magnitude of the contribution of each character to the discrimination amongst groups (ranked according to their eta value, in Table 3.1) was: F.DAT, C.ERE, C.HEI, RUST, O.DIS, C.DIA, L.WID and C.DEN.

The results from the multiple discriminant analysis are shown in Table 3.7. The 1st discriminant function obtained a canonical correlation of 0.4952. This showed that it produced a predictive potency of 0.4952 between the 8 original variables and the 160 group membership variables. The second discriminant function showed a canonical correlation of 0.4008, which is interpreted similarly, and so on. The R^2_{cj}

Box's M = 9086.5315

F-ratio = 1.37945 df_1 = 5724
df_2 = 774679
Prob. < 0.0001

χ^2 = 7963.099 df = 5724
Prob. < 0.0001

TABLE 3.6 The Results Of Equality Test Of MSCP Matrices, For AGROCHARA

j	R_{cj}	R^2_{cj}	$\emptyset_j$	% trace of $\emptyset_j$	λ	χ^2_L	df	Prob
1	0.4952	0.2452	0.32494	28.02	0.3449	3958.19	1272	0.0000
2	0.4008	0.1607	0.19142	16.51	0.4569	2912.08	1106	0.0000
3	0.3885	0.1509	0.17772	15.33	0.5444	2260.90	942	0.0000
4	0.3216	0.1034	0.11538	9.95	0.6411	1652.70	780	0.0000
5	0.3093	0.0957	0.10582	9.13	0.7151	1246.72	620	0.0000
6	0.2886	0.0833	0.09087	7.84	0.7908	872.75	462	0.0000
7	0.2842	0.0808	0.08785	7.58	0.8626	549.38	306	0.0000
8	0.2482	0.0616	0.06562	5.66	0.9384	236.32	152	0.0002

TABLE 3.7 The Results From Multiple Discriminant Analysis, For AGROCHARA

gave the eta-square of the j th discriminant function. The 1st eigenvalues, ϕ_1, was 0.32494, and these gradually reduced to 0.06562 for the 8th (last) eigenvalue. This decreasing trend was expected as the orthogonal functions were extracted according to their discriminating abilities. The % trace of ϕ_j showed that the first three discriminant functions collectively accounted for 59.8% of the total discriminating information available. χ^2_L (= χ^2_{j-1}) was the χ^2 for testing whether all the eigenvalues after L th can be regarded as zero value, as noted earlier. The results showed that even the 8th (last) eigenvalue (i.e. after removing the seven largest eigenvalues) was highly significant (P=0.0002). This suggested that all eight discriminant functions were required to retain the original variation.

Table 3.8 shows the S matrix (the correlations between the discriminant functions and the original characters). From this table, it was noted that the first two discriminant functions were associated mainly with C.ERE, F.DAT and L.WID. This was seen from the following. The first discriminant function correlated -0.8225 with F. DAT, 0.7016 with C.ERE, 0.2628 with C.HEI, and 0.1507 with L.WID. The second discriminant function correlated 0.5090 with C.ERE, 0.4848 with F.DAT, 0.2843 with C.DEN, and 0.2380 with L.WID.

From the S matrix and the nature of the original characters, the 8 discriminant functions appeared to be measuring, respectively, (1) clump elevation measures(+) versus F.DAT(-); (2) F.DAT, clump elevation and compactness measures; (3) C.DIA and disease measures(-) versus clump height measures; (4) overall size and disease measures; (5) C.DEN and O.DIS(+) versus L.WID(-); (6) horizontal size measures(+) versus C.DEN and disease measures(-); (7) clump measures(+) versus leaf and disease measures(-); and (8) clump and leaf measures(+) versus F.DAT and disease measures(-).

Table 3.9 shows the B matrix (the coefficients for producing standardized discriminant functions from group deviation vectors).

3.1.3 DISCCHARA

The results of the equality test of MSCP matrices are listed in Table 3.10. Both the F test and χ^2 test rejected the null hypothesis that MSCP matrices were equal. The main cause of the rejection was

	Discriminant Function							
Charac -ter	1	2	3	4	5	6	7	8
C.DIA	-0.0716	0.0265	-0.3879	0.3758	-0.0162	0.2245	0.7938	0.1455
C.DEN	-0.0775	0.2843	-0.0444	-0.0838	0.3013	-0.3366	0.3757	0.7473
C.ERE	0.7016	0.5090	0.3662	0.1888	0.0268	-0.0078	0.2784	0.0041
C.HEI	0.2628	0.0467	0.5637	0.6017	-0.1118	-0.1306	0.3723	0.2832
RUST	0.0573	0.1369	-0.4804	0.2951	-0.1819	-0.6186	-0.4466	-0.2104
O.DIS	-0.0060	0.0258	-0.3321	0.4281	0.6475	-0.1804	-0.4465	-0.2325
L.WID	0.1507	0.2380	-0.1963	0.3496	-0.3706	0.6124	-0.2472	0.4311
F.DAT	-0.8255	0.4848	0.1949	0.0711	0.0338	0.0793	-0.0789	-0.1621

TABLE 3.8 The S Matrix - The Correlations Between The Discriminant Functions and The Original Characters, For AGROCHARA

Character	Discriminant Function 1	2	3	4	5	6	7	8
C.DIA	-0.0800	0.0356	-0.4236	0.2163	-0.0299	0.1199	0.5569	-0.2367
C.DEN	-0.0320	0.1831	-0.1135	-0.2260	0.2469	-0.2851	0.0067	0.6628
C.ERE	0.3250	0.5087	0.0134	-0.1245	0.0766	0.0524	0.1075	-0.2343
C.HEI	-0.0954	-0.3900	0.4535	0.5515	-0.1162	-0.2041	-0.0145	0.1895
RUST	0.1043	0.2171	-0.4882	0.1992	-0.7155	-0.7900	-0.1048	-0.0930
O.DIS	0.0449	-0.0822	-0.0784	0.5492	1.0812	0.2501	-0.3283	-0.0654
L.WID	0.0348	0.4228	-0.2676	0.3976	-0.3940	0.8677	-0.7276	0.7870
F.DAT	-0.6690	0.6976	0.2490	0.1231	-0.0587	0.0565	0.0107	-0.2624

TABLE 3.9 The B Matrix - The Coefficients For Producing Standardized Discriminant Functions (Y_c) From Group Deviation Vectors ($X - \overline{X}$) For AGROCHARA

Box's M	= 3702.0385		
F-ratio	= 1.42572	df_1 =	2385
		df_2 =	825241
		Prob.	0.0001
χ^2	= 3411.044	df =	2385
		Prob.	0.0001

TABLE 3.10 The Results Of Equality Test Of MSCP Matrices, For DISCCHARA

believed to be the non-multivariate normality, as stated in section 3.1.2.

The results of the Wilk's Lambda test estimated a $\lambda = 0.469817$ and F approximation = 3.73398 with $df_1 = 795$ and $df_2 = 18194$. These results rejected the hypothesis that the centroids were equal, with a very high significance level ($P < 0.0001$). The complement of Wilk's λ, eta-square = 0.530183. This showed that the overall efficiency of the groups (considering 5 characters simultaneously) being represented by their own centroids was 53.02%.

The five characters in this set of data were the most discriminatory characters, in that they had the largest eta values (refer to Table 3.1). Their discriminatory ranking in descending order, was F.DAT, C.ERE, C.HEI, F.COL and RUST.

The results from the multiple discriminant analysis are listed in Table 3.11. The 1st discriminant function obtained a canonical correlation of 0.4944. This showed that it produced a predictive potency of 0.4944 between the 5 original variables and the 160 group membership variables. The % trace of ϕ_j showed the first two discriminant functions collectively accounted for 60.66% of the total discriminating information available in this set of characters. χ^2_L showed that even the 5th eigenvalue was highly significant ($P \angle 0.0001$). These suggested that all the 5 discriminant functions were required to retain the original variation.

Table 3.12 shows the S matrix. The first two discriminant functions accounted mainly for F.DAT, C.ERE and C.HEI, as noted (as before) by an examination of the correlation trends.

From the S matrix and the nature of the original characters, the 5 discriminant functions appeared to measure, respectively, (1) clump measures(+) versus flower measures(-); (2) F.DAT and clump measures; (3) clump measures(+) versus RUST and F.COL(-); (4) RUST and C.HEI measures;; (5) RUST and DAT(-) versus F.COL and clump measures(+).

Table 3.13 shows the coefficients for producing standardized discriminant functions from group deviation vectors.

j	R_{cj}	R^2_{cj}	ϕ_j	% trace of ϕ_j	λ	χ^2_L	df	Prob.
1	0.4944	0.2445	0.32356	38.97	0.4698	>1000	795	0.0000
2	0.3914	0.1532	0.18094	21.79	0.6218	>1000	632	0.0000
3	0.3437	0.1181	0.13395	16.13	0.7343	>1000	471	0.0000
4	0.3072	0.0944	0.10421	12.55	0.8327	680.94	312	0.0000
5	0.2838	0.0805	0.08757	10.55	0.9195	312.22	155	0.0000

TABLE 3.11 The Results From Multiple Discriminant Analysis For DISCCHARA

	Discriminant Function				
Character	1	2	3	4	5
C.ERE	-0.6975	0.6393	0.2626	0.0821	0.1693
C.HEI	-0.2466	0.2806	0.7709	0.4595	0.2336
RUST	-0.0723	-0.0496	-0.5827	0.7236	-0.3586
F.COL	0.3870	0.0546	-0.4757	0.1854	0.7656
F.DAT	0.8268	0.5377	-0.0826	0.0747	-0.1198

TABLE 3.12 The S Matrix - The Correlations Between The Discriminant Functions And The Original Characters, For DISCCHARA

	Discriminant Function				
Character	1	2	3	4	5
C.ERE	-0.3354	0.5123	-0.2141	-0.1525	0.0369
C.HEI	0.1488	-0.1810	0.5654	0.5366	0.1591
RUST	-0.1685	-0.0110	-0.4343	0.8421	-0.3879
F.COL	0.1348	-0.0282	-0.4977	0.2593	1.3088
F.DAT	0.6398	0.7762	0.0409	-0.0224	-0.3480

TABLE 3.13 The B Matrix - The Coefficients For Producing Standardized discriminant functions from group Deviation Vectors, For DISCCHARA

3.1.4 JACQCHARA

The results of the equality test of MSCP matrices are listed in Table 3.14. Both the F test and χ^2 test rejected the null hypothesis that MSCP matrices are equal. The main cause of rejection was believed to be the non-multivariate normality, as stated in section 3.1.2.

The results of Wilk's Lambda (0.527671), and F approximation (= 3.96824 with df_1 = 636, df_2 = 14560) rejected the hypothesis that the group centroids were equal, with a very high significance level ($p < 0.0001$). The complement of Wilk's λ (= 0.472329) showed that all the overall efficiency of the groups (considered all 4 characters simultaneously being represented by their own centroids was 47.23%.

The results of multiple discriminant analysis are shown in Table 3.15. The 1st discriminant function had a canonical correlation of 0.4887, indicating its potency. The decreasing trend of the eigenvalues again was usual. The % of trace of ϕ_j showed that the first two discriminant functions collectively accounted for 70.29% of the total discriminating information available in this set of characters. χ^2_L showed that even the 4th eigenvalue was highly significant ($p < 0.0001$).

Table 3.16 shows the S matrix. The first two discriminant functions were associated mainly with F.DAT and C.ERE.

From the S matrix and the nature of the original characters, the 4 discriminant functions appeared to measure, respectively, (1) C.ERE(+) versus F.DAT(-); (2) C.ERE, F.DAT and L.WID measures; (3) RUST and L.WID(+) versus C.ERE(-); and (4) L.WID(+) versus RUST and F.DAT(-). The B matix is shown in Table 3.17.

3.1.5 Additional Comments

For all sets of data (ALLCHARA, AGROCHARA, DISCCHARA and JACQCHARA), the MSCP matrices equality tests have shown a very highly significant result, with a consequent rejection of the null hypothesis that they were equal. This was believed to be due to the non-normality of the data distribution. However, the actual differences in dimension and orientation of the MSCP matrices could not be ruled out either. The large sample sizes involved (as shown by their large df in Table

Box's M = 2454.6582

F-ratio = 1.44317 df_1 = 1590
df_2 = 877095
Prob. < 0.0001

χ^2 = 2299.096 df = 1590
Prob. < 0.0001

TABLE 3.14 The Results Of Equality Test Of MSCP Matrices. For JACQCHARA.

j	R_{cj}	R^2_{cj}	ϕ_j	% trace of ϕ_j	λ	χ^2_L	df	Prob
1	0.4887	0.2388	0.31380	44.84	0.5277	2378.13	636	0.0000
2	0.3908	0.1527	0.18025	25.55	0.6933	1326.85	474	0.0000
3	0.3340	0.1116	0.12556	17.80	0.8183	746.35	314	0.0000
4	0.2812	0.0790	0.08583	12.17	0.9210	306.33	156	0.0000

TABLE 3.15 The Results From Multiple Discriminant Analysis. For JACQCHARA.

	Discriminant Function			
Character	1	2	3	4
C.ERE	0.6954	0.6896	-0.2011	-0.0017
RUST	0.0689	-0.0292	0.8172	-0.5710
L.WID	0.1506	0.1997	0.5061	0.8251
F.DAT	-0.8564	0.4887	0.0523	-0.1563

TABLE 3.16 The S Matrix - The Correlations Between The Functions And The Original Characters. For JACQCHARA.

	Discriminant Function			
Character	1	2	3	4
C.ERE	0.2626	0.4353	-0.0796	-0.1045
RUST	0.1851	0.0558	0.8362	-0.5601
L.WID	-0.0155	0.2703	0.8627	1.2189
F.DAT	-0.7151	0.7319	0.0377	-0.0697

TABLE 3.17 The B Matrix - The Coefficients For Producing Standardized Discriminant Functions From Group Deviation Vectors. For JACQCHARA.

3.2, 3.6, 3.10 and 3.14) could also have contributed to their apparent rejection, as these tests are sensitive to large sample size (Cooley and Lohnes 1973). In short none of these sets of data fulfilled the assumption of MANOVA. Many research workers prefer to ignore this fact, and make inferences from their results in the belief that MANOVA is robust enough (Cooley and Lohnes 1973). A similar approach is common in practice with ANOVA (Cochran 1947).

Non-normality of the multivariate distribution and non-equality of MSCP matrices, were believed to increase Type I error (Press 1971). However, the significances of group centroid were so high ($p < 0.0001$), the inferences concerning centroid differences could still be sufficiently correct in practice. These results were used as a descriptive guide for the data in this study. However, the main purpose of MANOVA here was to reduce the data, and to organise it for multiple discriminant analysis.

The χ^2 test for eigenvalues retention was also affected by the non-equality in MSCP matrices (Seal 1968). However, ordination of group centroids using all discriminant functions would not affected (because all discriminant functions would be retained in that case). The χ^2 test of this study suggested that retention of all the discriminant functions for all cases was needed in any case, so that this problem did not arise here. The main purpose of multiple discriminant analysis in this study was to transform the correlated original score vectors (x) to standardized uncorrelated discriminant score vectors (Y_c). The removal of these correlations between characters is important for estimating the similarity measure (SED) in cluster analysis (Refer to section 1.7.2).

The study of S matrix for each data set revealed two important points. (1) The first 5 discriminant functions of AGROCHARA accounted for similar characters to the first 5 discriminant functions of ALLCHARA. Furthermore, the 6th and 7th discriminant functions of AGROCHARA were similar to the 7th and 9th discriminant functions of ALLCHARA. This showed that the exclusion of the 3 characters. (F.COL, L.COL and L.ROL) from AGROCHARA did not alter the pattern of discrimination very much. This implied that the 3 characters were not very important for discrimination. (2) For all cases, F.DAT and C.ERE were the two major characters associated with the first two discriminant functions. This showed their importance as the discriminating characters. These findings agreed with the eta values of the characters (Table 3.1): F.DAT, C.ERE, F.COL, L.ROL and L.COL ranked 1, 2, 4, 9 and 11 respectively in their eta values.

For all cases, the communalities of all discriminant functions over every character was 0.9994742. This meant that, in all cases, over all characters, 99.95% of the original information was recovered in all the discriminant functions, collectively. This was expected, as $g-1 = 159 > p = 11$ (or 8, or 5, or 4) (see section 2.3.2). The slight deviation from 1.0 was probably due to rounding error.

3.2 Comparison of Different Clustering Strategies, Using ALLCHARA Attributes.

The standardized SED's (Square Euclidean Distances) obtained from SIMMAT were used for the clustering analysis (program CLUSAN). The value of the merging criteria at each stage of clustering and the dendrogram have been obtained for each of the seven strategies, for the ALLCHARA set of attributes. However only the dendrograms have been presented (Figures 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 & 3.9). In each cases, the full dendrogram, which involved 160 groups, has not been shown. Only the stages of clustering after the "cut off" point were shown (i.e. after the formation of clusters was well advanced). In the dendrogram the positions of the clusters along the vertical axis have no meaning, since any two clusters may be rotated about their point of fusion.

In order to compare the properties (clustering behaviour) of the different strategies on a same basis, it was desirable to "cut off" the clustering at the same stage, and to obtain the same number of clusters for each strategy. As the clustering criterion of Ward's method was the within-cluster sums of squares, the probability of the F-ratio (AMS/WMS) could be used, objectively, to decide the "cutting off" point for clustering (Program SEFWIG was used for this purpose and is discussed subsequently). A brief comparison of the seven strategies is presented in section 3.2.9.

3.2.1 Results of SEFWIG as Applied to Ward's Method

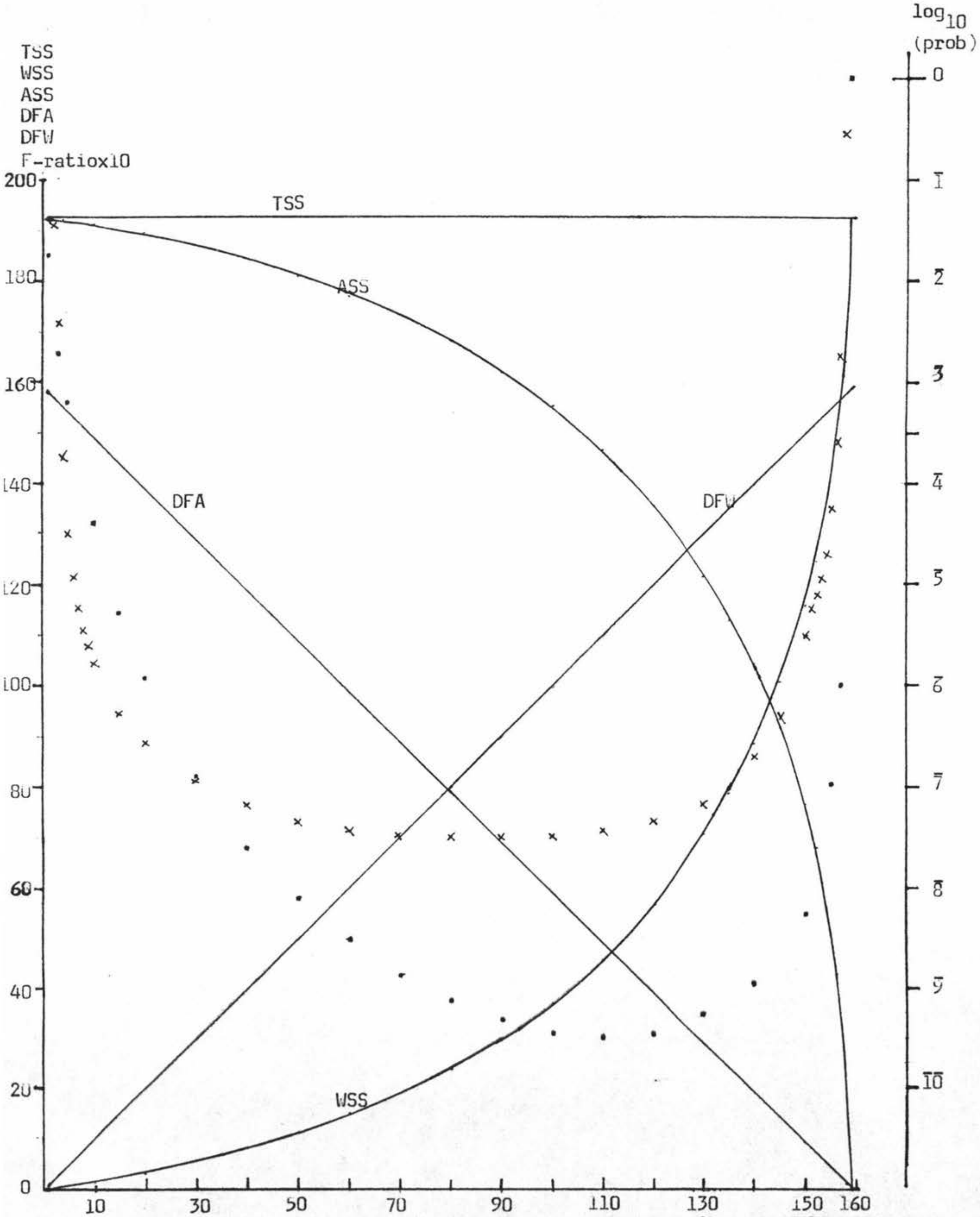
As clustering proceeded from stage 0 to stage 159, the number of clusters decreased from 160 to 1. The total sum of squares remained constant at 192.56. The pooled within-cluster sum of squares increased gradually from 0 to 192.56, whereas the amongst-cluster sum of squares decreased from 192.56 to 0. Also the degree of freedom for pooled within-cluster increased linearly from 0 to 159, as the degree of freedom for among-cluster sum of squares decreased linearly from 159 to 0. The overall F-ratio dropped from 22.9693 (at stage 1) to a minimum value of 7.01159 (at stage 79), then it fluctuated between 7.014 and 7.042. It next increased from 7.02613 (at stage 99) to 20.9164 (at stage 158). The associated probability dropped from 0.17056 (at stage 1) to a minimum value of 0.3326×10^{-8} (at stage 110) and increased to 0.69282×10^{-4} (at stage 158). All these trend have been shown in Figure 3.2.

The aim of clustering was to gather the most similar entities into the same cluster, and to segregate the dissimilar entities into different clusters, thus reducing the number of entities. However, as clustering proceeded and clusters merged, the internal homogeneity of clusters decreased, the sacrifice of this internal homogeneity is unavoidable as the number of cluster is reduced. The probability of the overall F-ratio showed the probability that the among-cluster dissimilarity (as measured by among-cluster mean squares) was equal to the heterogeneity within clusters (as measured by within cluster mean squares). When this probability is minimum, the chances of them being equal is least or the among-cluster varicance is most significant. This was the most "logical" compromise point for the maintenance of internal homogeneity and for maximising amongst-cluster differences. This stage was adopted as the most suitable cut off point in the clustering, and the clusters at this stage were adopted as the most relevant. This approach is examined in detail with the present data.

With the ALLCHARA attributes, it was found that the minimum probability occurred at stage 110, which resulted in 50 clusters. Thus all strategies were "cut off" at stage 110 in order to compare their clustering behaviours. These are considered in the following (see also Review, section 1.7.3.2.2).

TSS = Total Sum of Squares,
WSS = Within-cluster Sum of Squares,
ASS = Among-cluster Sum of Squares,
DFW = Degree of Freedom for WSS,
DFA = Degree of Freedom for ASS,
x = F-ratio,
. = Probability of F-ratio in $\log_{10}$ scale.

FIGURE 3.2 Changes in SS, and F-test for ALLCHARA, as examined by program SEFWIG.


TSS
WSS
ASS
DFA
DFW
F-ratiox10
log10 (prob)
TSS
ASS
DFA
DFW
WSS
Clustering Stage
0
20
40
60
80
100
120
140
160
180
200
10
30
50
70
90
110
130
150
160

3.2.2 Single Linkage

Of the 50 clusters obtained, 47 were single entity clusters (i.e. clusters with only one group), one cluster contained two groups, another contained three groups, and the remaining 108 groups were in one big cluster. This demonstrated the severe space-contraction, and the consequent chaining tendency (see reveiw). The dendrogram is shown in Fig 3.3.

3.2.3 Centroid Method

The results from program CONVER showed that, as clustering proceeded, the probabilities of the F-ratio (among-cluster mean squares to within-cluster mean squares) fluctuated. In view of the nature of this strategy (see review), this fluctuation was expected. Through minimum probability (0.5802×10^{-6}) occurred at stage 78, the clustering was "cut off" at stage 110 (with the probability of 0.2715×10^{-4}), for the reason already given.

Of the 50 clusters obtained, 40 were single entity clusters, eight clusters contained two groups each, one cluster contained three groups and the remaining 101 groups were in one big cluster.

Reversals occurred at various stages of clustering (at 30/159 stages). The most serious reversals were at stages 157 and 128, where the clustering criterion dropped from 2.876 to 2.411 and from 1.206 to 0.9699, respectively. The clustering criterion generally ranged from 0.106089 (stage 1) to 2.918606 (stage 159).

The results showed the severe space-contracting and non-monotonic properties of this method (see Review). These results agreed with those of the simplex test of Burr(1970). However, these did not agree with those of Lance and Williams(1967a). They regarded this as a space-conserving strategy. The dendrogram is shown in Figure 3.4.

3.2.4 Median Method

This method had similar results to those of the centroid method. Of the 50 clusters obtained, 42 were single entity clusters, six clusters contained two groups each, one cluster contained four

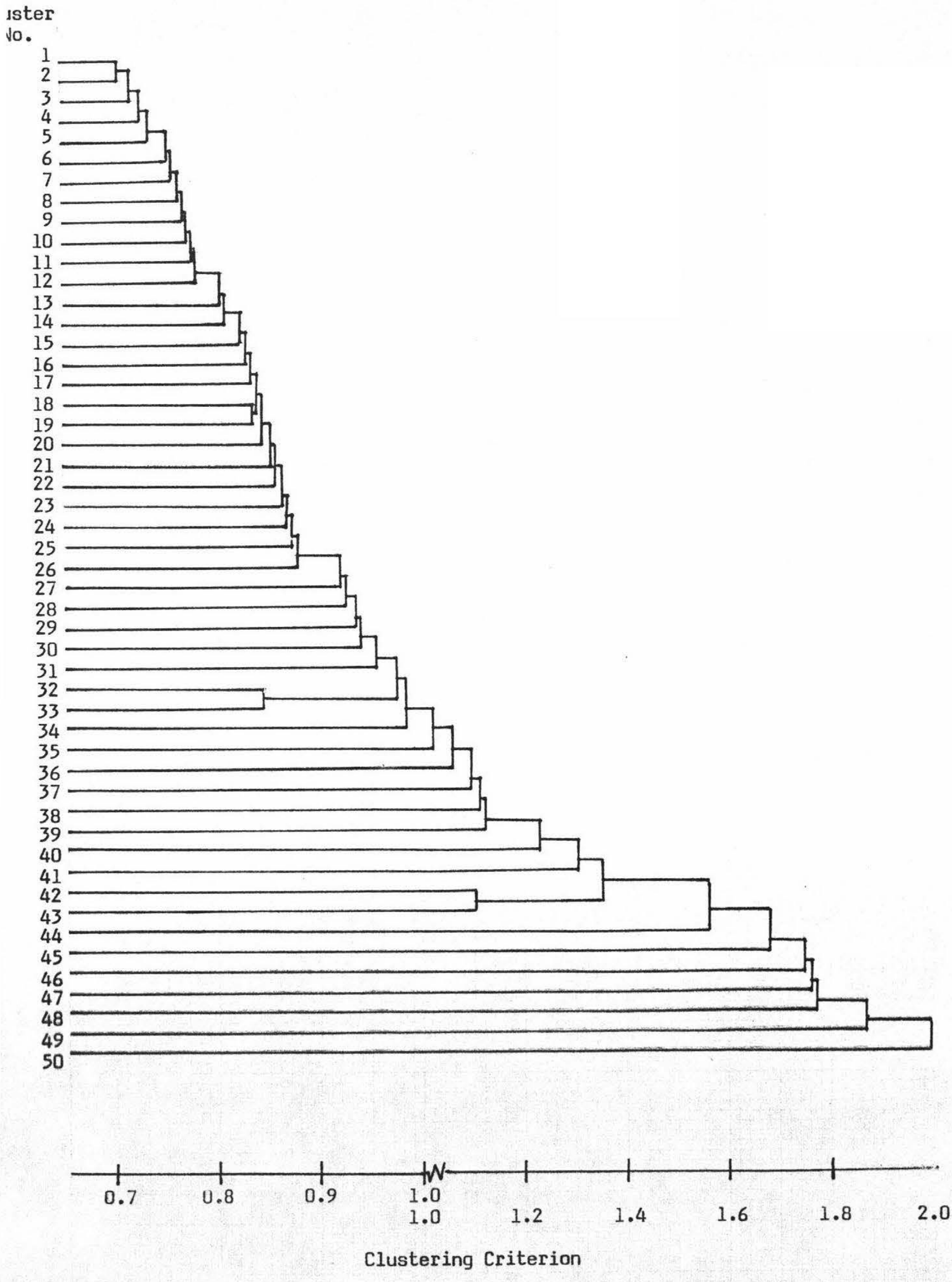


FIGURE 3.3 Dendrogram of ALLCHARA by Single Linkage Method.

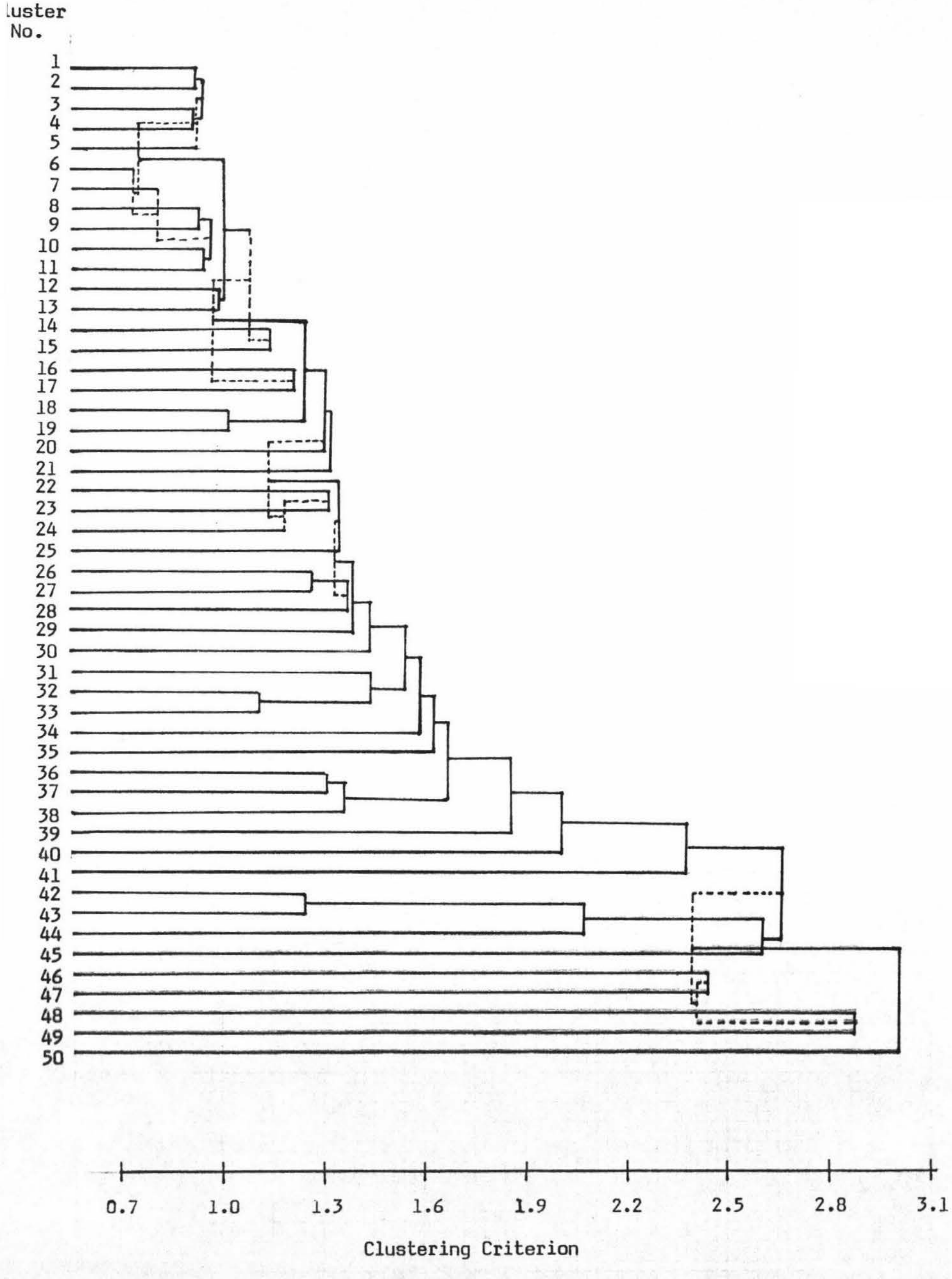


FIGURE 3.4 Dendrogram of ALLCHARA by Centroid Method, reversals are shown in dotted lines.

groups and the remaining 102 groups were in one big cluster.

Reversals also occurred at various stages of clustering. The dendrogram in Fig 3.5 shows space-contracting and non-monotonic properties of this method. The results did not agree with those of Lance Williams(1967a), who regarded this method as space-conserving and non-monotonic.

3.2.5 Average Linkage Between Merged Clusters

The results from program CONVER showed that, as clustering proceeded, the probabilities of F-ratio fluctuated. This fluctuation was expected because of the procedure (see Review). Though minimum probability (0.6337×10^{-8}) occurred at stage 101, the clustering was "cut off" at stage 110 (with the probability of 0.1160×10^{-7}). for uniformity in the comparsion.

Of the 50 clusters obtained, 21 were single entity clusters, 13 clusters contained 2 groups each, 6 clusters contained 3 groups each, 2 clusters contained 4 groups each, 3 clusters contained 5 groups each and the remaining 5 clusters contained 6, 8, 13, 20 and 25 groups each, respectively.

A very mild space contracting and chaining tendency was shown in the later stages of clustering. This occurred around the clustering criterion value of 2.0, as shown in Figure 3.6.

3.2.6 Average Linkage Within New-Cluster

Of the 50 clusters obtained, 20 of them were single entity cluster, 9 clusters contained 2 groups each, 4 clusters contained 3 groups each, 5 clusters contained 4 groups each, 5 clusters contained 5 groups each, 2 clusters contained 9 groups each and the remaining 5 clusters contained 6, 8, 10, 11, and 12 groups respectively.

This method showed a more even distribution of cluster size when compared to the previous one. Although it had 20 single-entity clusters, the largest cluster contained only 12 groups (as compared to 25 in Average Linkage Between Merged-Clusters). This showed a more intense clustering. The dendrogram is shown in Figure 3.7.

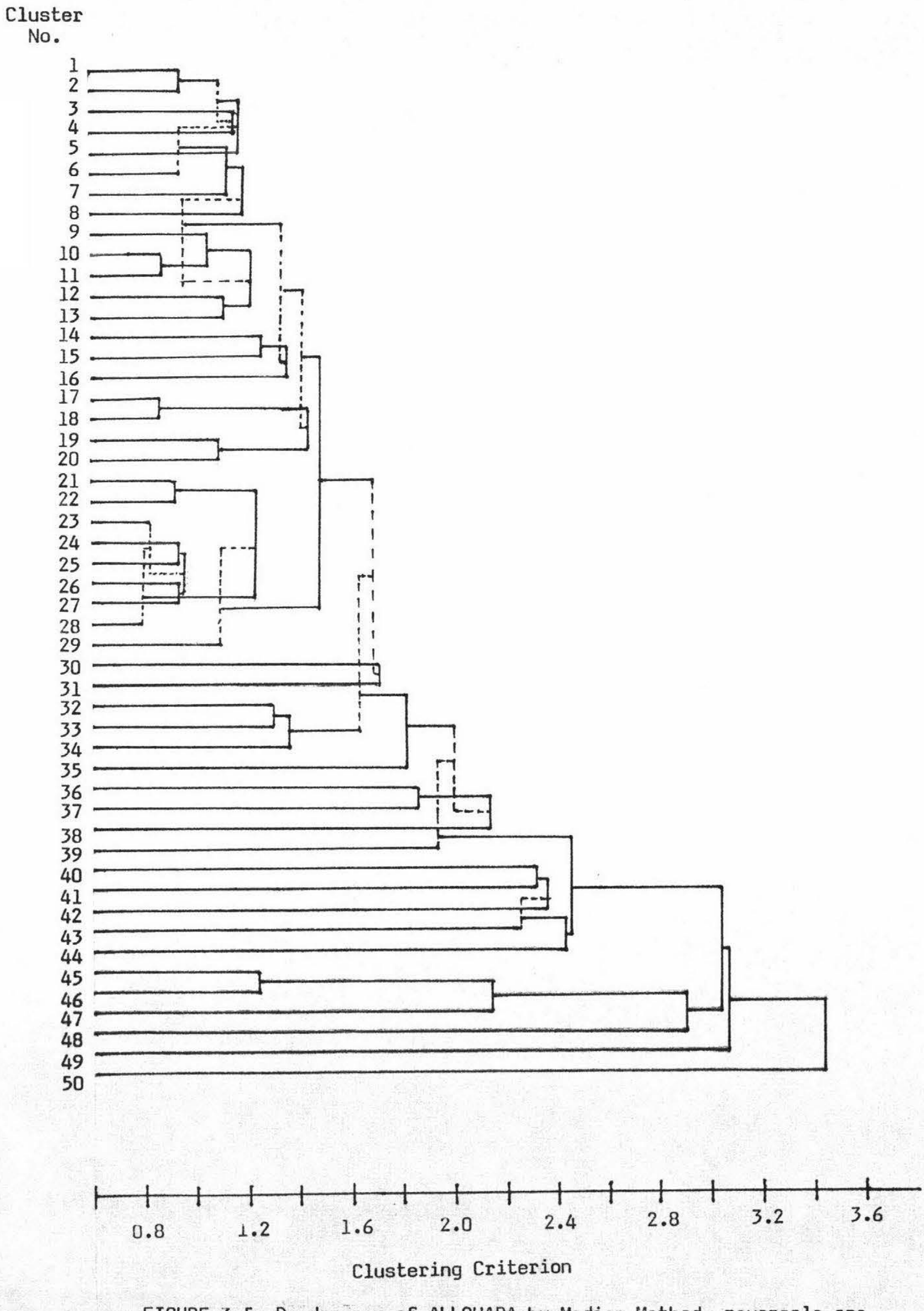


FIGURE 3.5 Dendrogram of ALLCHARA by Median Method, reversals are shown in dotted lines.

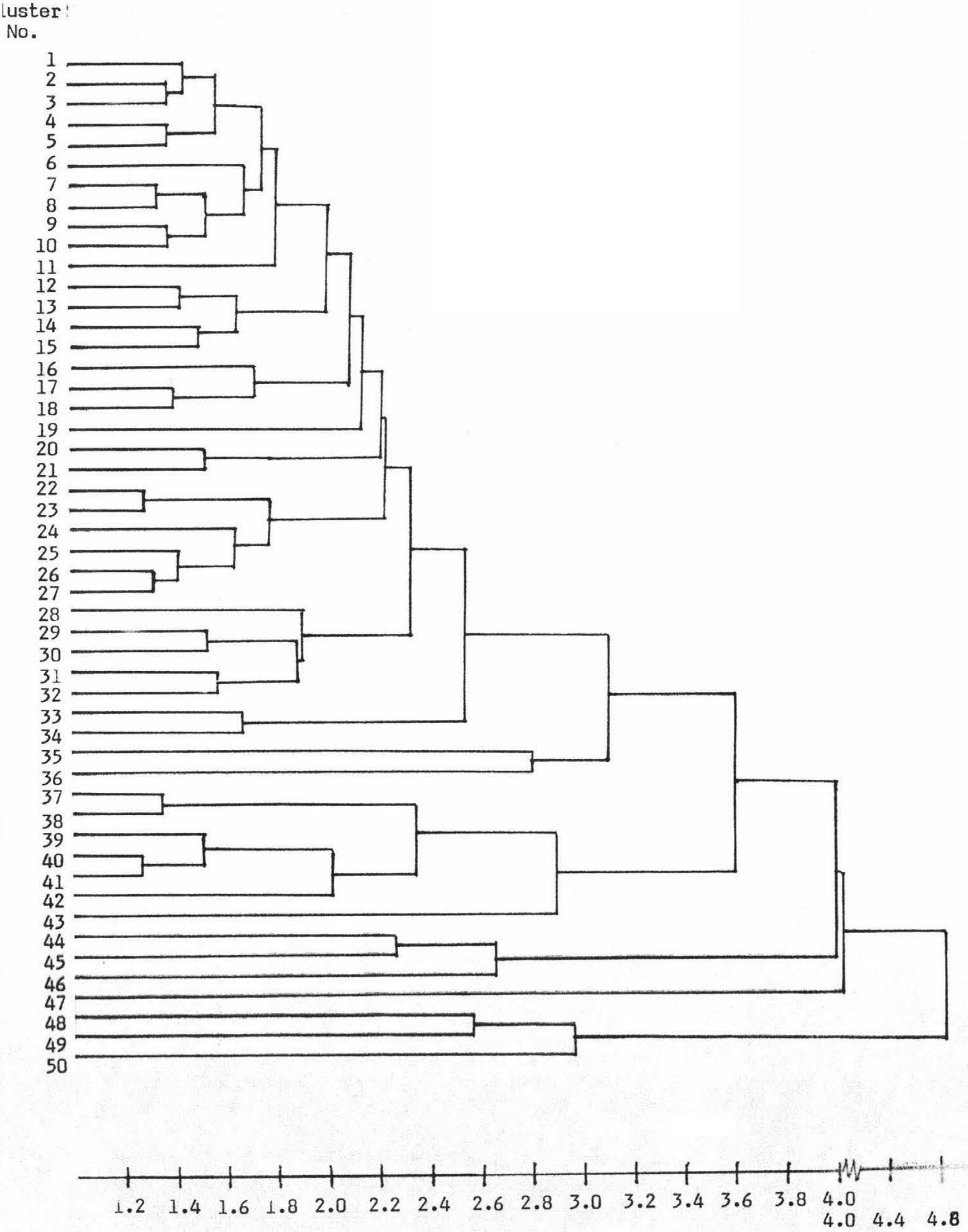


FIGURE 3.6 Dendrogram of ALLCHARA by Average Linkage Between Merged Clusters Method.

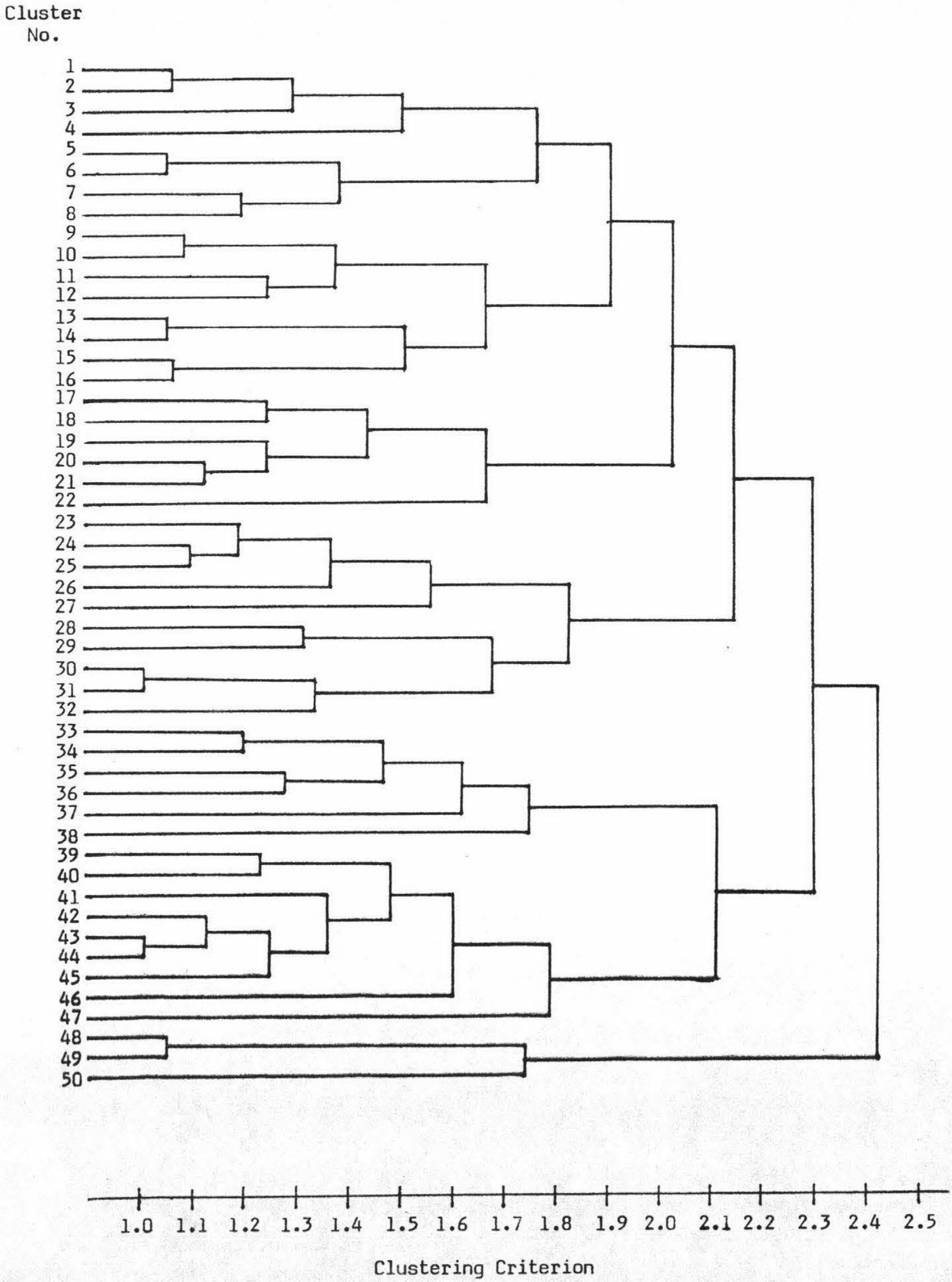


FIGURE 3.7 Dendrogram of ALLCHARA by Average Linkage Within New Cluster Method.

3.2.7 Complete Linkage

Of the 50 clusters obtained, 12 of them were single entity clusters. There were 10, 15, 2, 3, and 4 clusters containing 2, 3, 4, 5, and 6 groups each, respectively. The other 4 clusters each contained 7, 8, 9, and 12 groups, resectively. The dendrogram is shown in Figure 3.8.

3.2.8 Ward's Method

This was the most intense clustering strategy of the seven studied. It produced the most even distribution of cluster size, with the largest cluster containing 9 groups. Of the 50 clusters obtained, 9 of them were single entity clusters. There were 10, 12, 9, 5 and 3 clusters containing 2, 3, 4, 5 and 6 groups each, respectively. The remaining two clusters contained 7 and 9 groups each, respectively. The dendrogram is shown in Figure 3.9.

3.2.9 General Comparison

Single Linkage, Centroid, and Median methods were every "weak" clustering strategies, in that they produced heavily "chained" clusters. The chaining tendency of these methods produced one big cluster and many single entity clusters. The clusters obtained were of no practical use. Moreover, Centroid and Median methods showed the conceptual illogic of reversals. These findings agreed with those of Williams(1971, 1972), Lance and Williams(1967a, 1967b), Anderberg(1973), Cormack(1971), Clifford and Stephenson(1975), Boyce(1969), Burr(1970) and Pritchard and Anderson(1971). Thus, practically, these methods were of little use.

The Average Linkage Between Merged Clusters method produced 21 single entity clusters and mild chaining effect. It was not known whether this chaining tendency was due to the property of the method or to the weak structure of the data. It was suspected that the data was weakly stuctured, and, very probably this was a cause of the chaining. This method had been considered as a space-conserving strategy (Lance and Williams 1967a, 1967b), and a "most useful method" (Pritchard and Anderson 1971). However, its usefulness here was not obvious and it showed space-contraction.

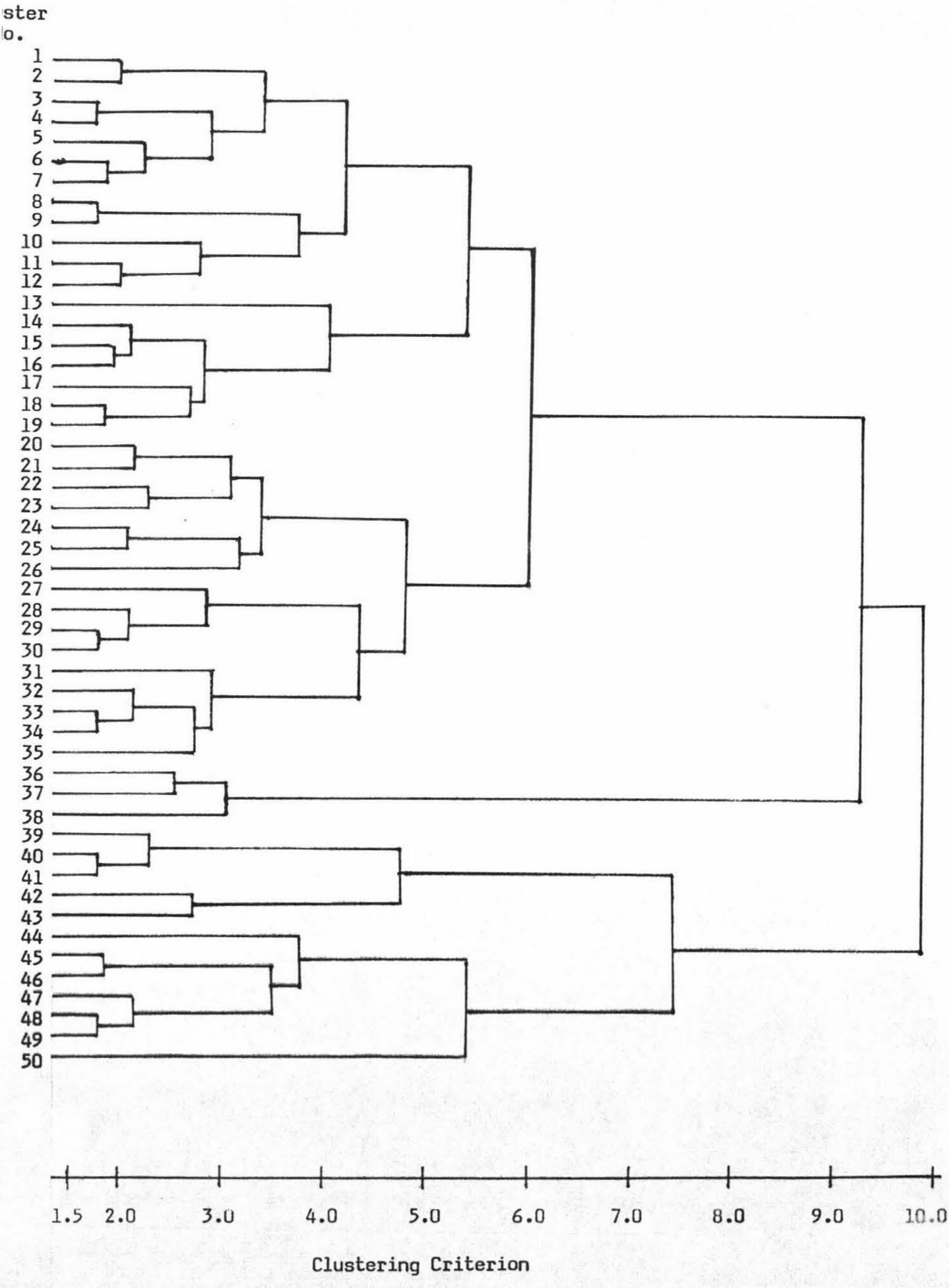


FIGURE 3.8 Dendrogram of ALLCHARA by Complete Linkage Method.

Cluster No.	Group No.
1	19, 66, 9, 39, 69.
2	140, 146, 139, 7, 126.
3	20, 94, 156.
4	76, 77, 38.
5	18, 145, 120, 122, 160.
6	31, 59, 8, 42.
7	28, 116, 128, 143, 82, 106.
8	89, 99, 35.
9	110, 135, 114, 118, 125.
10	25, 45, 44, 50.
11	72, 112, 73.
12	21, 127, 51.
13	49.
14	101, 107, 117, 130.
15	115, 119, 131, 85.
16	97, 121, 98.
17	36, 133, 6, 57.
18	14, 62, 41.
19	56.
20	15, 124.
21	102, 108.
22	86, 92.
23	26, 88, 65, 90.
24	70, 80, 27, 46.
25	23, 147, 87.
26	104, 123, 132, 48, 84, 109, 75.
27	16, 33, 17.
28	64, 152.
29	2, 47, 5, 136, 3, 37.
30	40, 58, 153.
31	43, 157, 111, 60, 52.
32	11, 24, 95, 96, 29, 141, 1, 71, 78.
33	67, 148, 74, 113, 83, 129.
34	63.
35	68.
36	81, 103, 105.
37	159.
38	10.
39	144.
40	53, 54.
41	30, 100.
42	91, 154.
43	22, 61.
44	138, 142, 134.
45	158.
46	137, 151.
47	34, 55, 93, 150.
48	79.
49	4, 13, 149, 155.
50	12, 32.

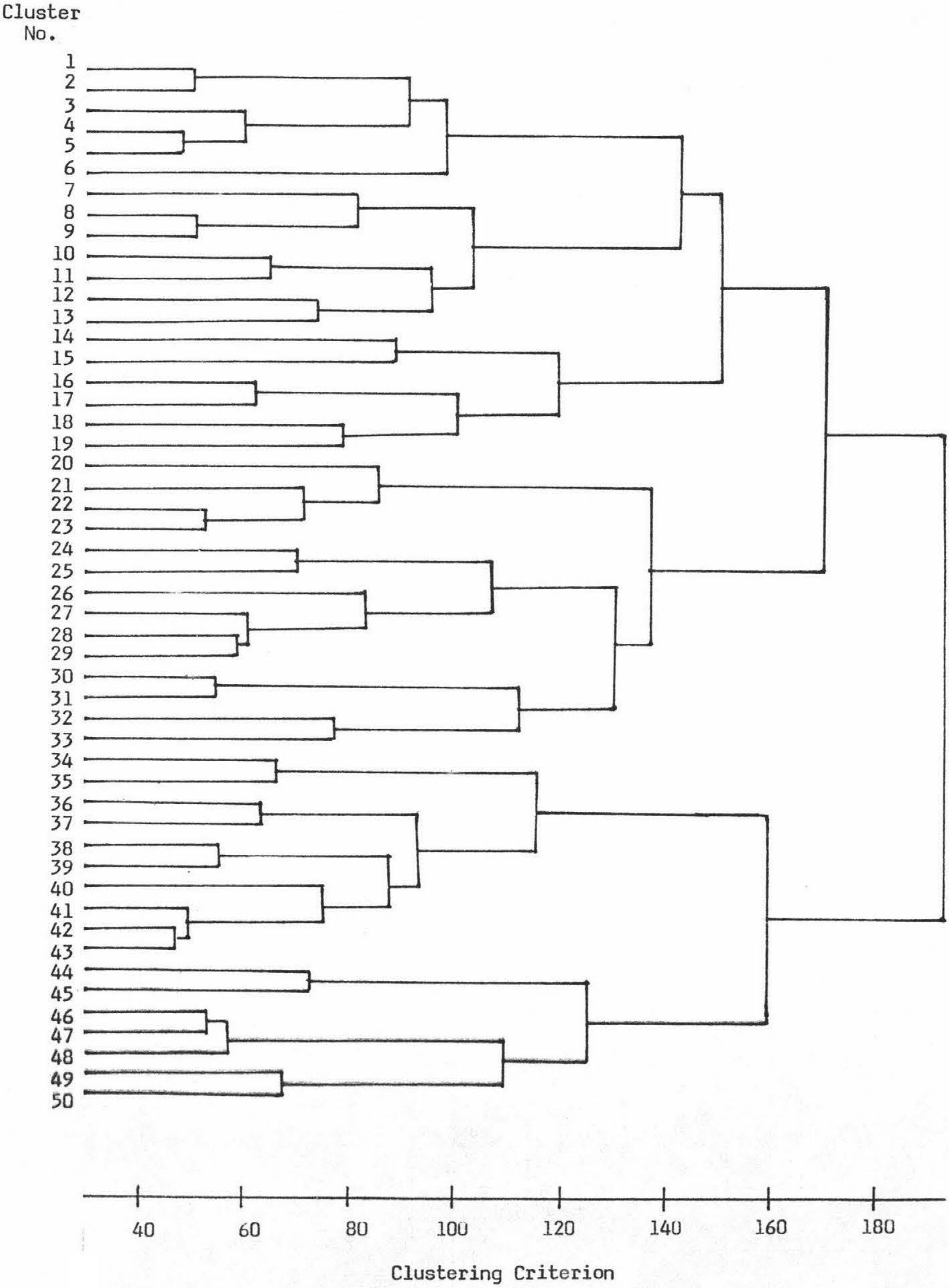


FIGURE 3.9 Dendrogram of ALLCHARA by Ward's Method.

The Average Linkage Within New Cluster and Complete Linkage methods were more intensely clustering strategies in that they produced more evenly sized clusters. The largest cluster contained only 12 groups. These two methods produced 20 identical clusters (clusters containing the same group or groups) out of the 50 clusters obtained. Furthermore, most of the other clusters amongst them were quite similar in their constituents. This showed that the two methods produced similar patterns of clustering. This finding concurred with that of Anderberg(1973).

Judging from the result of the "traditional" space-conserving strategy (Average Linkage Between Merged Clusters method), the set of data used in this study was weakly structured. That is, the groups were continuously spread out with no distinctive natural cluster. To facilitate description of such loosely defined clusters (with natural diffuse boundaries), Williams(1971) suggested the use of a space dilating, intensely clustering strategy to artificially "sharpen" the boundaries. Among the seven strategies studied, Ward's method was the most intense in its clustering. It's space dilating and intense clustering properties have been noted by Clifford and Stephenson(1975), Williams (1971, 1972), Lance and Williams(1967b), and Burr(1970). Hence, Ward's method was chosen as the principal clustering strategy of this study and has been used exclusively for other sets of attributes. Moreover, the logic of a "variability" approach (i.e. one based on the minimum increase of pooled within-cluster sums of squares) in weakly structured data, and the ready association with probabilistic "cut off" decisions (found in programmes SEFWIG and CONVER), confirmed the choice of Ward's method as the principal strategy for this data.

3.3 Clustering Analysis and Post Clustering Analysis

3.3.1 ALLCHARA

The clustering results of this set of attributes have been discussed briefly in section 3.2.8, and the dendrogram shown in Figure 3.9. There were nine single entity clusters. They were clusters 13, 19, 34, 35, 37, 38, 39, 45 and 48. A detailed investigation of these, with respect to the ranking of group means for each character separately, revealed that most of these groups had outlying values (i.e. they had at lease one extreme ranking (either first or last) in one of the characters). As SED will give extra weight to outlying values (Clifford and Stephenson 1975, and Cormack 1971), these groups became more remote from the others. This was the reason that they remained as single entity clusters, despite the intense clustering of Ward's method (Williams 1971, and Clifford and Stephenson 1975).

Table 3.18 shows the growth of the WSS/TSS ratio of each character as clustering proceeded. This ratio indicated the proportion of the total sums of squares of each character not explained by cluster differences. It is the complement of eta-square, and it is similar to the complement of the coefficient of multiple determination.

As the clustering was "cut off" at stage 110. Table 3.18 can be studied in two parts. The first part included stages 1 to 110, and described the changes of intra-cluster structure (pooled over all clusters) of each character. The second part included stages 111 to 159, and described the subsequent structure, when further merges occurred between the "accepted" 50 clusters. In this case, the relationships amongst the 50 clusters can be examined.

At stage 110, C.ERE was well partitioned across the 50 clusters; only 11.1% of the TSS was not accounted for by the among-cluster sums of squares. F.DAT was also well partitioned across clusters with only 13.2% of TSS not accounted for by clustering. The other characters were ranked in increasing order of proportion of TSS not accounted for by clustering, as follows. O.DIS(19.8%), RUST(22.5%), L.WID(25.9%), C.DIA(27.0%), L.ROL(28.1%), C.HEI(30.1%), F.COL(30.1%), C.DEN(30.2%) and L.COL(33.0%). These results implied that C.ERE and F.DAT were the two most dominant characters with more than 85% of their variation accounted for by differences amongst clusters. Conversely, C.HEI, F.COL, C.DEN and L.COL were the most dormant characters, with more than 30% of the variation not accounted for by clustering. This proportion (of 30%) due to within cluster variation still indicated, however, that a substantial proportion of the variation was due to cluster differences. In other words these character were only relatively dormant, but still contributed considerably to cluster structure.

Stage	Criterion	C.DIA	C.DEN	C.ERE	C.HEI	RUST	O.DIS	L.ROL	L.COL	L.WID	F.COL	F.DAT
1	0.053	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	1.226	0.009	0.006	0.002	0.007	0.008	0.013	0.006	0.008	0.005	0.007	0.004
20	3.065	0.015	0.015	0.009	0.017	0.018	0.026	0.024	0.017	0.014	0.014	0.007
30	5.340	0.031	0.028	0.014	0.023	0.033	0.038	0.041	0.026	0.034	0.024	0.012
40	8.107	0.046	0.046	0.021	0.039	0.049	0.048	0.051	0.061	0.048	0.028	0.021
50	11.391	0.071	0.090	0.026	0.054	0.056	0.060	0.089	0.085	0.066	0.039	0.029
60	15.069	0.088	0.102	0.042	0.075	0.076	0.077	0.104	0.128	0.088	0.054	0.048
70	19.299	0.114	0.145	0.055	0.106	0.106	0.096	0.136	0.171	0.102	0.133	0.059
80	24.293	0.127	0.201	0.072	0.146	0.130	0.125	0.148	0.195	0.130	0.144	0.071
90	30.179	0.169	0.236	0.085	0.183	0.174	0.153	0.195	0.216	0.177	0.212	0.088
100	37.399	0.197	0.274	0.104	0.227	0.211	0.175	0.206	0.276	0.218	0.268	0.101
110	46.010	0.270	0.302	0.111	0.301	0.225	0.198	0.281	0.330	0.259	0.301	0.132
120	56.957	0.455	0.349	0.147	0.338	0.300	0.247	0.299	0.371	0.307	0.410	0.156
130	70.789	0.569	0.401	0.191	0.405	0.389	0.331	0.347	0.422	0.420	0.503	0.211
140	88.732	0.696	0.610	0.268	0.493	0.510	0.391	0.486	0.619	0.538	0.593	0.238
150	115.845	0.854	0.841	0.309	0.616	0.689	0.573	0.665	0.718	0.721	0.719	0.337
159	192.560	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

TABLE 3.18 The Proportion Of Sum Of Squares (WSS/TSS) Not Explained By Clustering, At Different Stages Of Clustering By Ward's Method for ALLCHARA.

The ranks of the cluster means for each character are shown in Table 3.19. The notable feature was the outstanding ranking of cluster 35. It ranked 1st in C.DIA, 3rd in C.DEN, 3rd in C.ERE, 2nd in C.HEI, 2nd in RUST (i.e. high rust resistance), 5th in O.DIS (i.e. good overall disease resistance), last in L.ROL (i.e. flat leaf), last in L.COL (i.e. green leaf tip), 1st in L.WID (i.e. widest leaf) and 9th in F.COL. Clusters 34, 38, 39, 45 and 16 also had extreme ranking (good or bad) in some characters. Of these extremely ranked clusters, all except cluster 16 were single entity clusters.

With respect to an ecotype consideration, an extrinsically intrinsic study (Williams 1971, see section 1.7.2) has been carried out briefly also. This involved the examination of relationships amongst clusters' constituents and their respective external attribute (such as location, altitude and habitat from which the groups were collected, as listed in Appendix B-1). The object was to find whether the boundaries between clusters reflected any discontinuity in external attributes. The results failed to show any clear cut patterns. This implied that, as far as this analysis could reveal, there were no true ecotypes in the collection.

After stage 110, the growth of the WSS/TSS ratio for C.ERE and F.DAT remained slow. They were still the lowest even at stage 158. This indicated that they continued to be dominant. The growth of this ratio for C.HEI and L.ROL was also slow after stage 110, such that they became the 4th and 5th most dominant characters at stage 150. C.DIA and C.DEN remained as the most dormant characters. These results implied that further merging was influenced mainly by C.ERE and F.DAT. That is, the relationships amongst the 50 clusters at stage 110 can be expressed mostly by differences in C.ERE, and F.DAT; and least effectively by differences in C.DIA and C.DEN.

The clustering strategy used was polythetic, where similarity was based on all the characters as a whole (Williams 1971). Therefore, it was not appropriate to set up an identifying key (based on the 11 characters) for these 50 clusters. This can be done by a monothetic strategy (Williams 1971). However, the relationships amongst the 50 clusters can be described approximately by the dominant characters. These are presented briefly in Table 3.20. From this information, a descriptive partitioning could be devised similar to that

RANK	C.DIA	C.DEN	C.ERE	C.HEI	RUST	O.DIS	L.ROL	L.COL	L.WID	F.COL	F.DAT
1	35	39	38	39	8	34	17	19	35	20	21
2	5	34	39	35	35	11	3	16	13	15	22
3	6	35	35	50	40	12	16	20	44	25	25
4	22	23	37	42	11	50	15	29	41	12	14
5	43	22	45	37	4	35	12	28	6	8	23
6	34	33	34	38	38	10	7	12	36	22	15
7	31	3	42	43	13	14	28	23	43	13	28
8	14	49	50	24	6	8	23	27	37	2	26
9	12	50	43	49	18	18	40	39	1	35	36
10	49	21	41	3	12	48	14	13	11	16	12
11	17	7	46	36	44	7	5	18	9	18	8
12	10	6	36	25	48	40	4	17	27	14	20
13	11	36	30	27	46	13	8	40	10	34	4
14	3	16	44	41	7	16	25	31	45	40	7
15	7	20	40	34	9	38	18	33	7	21	13
16	1	46	49	19	25	15	22	22	30	23	11
17	18	26	3	11	14	17	21	32	31	36	24
18	29	30	48	46	41	27	6	25	49	7	9
19	45	17	13	5	10	43	29	24	38	26	29
20	13	31	7	26	16	33	2	5	2	9	2
21	46	8	33	29	47	46	33	43	24	41	16
22	33	45	25	45	34	47	50	41	17	49	33
23	2	15	31	1	22	49	26	49	12	3	34
24	19	29	5	33	5	4	43	48	8	10	30
25	20	5	32	30	23	32	49	30	25	24	5
26	4	43	19	32	1	9	1	9	42	5	43
27	24	14	6	14	17	26	37	36	20	19	3
28	26	2	47	47	3	36	13	14	39	29	32
29	44	11	1	31	39	29	32	8	23	32	1
30	23	10	11	4	15	6	27	37	3	43	17
31	50	9	23	13	43	25	9	10	15	33	40
32	9	1	9	2	32	23	30	6	28	1	27
33	42	25	4	6	27	44	31	42	29	4	10
34	41	12	24	40	33	39	11	15	47	11	37
35	16	37	27	20	42	37	24	26	14	47	35
36	30	4	20	18	36	28	46	3	5	17	31
37	32	42	26	44	50	22	42	2	18	38	41
38	50	27	10	7	49	20	36	47	50	44	49
39	36	28	8	9	29	19	10	21	26	31	18
40	47	47	2	21	24	42	47	11	4	48	38
41	39	41	22	23	2	5	41	38	46	6	6
42	21	48	29	8	20	30	20	44	22	27	42
43	25	24	12	17	26	2	48	4	34	45	39
44	27	40	21	28	28	41	34	50	21	37	50
45	38	32	17	48	21	1	19	1	32	42	47
46	8	38	18	10	19	31	44	7	33	28	19
47	40	18	28	22	30	3	45	45	40	30	46
48	28	44	15	15	31	24	39	34	16	50	48
49	37	19	14	16	45	45	38	46	19	39	45
50	48	13	16	12	37	21	35	35	48	46	44

TABLE 3.19 The Ranks Of The Means Of Each Character
Of 50 Clusters In ALLCHARA Analysis.

	Clusters									
Characters	1-6	7-13	14-15	16-19	20-23	24-29	30-33	34-43	44-45	46-50
C.ERE	M	M	L	L	M-	M-	M	H	H	H
F.DAT	M	M	H	M	H	M+	M	M	L	L
RUST	M	H	M	M	M-	M-	M-	S	S	M

TABLE 3.20 Brief Grouping Of The 50 Clusters Of ALLCHARA, And Their Approximate Average Ranking In Dominant Characters.
H = High, M+ = Medium High, M = Medium,
M- = Medium Low, L = Low and S = Spread Out.

achieved by a formal key. However, the separation is not always very distinct.

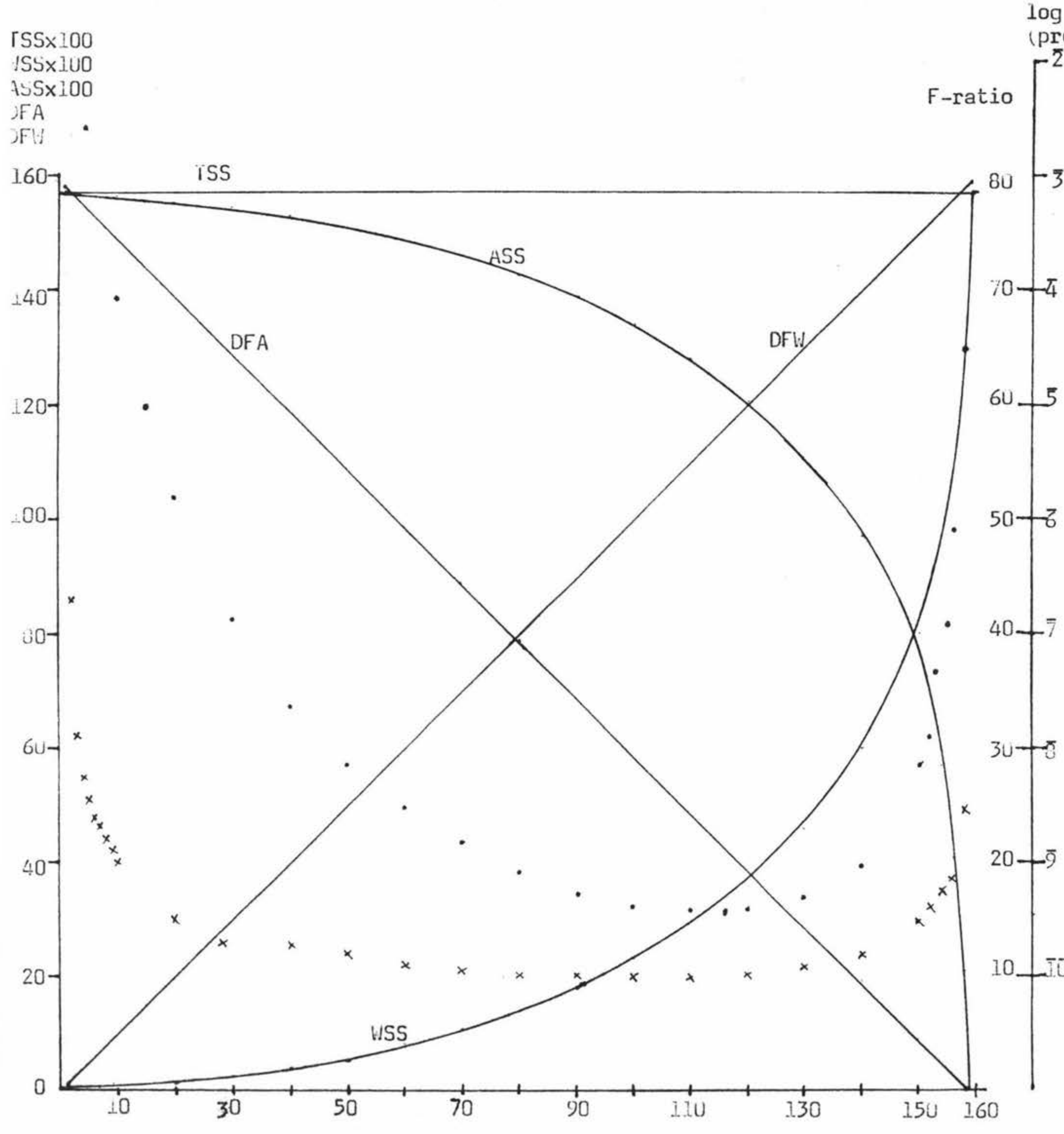
3.3.2 AGROCHARA

After clustering by Ward's method, SEFWIG was used to decide the "cut off" point for the clusters of the agronomic attributes. As shown in Figure 3.10, the TSS remained constant at 1.5699, and ASS and df_A decreased while WSS and df_W increased, as clustering proceeded. The F-ratio dropped from 71.4210 (at stage 1) to a minimum value of 9.90539 (at stage 103), then it fluctuated between 9.90660 and 9.92109. It then increased from 9.90713 (at stage 109) to 24.4873 (at stage 158). The associated probability dropped from 0.098197 (at stage 1) to a minimum value of 0.37027×10^{-9} (at stage 116), and then increased to 0.30295×10^{-4} (at stage 158). These trends were similar to those of ALLCHARA (in section 3.2.1). Thus, as discussed earlier, it was decided to "cut off" the clustering at stage 116 where the probability of the overall F-ratio was lowest. This resulted in 44 clusters.

Of these 44 clusters, 8 were single entity clusters (cluster nos. 3, 16, 31, 33, 34, 39, 40 and 44). There were 10, 6, 6, 5, 3, 4, 1 and 1 clusters each containing 2, 3, 4, 5, 6, 7, 8 and 11 groups respectively. The dendrogram is shown in Figure 3.11.

Table 3.21 shows the proportion of sums of squares (WSS/TSS) not explained by clustering. At stage 116, F.DAT was well partitioned amongst clusters, with only 12.3% of TSS not explained by clustering, and the other 87.7% accounted for by the among-cluster sum of squares. C.ERE was also well partitioned with only 15.4% of TSS not accounted for by clustering. The other characters were ranked as follows for this property: C.HEI(20.9%), L.WID(24.4%), C.DIA(25.7%), RUST(26.4%), C.DEN(30.3%) and O.DIS(31.0%). These results implied that F.DAT and C.ERE were the most dominant characters, whereas C.DEN and O.DIS were the most dormant characters. However, as stated before, this dormancy was not absolute.

The ranks of the cluster means for each character are shown in Table 3.22. The consistenly good rankings of cluster 40 were noted particularly. It ranked 1st in C.DIA, RUST and L.WID, 2nd in C.DEN,



TSS = Total Sum of Squares,
WSS = Within-cluster Sum of Squares,
ASS = Among-cluster Sum of Squares,
DFW = Degree of Freedom for WSS,
DFA = Degree of Freedom for ASS,
× = F-ratio,
• = Probability of F-ratio in $\log_{10}$ scale.

FIGURE 5.10 Changes in SS, and F-test for AGROCHARA, as examined by program SEFWIG.

Cluster No.	Group No.
1	29, 141, 38, 71, 95, 96, 40, 148.
2	11, 24, 9, 93.
3	54.
4	14, 41, 62.
5	1, 78, 55, 3, 56.
6	6, 57, 111, 52.
7	7, 69, 43, 157, 60.
8	42, 133, 31.
9	39, 120, 59, 34, 149, 8.
10	76, 77, 160.
11	18, 145, 19, 66, 20, 94, 122.
12	110, 135, 89, 99, 35.
13	36, 126, 5, 16, 33.
14	65, 83, 121, 131, 97, 98.
15	72, 112.
16	49.
17	25, 45, 21, 44, 50, 107.
18	51, 127.
19	114, 118, 125, 143, 128, 73, 85.
20	102, 108.
21	87, 88.
22	86, 92.
23	116, 153,
24	28, 106, 90, 67, 129, 15, 26.
25	27, 46, 17.
26	70, 80, 23, 147.
27	37, 101, 117, 130.
28	104, 132, 124, 146, 2, 140, 47, 123, 136, 84, 139.
29	109, 119, 115, 48, 75, 64, 152.
30	137, 151, 150.
31	79.
32	138, 142.
33	134.
34	158,
35	4, 13, 74, 113, 32.
36	91, 154, 155, 156.
37	12, 144.
38	58, 159.
39	63.
40	68.
41	81, 103, 105.
42	61, 82, 53, 22.
43	30, 100.
44	10.

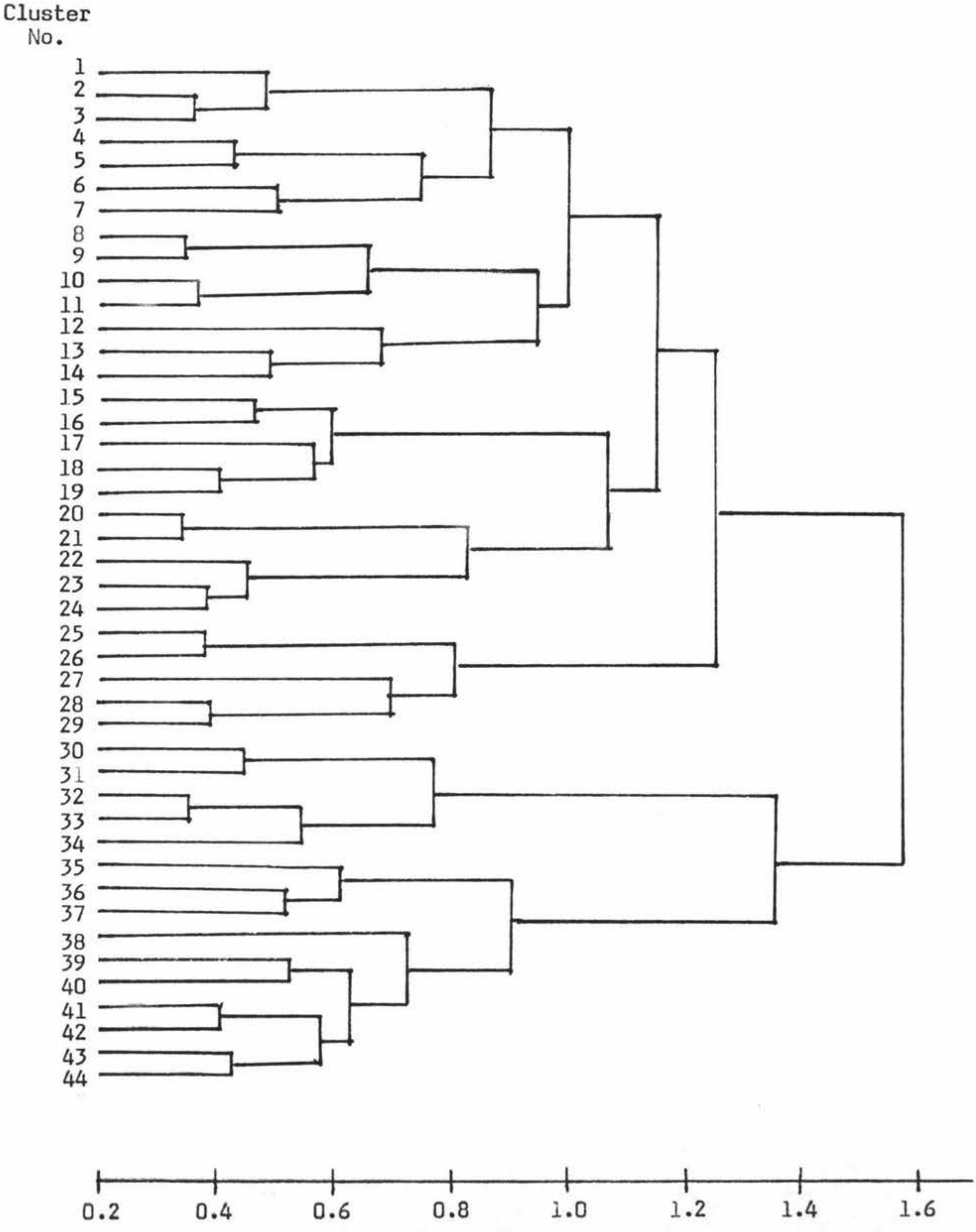


FIGURE 3.11 Dendrogram of AGROCHARA by Ward's Method.

Stage	Creter-ion	C.DIA	C.DEN	C.ERE	C.HEI	RUST	O.DIS	L.WID	F.DAT
1	0.0001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.0052	0.008	0.005	0.003	0.005	0.003	0.005	0.003	0.001
20	0.0145	0.015	0.018	0.005	0.009	0.008	0.012	0.009	0.006
30	0.0260	0.022	0.026	0.008	0.014	0.020	0.026	0.019	0.012
40	0.0398	0.035	0.040	0.016	0.024	0.028	0.042	0.027	0.013
50	0.0581	0.050	0.055	0.026	0.035	0.044	0.055	0.042	0.020
60	0.0807	0.072	0.079	0.038	0.055	0.062	0.076	0.057	0.031
70	0.1090	0.091	0.111	0.045	0.077	0.083	0.105	0.076	0.042
80	0.1421	0.112	0.140	0.058	0.105	0.100	0.127	0.106	0.052
90	0.1808	0.139	0.167	0.074	0.134	0.137	0.162	0.153	0.070
100	0.2290	0.178	0.192	0.090	0.150	0.203	0.221	0.196	0.086
110	0.2899	0.246	0.275	0.112	0.172	0.226	0.267	0.226	0.114
116	0.3323	0.257	0.303	0.154	0.209	0.264	0.310	0.244	0.123
120	0.3639	0.301	0.322	0.167	0.220	0.308	0.327	0.262	0.127
130	0.4603	0.446	0.390	0.227	0.270	0.395	0.405	0.334	0.153
140	0.5984	0.488	0.547	0.284	0.370	0.493	0.509	0.507	0.208
150	0.8329	0.665	0.648	0.379	0.492	0.702	0.737	0.700	0.301
159	1.5700	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

TABLE 3.21 The Proportion Of Sum Of Squares ($\frac{WSS}{TSS}$) Not Explained By Clustering, At Different Stages of Clustering By Ward's Method For AGROCHARA.

RANK	C.DIA	C.DEN	C.ERE	C.HEI	RUST	O.DIA	L.WID	F.DAT
1	40	39	44	37	40	39	40	21
2	33	40	40	40	3	15	16	20
3	15	24	37	36	15	18	33	22
4	22	22	34	44	12	17	32	27
5	8	37	39	26	33	40	43	18
6	39	35	38	38	8	27	19	29
7	9	14	36	25	44	4	38	41
8	18	8	43	35	16	35	41	28
9	11	20	41	41	10	42	8	26
10	27	41	42	15	18	31	15	24
11	7	36	33	42	4	37	7	19
12	6	21	30	11	31	19	17	12
13	17	11	32	43	42	16	9	14
14	35	30	35	5	30	12	34	10
15	10	23	23	39	32	44	23	16
16	4	42	3	30	17	30	44	23
17	42	12	31	1	43	2	26	39
18	34	34	1	7	19	24	13	15
19	16	7	16	9	21	14	12	25
20	19	9	11	28	9	8	42	11
21	24	19	7	27	14	13	25	42
22	36	13	9	34	2	23	11	13
23	28	28	26	16	39	29	36	1
24	5	26	21	21	22	41	37	38
25	23	6	25	33	26	6	21	17
26	30	15	15	13	13	3	18	7
27	26	27	5	10	37	1	28	35
28	14	38	24	29	11	9	6	6
29	43	29	12	4	27	10	2	2
30	13	18	2	3	24	32	27	3
31	1	1	19	12	23	28	35	40
32	41	10	28	8	1	5	24	8
33	25	17	10	20	41	33	4	43
34	20	5	8	32	35	38	29	36
35	2	43	6	19	29	22	5	4
36	21	31	17	14	36	26	30	44
37	37	44	22	6	5	21	22	5
38	29	4	13	17	28	25	39	9
39	38	33	20	31	25	36	20	37
40	44	3	18	2	7	11	10	30
41	32	25	4	24	20	43	14	31
42	12	32	29	22	6	7	1	34
43	3	2	14	23	34	34	31	33
44	31	16	27	18	38	20	3	32

TABLE 3.22 The Rank Of The Means Of Each Character Of 44 Cluster In AGROCHARA Analysis

C.ERE, C.HEI and 5th in O.DIS. As the scales ascended with increasing agronomic desirability, this cluster was outstanding. Clusters 37, 39, 44, 34, 31 and 33 also had extreme ranking (first or last) in some of the characters. Of these extremely ranked clusters, all except cluster 37 were single entity clusters.

The extrinsically intrinsic study did not reveal any clear cut pattern again, which implied that as far as the agronomic characters of this study were concerned, there were no ecotypic trends.

Although AGROCHARA was based on the 1st, 2nd, 3rd, 4th, 5th, 6th, 8th and 10th most dominant characters of ALLCHARA (at stage 110), the constituents of the 44 clusters obtained were different generally to those of the 50 clusters of ALLCHARA. It was found that clusters 13, 18, 21, 22, 34, 35, 36, 38, 41, 45 and 48 of ALLCHARA had the same constituents as clusters 16, 4, 20, 22, 39, 40, 41, 44, 43, 34 and 31 of AGROCHARA, respectively. However out of these 11 clusters, 6 of them were single entity clusters, 3 of them contained only 2 groups and another 2 of them contained only 3 groups. This showed that only a minor portion of the groups (18 out of 160) remained in the same clusters under these two analyses based on different sets of characters. This futher suggested that although the three characters (L.ROL, L.COL, and F.DAT) were comparatively dormant (in ALLCHARA), they contributed significantly to the overall similarity amongst groups. That is, as noted from the WSS/TSS ratios, their dormancy was not absolute.

After stage 116, the growth of WSS/TSS ratio for F.DAT, C.ERE and C.HEI remained slow changing. They were still the lowest even at stage 150 with 30.1%, 37.9% and 49.2% respectively. O.DIS remained as the most dormant character. The relationships amongst the 44 clusters can be described briefly through the dominant characters. These have been presented briefly in Table 3.23.

A comparison of the constituents of the "big" clusters (i.e. cluster amalgamations in Table 3.23) with those of ALLCHARA, revealed that the constituents of clusters 8-12 and cluster 19 were similar (but not identical) to those of clusters 1-9 of ALLCHARA. The constituents of clusters 25-29 were similar to those of clusters 24-29 of ALLCHARA. The constituents of clusters 30-44 were similar to those of 34-50 of ALLCHARA. The similarity was especially high in the latter

	Clusters						
Character	1-7	8-14	15-19	20-24	25-29	30-34	35-44
F.DAT	M-	M	M+	H	H	L	M
C.ERE	M	M	M	M	M-	H	H
C.HEI	M	M	M-	L	M+	H	M

TABLE 3.23 Brief Grouping Of The 44 Clusters Of AGROCHARA And Their Approximate Average Ranking In Dominant Characters. H = Hight, M+ = Medium High, M = Medium, M-:Medium Low and L = Low.

case, with 7 out of 15 clusters of AGROCHARA being same as the 7 out of 17 clusters of ALLCHARA. If these "big" clusters (clusters 30-44 of AGROCHARA and clusters 34-50 of ALLCHARA) were compared as a whole, there were 28 out of 32 groups of AGROCHARA in the same meld as the 28 out of 33 groups of ALLCHARA.

3.3.3 DISCCHARA

This set of characters was chosen because they (C.ERE, C.HEI, RUST, F.COL and F.DAT) were the five most discriminant characters amongst groups (i.e. they were the five with the largest eta values in MANOVA). The object in chosing them was to define the original structure in the collection with fewer characters than the total 11.

The value of TSS was 1.10505 for this analysis. The overall F-ratio changed from 60.5258 (stage 1) to a minimum value of 18.9742 (stage 99), and then fluctuated between 18.9825 and 19.2285. Finally, F increased from 19.1913 (stage 124) to 36.8662 (stage 158). The associated probability dropped from 0.10671 (stage 1) to a minimum value of 0.11831×10^{-10} (stage 114) and then increased to 0.33069×10^{-5} (stage 158). These trends were similar to those of ALLCHARA (section 3.3.1) and AGROCHARA (section 3.3.2). Using the previous criterion of minimum probability, the clustering cut off was at stage 114, which resulted in 46 clusters.

Of these 46 clusters, 8 were single entity clusters, which were nos. 5, 13, 25, 34, 36, 38, 39 and 42. There were 11, 8, 8, 4, 3, 2, 1 and 1 clusters each containing 2, 3, 4, 5, 6, 7, 8 and 14 groups, respectively. The dendrogram is shown in Figure 3.12.

Table 3.2 4 shows the proportion of sums of squares (WSS/TSS) not explained by clustering. At stage 114, C.ERE was well partition amongst clusters, with only. 7.5% of TSS not explained by clustering, and the other 92.5% accounted for by the among cluster sums of squares. F.DAT was also well partitioned amongst clusters with only 7.7% of TSS not accounted for by clusters. The other characters were ranked as follows, for this property: C.HEI(11.7%), F.COL(13.0%) and RUST(19.7%). These implied that C.ERE and F.DAT were the two most dominant characters, whereas F.COL and RUST were the two most dormant characters. However, as before, their dormancy was not absolute.

Cluster No. | Group No.

1 --141, 145, 95, 18, 74, 29, 4, 19, 122, 71, 160, 94, 120, 66.
2 -- 1, 149, 13, 56.
3 -- 70, 80.
4 -- 17, 27, 3.
5 -- 72.
6 -- 34, 42, 59, 55, 14, 31.
7 -- 96, 148, 25, 45, 39, 93.
8 -- 11, 44, 118, 9, 50, 6, 57.
9 -- 52, 64.
10 -- 69, 78, 111, 157, 60.
11 --129, 153.
12 -- 40, 43, 58.
13 --159.
14 -- 24, 67, 143, 116.
15 -- 26, 92, 28, 114, 51, 90.
16 -- 82, 110, 73, 106.
17 -- 76, 112, 38, 135, 77.
18 -- 25, 54.
19 -- 89, 127, 99.
20 -- 37, 115, 119.
21 --117, 130, 101.
22 -- 16, 33, 5, 133.
23 -- 83, 109, 48, 136, 2, 47, 75.
24 -- 36, 126, 107, 121, 98.
25 -- 62.
26 -- 15, 124, 139, 140.
27 -- 85, 131.
28 -- 7, 97, 21, 41.
29 -- 23, 147.
30 --103, 113, 81.
31 -- 65, 128, 20, 49, 125.
32 -- 46, 123, 84, 104, 108, 152, 146, 132,
33 -- 86, 87, 88.
34 --102.
35 -- 8, 137.
36 --151.
37 --134, 142, 79, 138.
38 --150.
39 --158.
40 --154, 156, 32, 155.
41 -- 12, 91.
42 --144.
43 -- 10, 68.
44 -- 63, 105.
45 -- 30, 53.
46 -- 22, 61, 100.

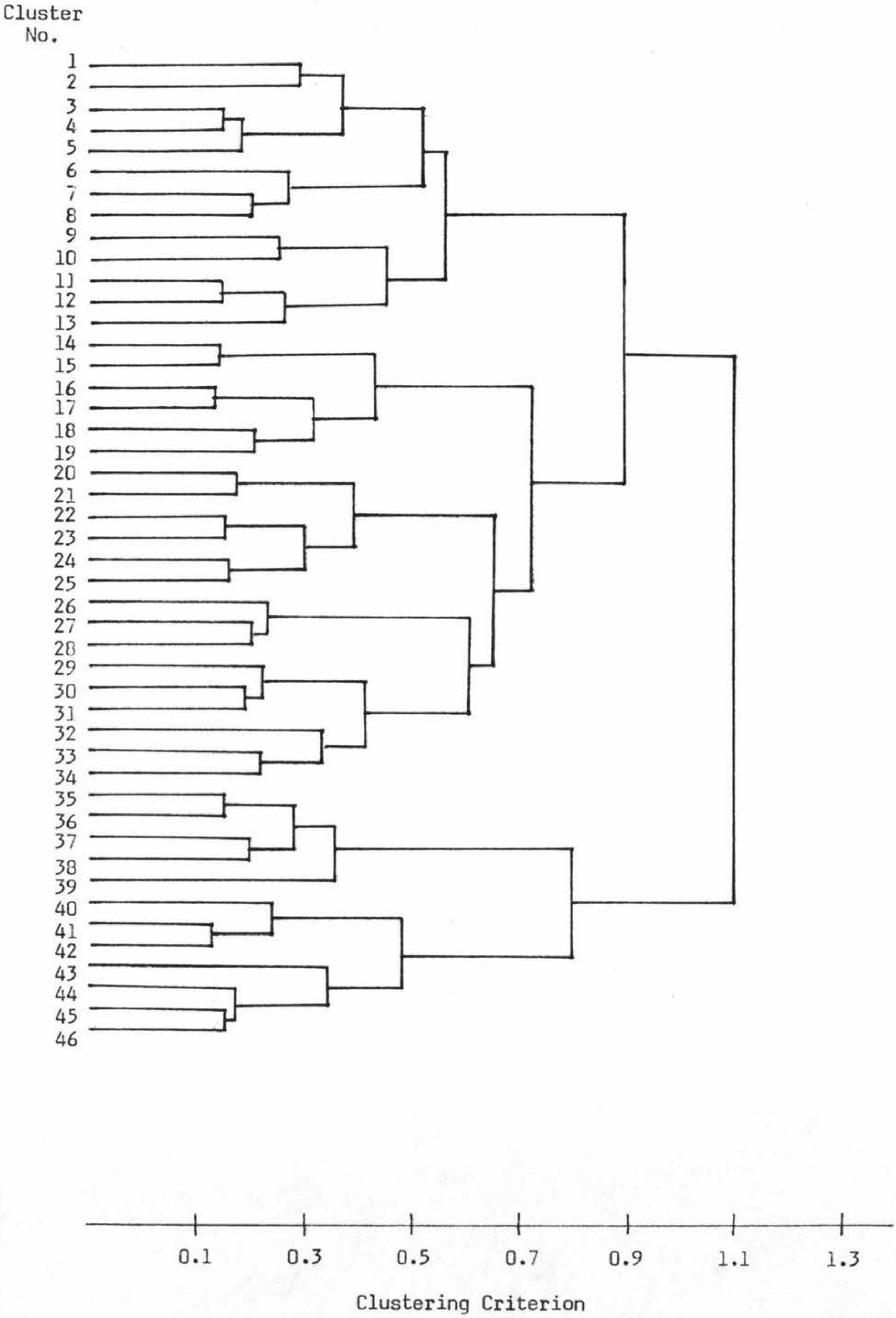


FIGURE 3.12 Dendrogram of DISCCHARA by Ward's Method.

Stage	Criterion	C.ERE	C.HEI	RUST	F.COL	F.DAT
1	0.0001	0.000	0.000	0.000	0.000	0.000
10	0.0016	0.001	0.001	0.001	0.002	0.001
20	0.0042	0.002	0.004	0.004	0.004	0.003
30	0.0082	0.005	0.010	0.008	0.007	0.005
40	0.0133	0.008	0.012	0.019	0.013	0.009
50	0.0197	0.013	0.021	0.029	0.018	0.012
60	0.0280	0.017	0.024	0.044	0.027	0.016
70	0.0386	0.024	0.038	0.063	0.033	0.023
80	0.0524	0.036	0.049	0.081	0.056	0.035
90	0.0700	0.048	0.064	0.090	0.074	0.043
100	0.0906	0.057	0.075	0.109	0.105	0.061
110	0.1159	0.070	0.094	0.164	0.123	0.072
114	0.1288	0.075	0.117	0.197	0.130	0.077
120	0.1524	0.094	0.153	0.216	0.147	0.090
130	0.2064	0.131	0.219	0.277	0.201	0.127
140	0.2905	0.196	0.311	0.367	0.270	0.200
150	0.4496	0.335	0.439	0.551	0.432	0.276
155	0.6550	0.437	0.667	0.780	0.654	0.343
159	1.1051	1.000	1.000	1.000	1.000	1.000

TABLE 3.24 The Proportion Of Sum Of Squares (WSS/TSS) Not Explained By Clustering At Different Stages Of Clustering By Ward's Method for DISCCHARA

The ranks of the cluster means for each character are shown in Table 3.25. Although DISCCHARA was based on the 1st, 2nd, 4th, 7th and 8th most dominant characters of ALLCHARA (at stage 110), the constituents of the 46 clusters obtained were different to those of the 50 clusters of ALLCHARA. This suggested that the other six characters were only comparatively dormant (in ALLCHARA), and they contributed significantly to the overall similarity. If, instead of these 5 most discriminant characters, the 5 most dominant characters (of ALLCHARA at stage 110) were used in the analysis, the pattern of the resultant clusters would still be expected to differ from those of ALLCHARA. This is because the 6 least dominant characters still contributed considerably to clustering, being only partially dormant (as discussed in section 3.3.1).

The extrinsically intrinsic study again did not reveal any clear cut pattern, which implied that, as before, there were no true ecotype evident, using this set of attributes.

After stage 114, the growth of the WSS/TSS ratio for F.DAT and C.ERE remained slow. However growth of the ratio for F.DAT was slower than that of C.ERE. The F.DAT ratio reached 27.6% and C.ERE reached 33.5% at stage 150. Rust remained as the most dormant character. The relationship between the 46 clusters can be described briefly from the dominant characters, as in Table 3. 26.

3.3.4 JACQCHARA

The results of SEFWIG again showed, as expected, that TSS remained constant (at 0.930705) during clustering, ASS and df_A decreased as WSS and df_W increased. The overall F-ratio changed from 251.672 (stage 1) to a minimum value of 27.5719 (stage 136), and then fluctuated between 27.6528 and 27.9386. Finally F increased from 27.9468 (stage 143) to 41.3563 (stage 158). The associated probability changed from 0.051586 (stage 1) to a minimum value of 0.22935×10^{-11} (stage 116), and then increased to 0.17575×10^{-5} (stage 158). These trends were similar to those of ALLCHARA (section 3.3.1), AGROCHARA (section 3.3.2) and DISCCHARA (section 3.3.3). Following these criteria (as before), the clustering was cut off at stage 116, where the associated probability of the F-ratio was lowest. This resulted in 44 clusters.

RANK	C.ERE	C.HEI	RUST	F.COL	F.DAT
1.	43	42	18	27	34
2	42	40	19	26	33
3	13	43	43	28	21
4	39	41	17	33	20
5	41	13	25	20	29
6	44	3	5	19	19
7	46	5	45	29	32
8	36	29	16	18	15
9	40	4	6	44	27
10	45	46	37	15	16
11	30	30	36	21	23
12	35	35	15	31	30
13	12	2	29	14	31
14	37	25	24	9	26
15	11	12	38	32	17
16	29	1	21	34	11
17	5	44	33	25	44
18	14	31	35	24	3
19	16	38	22	38	9
20	31	21	14	3	14
21	2	10	31	43	24
22	1	32	27	46	5
23	18	39	44	2	1
24	10	23	8	16	22
25	7	26	46	7	12
26	38	45	28	1	18
27	4	17	41	45	8
28	3	36	1	17	46
29	17	22	7	30	28
30	15	18	42	23	4
31	33	20	3	8	45
32	26	16	23	40	10
33	6	33	11	35	13
34	32	7	2	22	25
35	8	6	30	37	7
36	23	37	20	10	43
37	22	28	40	6	40
38	28	34	26	4	6
39	25	15	32	39	2
40	19	11	4	13	42
41	34	24	12	11	41
42	9	14	10	5	35
43	20	8	34	36	36
44	21	9	9	12	38
45	27	19	39	42	39
46	24	27	13	41	37

TABLE 3.25 The Ranks Of The Means Of Each Character Of 46 Clusters In DISCCHARA Analysis

Clusters									
Character	1-5	6-8	9-13	14-19	20-25	26-28	29-34	35-39	40-46
F.DAT	M	M-	M	M	M	M+	H	L	M-
C.ERE	M	M-	S	M	L	M-	M	M+	H
C.HEI	M+	M-	S	M-	M	M-	M	M	H

TABLE 3.26 Brief Grouping Of The 46 Clusters Of DISCCHARA And Their Approximate Average Ranking In Dominant Charaters. H = High, M+ = Medium High, M= Medium, M- = Medium Low, L= Low and S = Spread Out.

Of the 44 clusters obtained, 8 were single entity clusters (cluster Nos 9, 17, 19, 30, 40, 41, 43 and 44). There were 6, 11, 9, 4 and 2 clusters containing 2, 3, 4, 5 and 6 groups respectively. The other 4 clusters each contained 7, 9, 10 and 13 groups respectively. The dendrogram is shown in Figure 3. 13 .

Table 3.27 shows the proportion of the TSS not explained by clustering (WSS/TSS). At stage 116, F.DAT was well partitioned amongst clusters with only 6% of TSS not explained by clustering, and the other 94% being accounted for by the among cluster sum of squares. C.ERE was also well partitioned amongst clusters, with only 6.2% of TSS not explained by clustering. The other two characters in this analysis were L.WID(11.1%) and RUST(12.2%). These results implied that F.DAT and C.ERE were the most dominant characters, whereas RUST was the most dormant character. However, this dormancy was only comparative.

The ranks of cluster means for each character are shown in Table 3. 28 Although JACQCHARA was based on the 1st, 2nd, 4th and 5th most dominant characters of ALLCHARA (at stage 110), the constituents of the 44 clusters obtained were different to those of the 50 clusters of ALLCHARA. This further suggested that the other characters, though comparatively dormant, contributed significantly to the overall similarity amongst groups.

This set of characters had been nominated by Jacques(1962) as being ecocline indicators. However, this extrinsically intrinsic study did not reveal any clear-cut pattern. It did not show the ecoclinal trends proposed by Jacques(1962), and Munro(1961). Jacques(1962) proposed that in moving north-ward through New Zealand, there was an increasing degree of persistence, vigour, rust resistance and erectness. However this polythetic clustering of a more representative accession sample did not confirm his proposal. This matter is discussed further in section 4.2.

After stage 116, the growth of the WSS/TSS ratio remained slow for F.DAT and C.ERE. However C.ERE became the most dominant character with 22.5% of TSS not explained by clustering at stage 150. The other three characters were F.DAT(25.9%), RUST(49.7%) and L.WID(62.8%). L.WID became the most dormant character. The relationships between the 44 clusters can be briefly described as in Table 3. 29 .

Cluster No.	Group No.
1	123, 124, 15, 146, 84, 139, 140, 27, 136, 46, 152, 47, 104.
2	86, 87.
3	108, 132, 102.
4	121, 131, 75, 98.
5	70, 90, 2, 109, 119.
6	51, 130, 127.
7	37, 115, 64.
8	101, 117.
9	85.
10	5, 126, 57, 97.
11	36, 107, 33, 133, 9, 16, 50, 21, 44, 118.
12	14, 62, 41.
13	31, 35.
14	42, 45, 34, 59.
15	3, 78, 6, 1.
16	55, 93, 149.
17	56.
18	60, 111, 58.
19	52.
20	17, 157, 43, 13.
21	80, 125, 7, 69.
22	53, 54.
23	95, 96, 29, 141, 67, 113.
24	92, 129, 83, 160.
25	24, 71, 11, 94, 18, 74, 48.
26	76, 77, 28, 106, 38.
27	20, 135, 82, 116, 110, 23.
28	26, 65, 88, 147.
29	73, 114, 89, 112, 99.
30	49.
31	22, 61, 105.
32	145, 153, 81, 103, 128.
33	30, 100, 8.
34	72, 143, 25, 39, 66, 120, 4, 19, 122.
35	138, 142, 134.
36	79, 150.
37	12, 151, 137.
38	40, 148, 32.
39	91, 154, 155, 156.
40	159.
41	158.
42	10, 144.
43	63.
44	68.

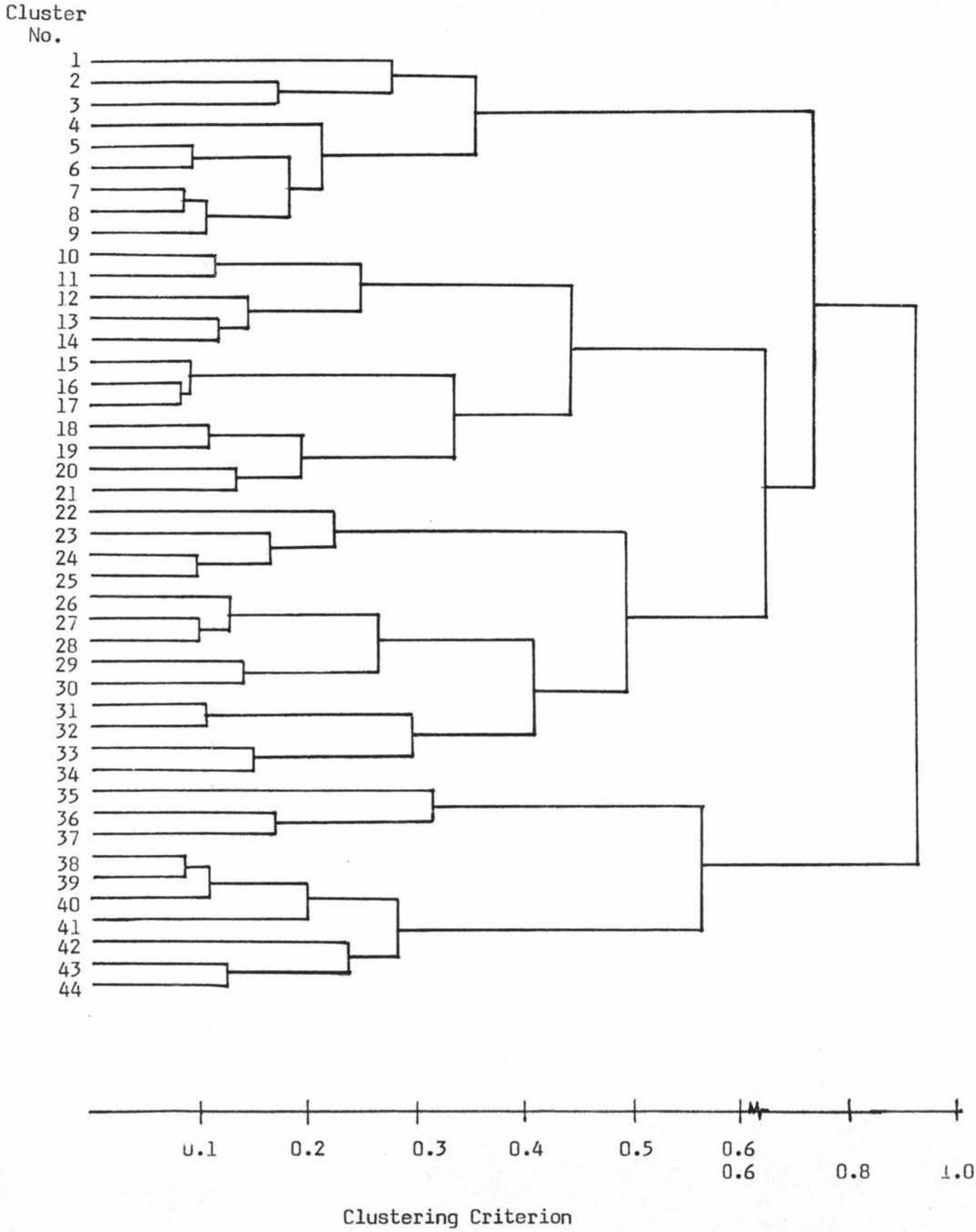


FIGURE 3.13 Dendrogram of JACQCHARA by Ward's Method.

Stage	Criterion	C.ERE	RUST	L.WID	F.DAT
1	0.00002	0.000	0.000	0.000	0.000
10	0.00057	0.001	0.001	0.001	0.000
20	0.00203	0.002	0.003	0.004	0.001
30	0.00418	0.004	0.007	0.008	0.002
40	0.00680	0.006	0.011	0.010	0.005
50	0.01018	0.008	0.018	0.017	0.007
60	0.01453	0.012	0.022	0.025	0.010
70	0.01980	0.017	0.026	0.033	0.015
80	0.02668	0.023	0.039	0.038	0.022
90	0.03610	0.030	0.055	0.052	0.027
100	0.04819	0.039	0.073	0.072	0.036
110	0.06622	0.056	0.111	0.096	0.044
116	0.08010	0.062	0.122	0.111	0.060
120	0.09126	0.067	0.144	0.135	0.070
130	0.12913	0.101	0.205	0.214	0.096
140	0.19423	0.180	0.323	0.310	0.117
150	0.33457	0.220	0.497	0.628	0.259
155	0.49728	0.367	0.847	0.954	0.318
159	0.93071	1.000	1.000	1.000	1.000

TABLE 3.27 The Proportion Of Sums Of Squares (WSS/TSS) Not Explained By Clustering, At Different Stages Of Clustering By Ward's Method For JACQCHARA.

RANK	C.ERE	RUST	C.WID	F.DAT
1	42	13	44	2
2	44	29	30	3
3	40	44	35	8
4	41	22	9	6
5	43	30	21	28
6	39	26	33	7
7	31	6	31	29
8	37	12	14	9
9	33	35	40	5
10	38	14	18	24
11	35	36	34	26
12	22	27	32	1
13	32	42	29	32
14	18	37	11	27
15	23	34	41	30
16	20	2	20	4
17	27	33	27	43
18	30	11	42	31
19	34	4	13	25
20	36	43	39	10
21	28	24	5	23
22	17	8	1	21
23	16	5	2	11
24	25	25	19	19
25	13	28	6	34
26	21	31	28	22
27	26	16	16	18
28	14	10	7	13
29	1	9	10	44
30	29	23	12	40
31	15	21	26	38
32	24	32	37	15
33	2	39	15	20
34	5	15	25	39
35	3	20	8	12
36	11	1	38	33
37	19	38	43	14
38	10	7	3	42
39	12	3	36	16
40	6	17	22	17
41	7	41	17	37
42	4	40	24	36
43	9	18	4	41
44	8	19	23	35

TABLE 3.28 The Ranks Of Means Of Each Character Of 44 Clusters In JACQCHARA Analysis

	Clusters									
Characters	1-3	4-9	10-14	15-17	18-21	22-25	26-30	31-34	35-37	38-44
F.DAT	H	H	M-	M-	M-	M+	M+	M	L	M-
C.ERE	M-	L	M-	M-	M	M	M	M+	M+	H

TABLE 3.29 Brief Grouping Of The 44 Groups Of JACQCHARA And Their Approximate Average Ranking In Discriminant Characters. H = High, M+ = Medium High, M= Medium, M- = Medium Low, and L = Low.

CHAPTER 4. GENERAL DISCUSSION

4.1 Multivariate Analysis

Some general issues concerning the statistical methods used in this study are discussed further.

4.1.1 Multivariate Versus Univariate Analyses

Multivariate analysis is preferable to a series of univariate variance analysis because the latter ignores correlation amongst characters. Because multivariate analysis considers these covariances, it regards the relationships, interdependence and relative importance amongst all characters (Kshirsagar 1972).

4.1.2 Model Used

In this study a one-way MANOVA model was used. The model (as shown in section 2.3.1) implied that the W-MSCP matrix was composed of the variation due to replicates (blocks), group x replicate interaction (experimental error) and within-plot variation. The replicate variation could have been partitioned out by a two way MANOVA. But for the present purposes, it seemed sufficient to amalgamate all these sources of variance into one component (the "within-group variance"), as the object was to contrast amongst-group variance against the rest.

4.1.3 Data Transformations

There was no attempt to transform the original data, even though the results revealed that the MSCP matrices of each group were not equal. This approach was adopted because: (1) the effect of marginal (character by character) transformations was not certain (as discussed in section 1.4.5); (2) joint transformation would be complex, and its validity was doubtful (as discussed in section 1.4.5); (3) complex transformations would reduce the flexibility and interpretability of the original data. The use of discriminant functions amounts to a form of transformation, but this overcame only character covariances, this being a prerequisite for the correct calculation of SED.

4.1.4 Data "Crunching"

In this study, MANOVA was used to summarize the bulk data for multiple discriminant analysis. For ALLCHARA, it has summarized the 160 groups x 24 plants x 11 characters data matrix into a 160 x11 character-means data matrix. Multiple discriminant analysis then transformed the correlated characters (i.e. 11 original means of each group) into uncorrelated discriminant functions. These uncorrelated discriminant functions were then used to calculate the SED for the clusering analysis. The clustering analysis compressed the data further into a 50 clusters x 11 characters data matrix. Under this series of statistical methods, a huge amoung of data (approximately 42,000 elements) has been reduced to a managable and interpretable size (550 elements). For AGROCHARA, the data matrix was reduced from 160 x 24 x 8 to 44 x 8; for DISCCHARA, it was reduced from 160 x 24 x 5 to 46 x 5; and for JACQCHARA, it was reduced from 160 x 24 x 4 to 44 x 4. This illustrates well the power of these methods in extracting the essential information from large data sets and in reducing them to a size which can reasonably be examined and comprehened.

4.1.5 Squared Euclidean Distance as a Similarity Measure

In this study only standardized SED was used as the similarity measure, because of its advantages over other measures. It is additive over attributes, it is a size measure, and it possesses combinatorial properties (as discussed in section 1.7.3). However, the typical property of SED (giving extra weight to outlying values) was obvious in this study. This can be seen from the dendrograms (Figures 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 3.11, 3.12 and 3.13), in that most of the single-entity clusters possess an extreme value in one or more characters. This property could be desirable in some studies, as it isolates the outlying groups. However, it might not be preferred in other studies, such as in ecology.

4.1.6 Probabilistic Decision on Clustering Cut-Off

For these four sets of attributes, SEFWIG seemed to provide a useful decision-base for choosing a clustering cut off point, the use of which has been discussed previously (section 3.2 and 3.3). It always defined a clear minimum probability point (Figure 3.2 and 3.10), but this behaviour could have been a property of this set of data. The method needs further use and evaluation, but it seems very promising.

These "cut-off" points were defined more objectively than with other methods. This was particularly useful with this set of data, for this study was a pioneering one in this species. There was no a priori information from which to judge a suitable cut-off point by more common, subjective methods. The a posteriori examination of the resultant clusters revealed: (a) virtually no contentious memberships, and (b) an acceptable structure of association and division amongst the clusters. No adjustment of cluster boundaries or membership seemed necessary. Further use of the procedure will be interesting to see if it's efficacy is general, particularly with Ward's method and weakly-structured data.

4.2 Ecotype Studies and Ecoclinal Trends

This study did not reveal any ecotypes irrespective of which set of attributes was used. This result could arise from two possibilities: firstly, there were no ecotypes, secondly there were ecotypes, but the approach used in this study could not reveal them. The latter needs serious consideration because of the following reasons. Firstly the external attributes available were neither complete nor detailed enough (see Appendix B-1). e.g. Nearly half of the 201 accessions were without information on the site altitude. Secondly, and in consequence of the previous reasons, an extrinsically intrinsic study was carried out. This study compared only one external attribute, at a time, with clusters formed on the basis of internal attributes. Thirdly, the clustering strategy used in this study (Ward's method) might not be appropriate for this ecological purpose. The intense clustering of Ward's method was preferable in this study, with the aim of "artificially sharpening" the boundaries of the weakly structured data. This aim was similar to that of a taxonomist, whose main interest is primarily in "homostate" or "stat" (Williams 1971) or "internal cohesion" (Cormack 1971) (i.e. clusters defined entirely by internal similarities). This is in contrast to the ecologist's aim, which is primarily in "segregate" or "ait" (Williams 1971), or "external isolation" (Cormack 1971) (i.e. a cluster which may or may not be internally homogenous, but which is defined by its extrinsic separation from other clusters). Thus, the main aim of this study was contrary to that of the ecologist. Furthermore, it should be recalled that the properties of SED might not suit the ecologist's purpose. Fourthly, most of the internal attributes, forming the present bases of clustering, were agronomic characters (8 out of 11 characters were agronomic). Agronomic characters are known to be non-stable in that they are location and time dependent (Burt <u>et</u>. <u>al</u>. 1971). That is they have a comparatively large genotype x environment interaction than morphological characters. In this study, where the accessions have been grown for several years in one site, genotype and genotype x environment interaction effects would be confounded.

There are some suggestions,which might be considered, for a future ecotype study. (1) more detailed and comprehensive records of external attributes are needed, such as the latitude, altitude, soil type, essential soil properties, aspect, and seasonal properties. (2) More internal attributes should be measured, especially morphological characters (such as floral characters, tiller and leaf characters). If possible, some intrinsic characters (especially agronomic) of the plants should be measured at the accession site as well as at the experimental site. This would enable some measure of genotype x environment interaction and adjustment to be done. (3) Relationships between internal and external attributes shoud be examined by canonical correlation analysis. (4) Different clustering strategies (such as group average, minimum information gain or hierarchical divisive mehtod) could then be applied to these two sets of attributes separately. A subsequent comparison of the resultant clustering patterns should show the ecotypes distributions, if any existed.

In this JACQCHARA study, the results did not support the ecoclinal hypothesis of Jacques(1962) and Munro(1961). They proposed that there were ecoclinal trends from South(cold) to North(warm) of New Zealand, indicated by an increasing degree of persistency, vigour, resistance to rust and erectness in growth forms. The present disagreement did not necessarily disprove their hypothesis, as discussed already. However, the earlier studies also suffered from experimental difficulties. They examined the ecoclinal trends univariately, thereby ignoring correlations amongst characters. This study did not omit such correlations. The present study also examined a wider sample than the previous ones (refer to Introduction).

Despite the possible inappropriateness of the methods of the present study for ecological purposes, it was clear that these results revealed no cluster distribution which concurred with the ecoclinal trends suggested by Jacques(1962) and Munro(1961). If there are genuine ecotypes present in this species in New Zealand, they appear not to be distinctly separated, as indicated by the fact that this extensive data was weakly structured. None of the New Zealand studies to date on Yorkshire Fog have been ideal for examining ecotypes and so the matter is unresolved. The present evidence does suggest, however, that eco-clinal trends may be weak and perhaps are only incipient. Further research along the lines discussed earlier, needs to be done to examine this question critically.

4.3 Agronomic and Plant Breeding Aspects

4.3.1 Agronomic Relevance of Characters Assessed

The field collection of this study (working collection) was also the genetic resources (active collection or base collection) of Yorkshire Fog in New Zealand. Therefore the statistics obtained from ALLCHARA not only evaluated the working collection, but also provided important information about the genetic resources. However the set of attributes was fairly restricted for both purposes. To be more useful, especially as regards genetic resource evaluation, a larger set of attributes should be collected, in order to describe the variation more thouroughly, other agronomic attributes, such as duration of flowering time, actual performance as spaced palnts (annual yields and/or seasonal yields), and tillering habit, would be useful. The practicability of obtaining them for such a large collection may be questionable, however, other morphological attributes, such as pubesence on leaf, and leaf shape, would also be of value.

The clustering analysis of this collection was aimed at both agronomic and plant breeding use. As pointed out by Burt et. al.(1971), two morphologically distinct plants may be similar in agronomic performance, and conversely two morphologically similar plants may have distinctly different agronomic performances. Clustering based on morphological attributes has principally a taxonomic application, being of limited use agronomically. Conversely, clustering based only on agronomic characters may be too dependent on location and time. The attributes of the ALLCHARA analyses were both morphological and agronomic. These are considered briefly in the following.

F.DAT reflected the date of infloresence emergence and hence summer maturation. C.ERE reflected the growth forms of the plant. This character has been suggested as being related to the palatability (i.e. acceptability by grazing animal) of the grass (Jacques 1962). He suggested that the postrate from (Low C.ERE) was non-palatable, and the erect or semierect (high C.ERE) forms were preferred. F.DAT and C.ERE have been found to be the most discriminating characters amongst groups (section3.1). They also had the highest correlation (positive or negative) with the first discriminant function. From the breeder's point of view, this suggests that selection amongst groups for thses two characters should be promising; but this also depends on their having moderate-high predictive heritability.

L.WID was one of the character indicating herbage yield, and also indicated light intercepting ability (Jacques 1974). Jacques (1974) suggested that the wider leaf of Yorkshire Fog utilized incoming light more efficiently, making it comparatively more aggressive than perenniel ryegrass under zero grazing condition. This suggested that broad leaf (high L.WID) was agronomically preferable.

L.ROL was considered as a xerophytic character. Leaf roll could be considered an adaptation to arid conditions, as it may reduce water loss. This would be an important character for drought resistant cultivars.

L.COL and F.COL reflected putative pigment content. This pigment has tentatively been assumed to be flavonoid. If this also reflects tannins, such as catechins, it may be related to lack of palatability, for which Yorkshire Fog has a reputation (Jacques 1962).

RUST and O.DIS may also be connected with non-palatability (JACQUES 1962, 1974), as well as being of obvious importance with respect to yield. The main indicators of herbage yield (in clumps) were C.HEI, C.DIA and C.DEN together with L.WID. Direct measurement of yield was not practicable in view of the large number of genotypes.

4.3.2 Limitations of The Study and Subsequent Analyses

As noted previously, the pattern of clusters relies not only on the strategy but also on the set of attributes used. This was illustrated by the comparison of clustering for ALLCHARA and AGROCHARA. Here the exclusion of the more dormant characters altered the clustering pattern considerably. This is especially so for intense clustering methods, as they are very sensitive with respect to changes (Cormack 1971). Thus, all the clustering patterns obtained in this study were unique not only because of the clustering strategy (Ward's method), but also because of the set of attributes used (also to some extend because of the similarity measure used). With most of the attributes being "unstable" agronomic characters, these clustering patterns should also be considered as location and time dependent. A subsequent "goal oriented" clustering analysis for agronomic or plant breeding use could use only the attributes which influence the ultimate goal. For example, if the goal is for increasing yield, then those attributes that will affect yield should be used alone, such as, clump diameters, clump height, clump density, leaf width and dry matter %. Although the other attributes will not have been used in the clustering analysis, they could be used as secondary attributes for selecting a particular group within the chosen cluster (refer to section 4.3.5). Probably an "all characters" analysis should always be included and an overall "agronomic" analysis also has obvious utility.

The set of individuals (groups) used, will affected the pattern of clusters, also. This is especially so for an intense clustering strategy (Cormack 1971). The collection used in this study was a highly representative sample of Yorkshire Fog in New Zealand. They were sampled from most parts of the country (as seen in Appendix B-1). Thus the pattern of clusters obtained should reflect well the New Zealand situation.

4.3.3 Variation Amongst Individuals Within Groups

In this collection, each group consisted of 24 (or less) individuals, each of which was potentially a different genotype, because of cross pollination. However, the model indicates that the clustering analysis in this study was based on the discriminant function of means of each group. Therefore th genetic variation amongst individuals within each group was not partitioned out, but was included in W-MSCP. A suggested further study might be of interest. It would consider each individual (3803 of them) as a different genotype and subject this to the same analytical sequence as was used for groups. However, this analysis on the individual genotype variation, is too large for many present computers. e.g. There would be a total of ½(3803 x 3802) (=7,229,503) interindividual similarity measures in the present data, and the maximum array permissable in the local B6700 computer is only 65,535 words. A study of groups was therefore more readily accomplished. In any case, data "crunching" with individuals may conclude with still too many clusters for it to be of value, or to be comprehended.

4.3.4 Agronomic Evaluation of The Groups (Accessions)

In this study a few outstanding groups were found. The most outstanding one was group 68 (from Pioneer Highway of Palmerston North, Manawatu), which had an erect, tall, compact, large clump, with good disease resistance, broad and flat leaf with green tip and a medium-early flowering date. Group 63 (from 3 miles North of Putaruru, near Hamilton) had a similar performance. It had an erect and compact clump, of moderately high, flat and narrow leaf with green tip, and good overall disease resistance. It was not as outstanding as group 68, because of it's narrow leaf and moderately high clump. Group 144 had a compact, erect, high but small clump, with medium-low disease resistance, flat and medium broad leaf, and early flowering. From the preliminary results, these groups looked to be promissing breeding materials. They could be utilized as line selections, following further evaluation.

4.3.5 Cluster Analysis and The Choice of Parents

Clustering analysis can be used to help identify parental groups useful in planning crossing program for plant breeding. Parental groups within clusters were phenotypically similar with respect to all attributes examined. A simplifying assumption is that phenotypic similarity reflects genotypic similarity. Consequently, crossing of groups within any one cluster is not expected to provide great genetic variation in the F_2 and later generations. Conversely, the greater phenotypic variation amongst clusters is assumed to reflect greater genotypic diversity also. Therefore, if the parents belong to different clusters, a much wider genetic variation is expected, for selection to operate upon in segregating generations. This does not imply that one has to choose necessaily from the extremes of a character to create genetic diversity, because this could included undesirable alleles. To illustrate this, the clustering pattern of ALLCHARA has been used as an example. Assuming the aim of a breeding project in for early flowering date (refer to section 3.3.1), a cross between parents chose from clusters 44-50 (clusters which had early flowering date) and parents chosen from clusters 20-23 (clusters which had late flowering date) would be expected to produce great genetic variation for this charcter. However, as the aim is for early flowering, the "late" tail of variation will not be of much use. Conversely, if a cross between parents chose from different clusters of the amalgam containing clusters 44-50, the variation will be expected to be smaller but mainly at the useful "early" end of this character. In order to decide which of the several groups within a cluster may be used for crossing, the secondary characters (such as disease resistance, herbage yield, or seed yield) should be taken into account.

CONCLUSIONS

1. The MSCP matrices of the 160 groups were found to be heterogeneous. Non-multivariate normality was believed to be one of the causes. Despite of these, the differences among groups were highly significant.

2. There are considerable phenotypic divergences among groups. Group 68 (from Pioneer Highway of Palmerston North), group 63 (from 3 miles North of Putaruru, near Hamilton) and group 144 ("Massey Basyn" of Massey University) possessed most of the agronomic desirable characters.

3. Among the characters studied, flowering date and clump erectness were the two most important characters. They had the largest eta-values, hence were the two most discriminating characters among groups. They had the highest correlation with 1 st and 2 nd discriminant functions. Also they were the most dominant characters in clustering, which influenced the clustering pattern most.

4. For all set of attributes, all the discriminant functions were retained, as they were significant.

5. The clustering behaviours of the seven agglomerative clustering strategies, using ALLCHARA, agreed with the finding of most of the other authors. The reversals of Median and Centroid Methods, the chaining effects of Single Linkage Method and the intense clustering of Ward's Method were obvious in this study.

6. This study did not reveal any ecotypes irrespective of which set of attributes used. Also in the JACQCHARA study, the results did not support the ecoclinal trends hypothesis of Jacques. It was suggested that the ecoclinal trends might be weak and perhap were only incipient. Further reserarch need to be done.

APPENDIX A-1 KEY TO MULTIVARIATE ANALYSIS

MULTIVARIATE ANALYSIS

I. Observations are sampled from one population (e.g. one cultivar, or one ecotype); and concern is only with the pattern of variation and covariation of this single sample.

A. Observations are described by one homogeneous set of attributes (i.e. one set of characters).

1. The main purpose is to described the total variance-covariance in a sample in few dimensions, i.e. to reduce the dimensionality of the original data while minimizing any loss of information. The few dimensions are the linear combinations of the original attributes that successively account for the major independent pattern of variation in the original attributes of the population.

(a) The observations are described by a series of P-axes, each representing a separate attribute.

PRINCIPAL COMPONENT ANALYSIS

(b) The observations are described by $\frac{1}{2}n(n-1)$ inter-observation similarity (ordissimilarity) measures (N = no. of observations.

PRINCIPAL COORDINATE ANALYSIS

2. The main purpose is to study the correlation structure underlying the inter correlations amongst the observed attributes; i.e. to reproduce only the inter correlations rather than the total variance.

FACTOR ANALYSIS

B. Observations are described by more than one set of attributes (e.g. (1) dependent and independent characters; or (2) intrinsic and exintrinsic characters).

1. The main purpose is to establish maximal linear functional relationships between dependent and independent attributes.
   MULTIPLE REGRESSION AND CORRELATION

2. The main purpose is to establish relationships between a series of observations described by these sets of data.
   CANONICAL CORRELATION ANALYSIS

II Observations are sampled from more than one population (e.g. several cultivars, or ecotypes).

A. Observations are described by one homogeneous set of attributes.

1. The main purpose is to determine if the samples could have been drawn from a single statistical population; i.e. are the mean vectors of the sample populations equal?
   MULTIVARIATE ANALYSIS OF VARIANCE

*2. The main purpose is to find a set of linear functions for the variables that maximize differences among sample populations.

a. To maximize the ratio of among group sums of squares to within-group sums of squares, subject to the condition that the coefficients are orthogonal.
   MULTIPLE DISCRIMINANT ANALYSIS

b. To maximize the among-group variance and covariance, subject to the condition that the within-group variances are unity and within group covariances are zero.
   CANONICAL VARIATE ANALYSIS

*3. The main purpose is to find a set of g linear functions that serve as indices for classifying new observations into one of g pre-defined populations.

a when g = 2

DISCRIMINANT ANALYSIS

b when $g > 2$

GENERALIZED DISCRIMINANT ANALYSIS

4. The main purpose is to sort a previously unpartitioned heterogeneous collection of objects into a series of sets.

   CLUSTER ANALYSIS

5. The main purpose is to arrange the objects graphically in few dimensions, while retaining maximal fidelity to the original inter object relationships.

   NON METRIC SCALING

B. Observations are described by more than one set of attributes.

1. The main purpose is to determine if the samples could have been drawn from a single statistical population after covariance adjustment on one set of variables by the other sets.

   MULTIVARIATE COVARIANCE ANALYSIS

* When g = 2, canonical variate analysis and discriminate analysis is the same. Since the number of functions extracted depend on g-1 (when $(g-1) < p$) or p (when $p \leq (g-1)$). Therefore when g = 2 there exists only one linear function. The canonical variable is then the discriminate function. The new observation will be allocated to one or other group depending on the sign of its canonical variable (positive or negative) (Seal 1968).

When $g > 2$ canonical variate analysis is similar to multiple discriminate analysis.

References used in the key

Anderson (1958)
Bryant & Atchley (1975)
Cooley and Lohnes (1971)
Gower (1966, 1968)
Rohlf (1971)
Seal (1968)
Clifford & Stephenson (1975)

APPENDIX A-2

Special Cases of F Approximation for Wilks' Lambda

Parameter				
p	g	$F_{(n_1,\, n_2)}$	n_1	n_2
Any	2	$\frac{1-\lambda}{\lambda} \; \frac{n-p-1}{p}$	p	n-p-1
Any	3	$\frac{1-\lambda^{\frac{1}{2}}}{\lambda^{\frac{1}{2}}} \; \frac{n-p-2}{p}$	2p	2(n-p-2)
1	Any	$\frac{1-\lambda}{\lambda} \; \frac{n-g}{g-1}$	g-1	(n-g)
2	Any	$\frac{1-\lambda^{\frac{1}{2}}}{\lambda} \; \frac{n-g-1}{g-1}$	2(g-1)	2(n-g-1)

That is

If $1 \leq p \leq 2$

$$F_{(n_1,\, n_2)} = n_2\,(1-\lambda^{(1/p)})/(n_1\,\lambda^{(1/p)}) \qquad \begin{aligned} n_1 &= p(g-1) \\ n_2 &= p(n-g-p+1) \end{aligned}$$

If $2 \leq g \leq 3$

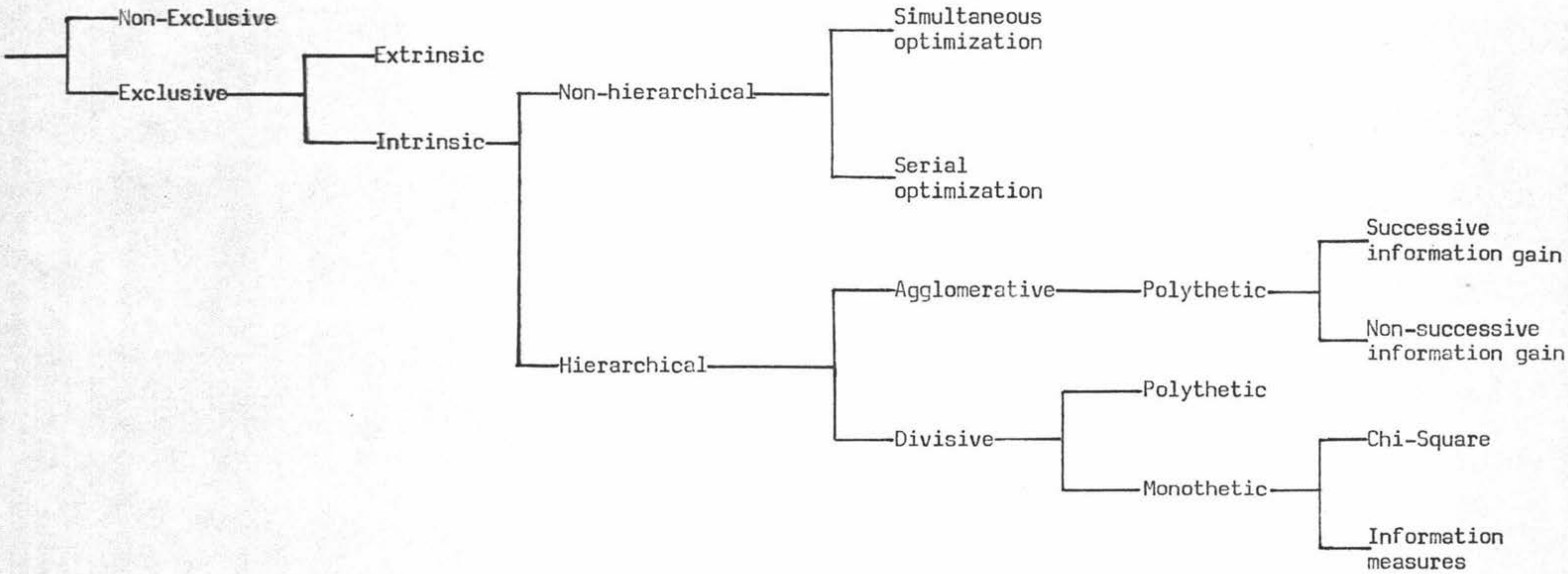
$$F_{(n_1,\, n_2 0} = n_2\,(1-\lambda^{(1/(g-1))})/(n_1\,\lambda^{(1/(g-1))}) \qquad \begin{aligned} n_1 &= p(g-1) \\ n_2 &= (g-1)(n-g-p+1) \end{aligned}$$

APPENDIX A-3 DIFFERENT TERMINOLOGY FOR ATTRIBUTES

<table>
<tr><th>Clifford and Stephenson</th><th>Conover</th><th>Burr</th><th>Goodall</th><th colspan="2">Gower</th><th colspan="2">Lance and Williams</th></tr>
<tr><td rowspan="2">Binary</td><td rowspan="4">Nominal</td><td rowspan="4">Nominal</td><td rowspan="2">Binary</td><td rowspan="2">alternative</td><td>symmetric</td><td colspan="2" rowspan="2">Binary or Qualitative</td></tr>
<tr><td>asymmetric</td></tr>
<tr><td rowspan="2">Disordered Multistate</td><td rowspan="2">Qualitative</td><td colspan="2" rowspan="2">Qualitative</td><td rowspan="2">Disordered Multistate</td><td>exclusive</td></tr>
<tr><td>non-exclusive</td></tr>
<tr><td>Ordered Multistate</td><td>Ordinal</td><td>Ordinal</td><td>Ordered</td><td colspan="2"></td><td colspan="2">Ordered Multistate</td></tr>
<tr><td rowspan="2">Continuous</td><td>interval</td><td rowspan="2">Metric</td><td rowspan="2">Metrical</td><td colspan="2" rowspan="2">Quantitative</td><td colspan="2" rowspan="2">Quantitative</td></tr>
<tr><td>ratio</td></tr>
</table>

APPENDIX A-4

DICHOTOMOUS CHOICE OF CLUSTERING PROCEDURES (BASED ON WILLIAM 1971)



APPENDIX B-1 The Geopgrphical Location, Altitude and Habitat of the 201 Fog Accessions

Collection No.	Group No.	Location	Altitude	Habitat
1	1	Tara Hills, Mckenzie Basin, near Omarama	3700'	Roadside, pasture, localized dense
2		Mt. Coom airfield, Hermitage	2500'	Shortgrassland, brown top Fescue, scattered plants
3		Mt. White Station, Waimakariri River	2500'	Fescue grassland, scattered plants
4	2	Upper Rees River (Hd L. Wakatipu)	1500'	Bush edge & open shortgrassland, scattered plants
5	3	Warepa, S. Otago (near Balclutha)	200'	Pasture (commercial sample), scattered plants
6	4	Tara Hills, Mckenzie Basin	3100'	Fescue grassland, scattered plants
7	5	Mt. Fyffe (hut site) Kaikoura Range	3800'	Snowtussock grassland
8	6	Shotover River, (near Queenstown)	1000'	Roadside, pasture, localised dense
9	7	Kiwi Flat, Makarora River	1250'	Hay paddock, red clover, Timothy, scattered plants
10	8	Mossburn, Southland	1000'	Pasture (commercial sample), scattered plants
11	9	Athol, Southland	1000'	Pasture (commercial sample), scattered plants
12		William's Stream, Clarence River, North Cantabury	4000'	Fescue grassland, scattered plants
13	10	TGMLI plot, Mid Dome Soil Con. Res. Southland	4500'	Plot on bare soil established 1965 "Massey Basyn"
14	11	Styx River, Clarencer, North Cantabury	2600'	Pasture, brown top, scattered plants
15	12	TGMLI plot, Island Gully Pass Clarence/Wairau Rivers	4600'	Plot established 1965 "Massey Basyn"
16	13	Manapouri Station, Manapoure, Southland	700'	Pasture(commercial sample), scattered plants
17	14	Mossburn, Southland	1000'	Pasture (commercial sample), scattered plants
18	15	Williams Stream Clarence River, (Amuri Ski Club Road)	3400'	Stream side, scattered plants
19	16	Upper Rees River (Hd L, Wakatipu)	1500'	Open short grassland, Scattered Plants
20	17	Mid Rivers Glenorchy (Hd L. Wakatipu)	1100'	Rank pasture (commercial sample), scattered plants
21		Crawford Junction Hut, Kokatahi River, Hokitika River	1000'	Hut site, alluvial terrace, locally dense
22	18	Lumsden, Southland	750'	Pasture (commercial sample), scattered plants
23	19	Omarama, Mckenzie Basin	1400'	Road side, Pasture, localised dense

Collection No.	Group No.	Location	Altitude	Habitat
24		William's stream, Clarence River North Cantabury	5400'	Roadside, Snow tussock grasslands, scattered plants
25	20	Tara Hills, Mckenzie Basin	3700'	Fescue grassland, scattered plants
26		Ruataniwha Station, Mckenzie Basin	1550'	Hayshed, pasture, scattered plants
27	21	Upper Clarence River, Bridge (L. Tennyson)	3500'	Alluvial terrace, Brown Top, red tussock, scattered plants
28	22	Dipton, Southland	150'	Scattered plants in grazed pasture
29	23	Comp stream, Craigieburn Range	3400'	Roadside Mountain beech, localized dense
30	24	TGMLI plot, Black Birch, Awatere River	4500'	Sown plot established 1969 on bare soil, dense
31	25	Flora Hut, Hd Takaka River, Nelson	2600'	Hut clearing in Silver Beech forest, localized dense
32		Griegs stream, Branch River, Marlborough	4000'	Tussock grassland, scattered plants
33		North Crown Terrace, (near Arrow town)	2400'	Roadside, pasture, scattered plants
34		South Crown Terrace	2400'	Roadside, pasture, scattered plants
35	26	William's stream, Clarence River (Amure Ski Club Road)	3400'	Stream side, scattered plants
36	27	Wards Pass, Moleswarth, Marlborough	3750'	Open Tussock grassland, scattered plants
37	28	Camerons flat, Matukituki River (near Wanaka)	1400'	Riverbed, sandbank, scattered plants
38		Upper Cleddan River (near Homer Tunnel), Fiordland	2000'	Roadside, broad leaf forest, localised dense
39		Glenorchy airfield, Head of Lake Wakatipu	1100'	Shortgrassland, Brown Top Fescue,scattered plants
40	29	Cattle Flat Station, Matukituki River (near Wanaka)	1300'	Large wet alluvial terrace with bushes etc., scattered plants
41		Upper Makarora River (Tourist lodge)	1000'	Bush clearings, scattered plants
42	30	Mid Crown Terrace (near Arrowtown)	2400'	Roadside, pasture, scattered plants
43	31	Tara Hills, Mackenzie Basin	3800'	Pasture short Tussock grassland, scattered plant
44	32	Manapouri, Southland	700'	Pasture (commercial sample), scattered plant
45		Hooker Flats, Hermitage	2600'	Fescue grassland, scattered plant
46		Makarora Township (Hd L. Wanaka)	950'	Rank sward, scattered plant
47	33	Hd Ahuriri River Mckensie Basin	2800'	Fecue grassland, scattered plant
48	34	Cardrona Valley (near Wanaka)	1000'	Roadside, run country, scattered plant
49		Skippers, Shotover River (near Queenstown)	2100'	Short grassland, scatterd plant
50	35	Cardrona Valley (near Wanaka)	1100'	Roadside, run country, scattered plant

Collection No.	Group No.	Location	Altitude	Habitat
51	36	Hermitage Postoffice	2600'	Recent alluvial deposit, scattered plants
52		Pukaki, Mckensie Basin	1700'	Short grassland, scattered plants
53	37	Greigs stream, Branch River, Marlborough	1800'	Alluvial terrace in open Tussock and Manuka, scattered plants
54	38	"Nursery" Cave stream, Craigieburn Range	3200'	Drained Red Tussock swamp, dense
55	39	Clorence River, North Cantabury	2600'	Fescue grassland, scattered plants
56	40	Cameron's Flat, Matukituki River (near Wanaka)	1400'	Pasture, scattered plants
57	41	Forks River,(Hd of Hollyford River) Fiordland	2700'	Roadside scrubs, localized dense
58		William's stream, Clarence River (Amuri Ski Club Road)	5300'	Roadside, Snow Tussock grassland, scattered plants
59	42	Mt. Cargill (2 mile N W of Port Chalmers)	1100'	Roadside, ridgecrest, Brown Top
60	43	Flag swamp, Main South Road between Palmerston and Waikouaiti	50'	Flat roadside, coastal plain recent alluvial
61		Berwick Forest	1900'	Flat exposed, unimproved Tussock plateau,
62	44	Abbotsford (Dunedin)	700'	Roadside, scattered plants in Brown Top/ gorse association
63	45	Ben Ohau Station, Lake Pukaki	1000'	Amajor component of very old pasture/hay paddock
64		Ryans Beach, Otago Penisular	300'	Laxly grazed pasture on coastal cliffs
65	46	Kurow, Waitakei River	600'	Groved river flats with Brown Top
66		Lake Ohau	1800'	Roadside by shore-stony
67	47	Luggate/Hawea Flat/Tarros Juntions	900'	Dry roadside with Brown Top
68	48	Haast Bridge, South Westland	0'	Waste area off road
69	49	Franz Joseph Glacier, South Westland	0'	Roadside
70	50	Fergusson Bush, Ross, Westland	300'	Waste area near road
71	51	Orowaiti, West Port	0'	Roadside
72	52	Ohikonui River Jen. Buller Gorge	300'	Roadside in Bush
73	53	Inangahua Jen. Buller Gorge	300'	Pasture
74	54	Springs Jen, South Nelson	1500'	Roadside
75	55	Maruia Saddle, South Nelson	2100'	Bush clearing
76	56	Waihopai River (Wairau Valley) Marlbouough	600'	Riverside
77	57	Tongariro N.P. (near Wanganui River)	2700'	Roadside Tussock
78		Huka Falls	1500'	Beside river track

Collection No.	Group No.	Location	Altitude	Habitat
79	58	Kirikiri Saddle, Coromondel	1800'	Roadside in bush
80		Kaimarama, Coromondel	0'	Pasture, in valley
81		5 miles W. of Whakatane, Bay of Plently	0'	Roadside
82	59	Toatoa- Motu Road, Gisborne	2400'	Roadside near bush
83	60	Rotoehu, Rotorua	900'	Lakeside
84	61	Ekatahuna, Wairapa	900'	Roadside
85	62	Rimutaka Pass, Wellington	1800'	Track in secondary bush
86	63	Putaruru (3 miles North)		Pasture
87	64	Punga Road (1 mile east of top of)		
88	65	Andy Hill, Owhango		Ryegrass,, White Clover
89		Owhango		Rotationally grazed pasture, eyegrass Brown Top, White Clover
90	66	Owhango		Continually grazed pasture, ryegrass Brown Top, White Clover
91	67	Old West Road, Palmerston North		Short pasture
92	68	Pioneer Highway (near Rongotea Road turn off)		
93	69	Aorangi Field Station		Improved pasture
94	70	Fielding (1½ miles East of)		Dairy pasture, White Clover and Creeping Fog
95	71	Menzies Ford (between Colyton & Fielding)		
96	72	Valley Road, 2 miles East of Colyton		Pasture , Scattered Plants
97	73	Pohangina Valley Road		Poor hill pasture, Brown Top & weeds
98	74	Saddle Road, Summit		Short grazed Pasture, White Clover, Brown Top, Yorkshire Fog, Dogstail, Scattered Plants
99	75	DSIR Hill Station, Saddle Road		Short grazed Pasture, Yorkshire Fog, mainly Brown Top & Dogstail
100	76	4 miles South of Pahiatua		Improved pasture, Yorkshire Fog, White Clover, rye grass, Creeping Fog and Brown Top
101	77	1½ miles South of Eketahuna		Old pasture on flat, some improved spp. but Brown Top and dogstail

Collection No.	Group No.	Location	Altitude	Habitat
102	78	12½ miles South of Eketahuna, 12 miles North of Masterton		Old pasture on stony soil, Yorkshire Fog, White Clover, Brown Top, dogstail & Creeping Fog
103	79	7 miles North of Masterton, Opaki		Newish pasture, Yorkshire Fog dominant with White Clover, Creeping Fog, many seed head
104	80	7 miles North of Raetihi		Poor, grazed sheep pasture, short
105		Wanganui, Victoria Park		Isolated plants
106	81	5 miles North of Maxwell, Wanganui		Sheep/cattle pasture, mainly Tim, abundant plants isolated head
107	82	North of Wanganui		Roadside, pasture
108	83	Wanganui (2 miles South of Waverley)		Sheep/cattle pasture, Crested dogstail
109	84	1 mile South of Hawera		Grazed pasture, ryegrass, White Clover
110	85	2 miles North of Stratford, Midhurst		Lightly grazed pasture, abundant Yorkshire Fog, Brown Top, ryegrass etc.
111	86	6 miles South of New Plymouth		Grazed pasture
112	87	IWD 'Waireka' Res. Station , New Plymouth		Hedgerow plants
113		Andy Hills, Owhango		Railway side, unploughed
114	88	Andy Hills Owhango		Pasture, ploughed
115		Atawhai Heights, Palmerston North		Housing development area,previously poor pasture
116	89	Holden Station, Mckensie country		Grazed Tussock grassland (semiarid)
117	90	Birkes Pass, Mckensie country		Roadside, (semi-swamp), Craige burn soil
118	91	Black Birch, Marlborough	4800'	Sandy loam
119	92	Katahu Frest, (Fairlie-Geraldine Highway)		Grazed pasture & adjacent roadside near waterway
120	93	6 miles East of Woodville		Grazed hill pasture, scattered plants, "watergrass" dominant
121		Opapa (near L. Pou Kawa)		Roadside on dry hill country
122	94	Wairoa		Sheep grazed pasture mainly paspalum/Yorkshire Fog isolated heads
123	95	Captain Cook Statue, Gisborne		Pasture highly grazed, scattered plants, Yorkshire Fog with Creeping Fog, Kikuyu, Paspulum
124	96	Gray's Bush, Gisborne (6 miles North-West of City)		Pasture highly grazed, scattered plants, Yorkshire Fog with Creeping Fog, damp site

Collection No.	Group No.	Location	Altitude	Habitat
125	97	Waioeka Gorge (28 miles S. of Opotiki)		Ungrazed, dense
126	98	Waioeka Road (8 miles S of Opotiki)		Ungrazed, Moderate dense
127	99	Waioeka River, Opotiki Park		Mown area, Paspalum and Phalaris
128	100	Ohope beach, Whakatane		Ungrazed on sandy soil, dense large plants
129	101	Ohope beach (hill slope) Whakatane		Moderate short sheep/cattle pasture, plentiful Yorkshire Fog, White Clover & poor grasses
130	102	Rongitaikei Plains, 2½ miles North of Edgecumbe)		Roadside, tall Fescue & Paspalum
131	103	Edgecumbe (near factory)		Roadside, in dense Paspalum
132	104	Edgecumbe (near factory)		Grazed Paspalum pasture, isolated plants
133		2 mile Edgecumbe Whakatane		Roadside (occassionally grazed by cattle), Pumice area, Paspulum
134		Rotorua		Vaccant lot, dense large Yorkshire Fog with Fescue and weeds
135	105	Rotorua- 6 miles towards Paradise Valley Springs		Fertile pasture abundant Yorkshire Fog with ryegrass & White Clover, grazed by cattle
136	106	Rotorua - 5 miles on Lake Okereka Loop Road		Sheep pasture on Pumice, White Clover, ryegrass, Creeping Fog & Brown Top
137	107	Rotorua - 6 miles between Blue and Green Lakes		Pinus radiata forest fringe, large plants (do not grown in forest)
138	108	Rotorua - 17 miles Waimangu Thermal Valley		Pathway in native bush, abundant plants
139	109	Ruaumpoko's Throat, Waimangu, 17 miles Rotorua		Plants bordering on boiling lake
140	110	Taupo-Turanai (midway)		Rest area abundant Yorkshire Fog, White Clover, Pairie grass
141	111	Desert Road, 15 miles N. of Waiurou		Roadside
142	112	Taihape (Hautapu River)		Caravon park on river bank, large plants, Creeping Fog Fescue, Clovers, Weeds, ryegrass
143	113	Wanganui (10 miles South of)		Hill pasture, dense, Yorkshire Fog but few seed heads, Brown Top , ryegrass, dogstail, White Clover, Creeping Fog

Collection No.	Group No.	Location	Altitude	Habitat
144	114	Makirikiri, 12 miles North of Wanganui		Old sheep and horse paddock on flat, Brown Top dogstail and thistles
145	115	Otoko, 24 miles North of Wanganui		Sheep pasture, localized dense
146	116	Kakatahi (20 miles South of Raetihi)		Laxly grazed cattle pasture, Paspulum, White Clover, Brown Top, some Creeping Fog, Timothy Yorkshire Fog, dense
147	117	Oreore (10 miles South of Raetihi)		Fairly closely grazed pasture, ryegrass, White Clover, moderate Yorkshire Fog
148	118	Horopito (12 miles North of Raetihi)		Laxly grazed sheep pasture, abundant Yorkdhire Fog, Brown Top with White Clover
149	119	Kuratau Junction (Lake Taupo)		Heavily stocked sheep pasture, Tussock, improved spp;, White Clover, abundant Yorkshire Fog
150	120	Kuratau Junction, 7 miles North of Lake Taupo		Poorly grazed sheep pasture, almost pure stand Yorkshire Fog
151	121	Kuratau Junction, 15 miles North of Lake Taupo		Mekium grazed sheep pasture predominantly chewing Fescue, ryegrass and Yorkshire Fog
152	122	Kuratau Junction, 24 miles North of Lake Taupo		New pasture, ryegrass, Creeping Fog, White Clover, some Red Clover, Volunteer Yorkshire Fog, laxly grazed
153	123	Taupo (23 miles West of)		Laxly grazed sheep pasture, predominantly ryegrass, some Brown Top, White Clover, scattered plants
154	124	Taupo (11 miles West of)		Laxly grazed pasture, Creeping Fog, Yorkshire Fog, White Clover, some Brown Top
155	125	Rangitaiki (½ mile West of)	2500'	Short pasture, Yorkshire Fog dominant, White Clover, some Creeping Fog
156	126	Rangitaike, on plateau	2400'	Tussock grasslands, closely grazed (sheep) Creeping Fog, ryegrass, White Clover, abundant Yorkshire Fog
157	127	Turangakumu, Central North Island Hill country	2000'	Rough pasture, some improved grasses
158	128	Titiokura, Central North Island Hill country	2000'	Laxly grazed, sheep pasture, White Clover, Yorkshire Fog, few improved species

Collection No.	Group No.	Location	Altitude	Habitat
159		Eskdale (17 miles West of Napier)		Sheep pasture, Brown Top, native Tussock, scatter Yorkshire Fog, some Creeping Fog on ridge
160	129	Eskdale (17 miles West of Napier)		Sheep pasture (short) Couch/Paspulum on creek flat
161	130	Rissington (5 miles South-east of)		Closely grazed sheep/cattle pasture, sweet vernal, Brown Top, White Clover, moderate Yorkshire Fog, (dry)
162	131	Rissington (5 miles South-east of)		Closely grazed sheep/cattle pasture, Creeping Fog, ryegrass, Brown Top, abundant Yorkshire Fog, (moist)
163	132	Rissington (5 miles South-east of)		Cattle/sheep pasture, Paspulum reeds, sedge, Couch, abundant Yorkshire Fog, (swamp)
164		Maraekakaho (6½ miles South-west of)		Laxly grazed cattle pasture, ryegrass dominant, Brown Top White Clover, Creeping Fog, few Yorkshire Fog seed heads
165	133	Maraekakaho (6½ miles South-west of)		Stream bank, predominantly Creeping Fog, some Paspulum, few Yorkshire Fog seedheads
166	134	Ohaupo (6 miles on Cambridge Road)		Laxly grazed cattle pasture, predominantly Timothy, Creeping Fog, Yorkshire Fog, White Clover
167	135	Ohaupo (4 miles on Cambridge Road)		Well grazed dairy pasture, predominantly Paspulum, Timothy, Creeping Fog, ryegrass, isolated Yorkshire Fog
168		Rukuhia Swamp (1 mile West of Rukuhia)		Dominantly Yorkshire Fog, Brown Top, heavy peat few Yorkshire Fog seed heads
169	136	Rukuhia Swamp edge (near Ngahinapauri)		Cattle pasture, predominantly Paspulum, some Timothy, short Yorkshire Fog
170	137	Spain		
171	138	Crookwell		
172	139	Kuripapange (Gentle Annie Road)	2800'	Laxly grazed sheep/cattle pasture, Yorkshire Fog common, White Clover, Brown Top, Creeping Fog
173	140	38 miles from Taihape (Gentle Annie Road)	3100'	Improved pasture, abundant Yorkshire Fog, Creeping Fog, White Clover & Brown Top, medium grazing

Collection No.	Group No.	Location	Altitude	Habitat
174	141	Erewhon Station, 24 miles from Taihape (Gentle Annie Road)	2400'	Quite closely grazed, abundant Yorkshire Fog, Brown Top dominant, White Clover
175		Crail Bay, Pelorous sounds	0'	Garden, and sharf area
176	142	"Dundee" NSW		
177	143	"Colencoe", NSW (15 miles South-East Glen Innes)		
178	144	Massey University		Massey Basyn
179	145	Lincoln - Coes Ford		Roadside
180	146	Lincoln - Coes Ford		Roadside
181	147	Green Ford, Cantabury		Roadside
182	148	Lake Ellesmere, Cantabury		Roadside
183	149	Junction Okaihau - Kerikeri and Waimate North - Kabo Roads		Grazed pasture (Paspalum, Creeping Fog White Clover, Red Clover,) scattered fog
184	150	No. 10 highway, South of Kaeo		Grazed pasture (Paspalum, Azonopus) Crested dogstail etc.
185	151	No. 1 highway, 16 miles North of Okaihau (Mangamuka)		Grazed pasture (Paspalum, Creeping Fog, Axonopus and Clovers
186		10 miles South Okaihau towards Maungatapere		Grazed pasture (ryegrass, Paspalum, Axonopus, Clovers
187	152	Balclutha ; Inchclutha soil type(very fertile)		Continuouly grazed pasture, Tiller sample
188	153	Invermay Agricultural Research Station		Old, heavily -grazed, heavily -fertilized sheep pasture
189	154	1 mile West of Brynderwyn turnoff, Northland		Dairy pasture, isolated fog, well grazed
190	155	2 miles South Ruawai Flat, Northland		Well-grazed dairy pasture, Sparse Fog
191	156	1 mile East Mamaranui, Northland		Well-grazed dairy pasture, Sparse Fog
192	157	Top of Bombay Hill (Auckland side), Red Ash Soil		Closely-grazed dairy pasture, Sparse Fog
193	158	7 miles East Maramarua, Sunny Hill Side		Laxly-grazed sheep pasture, abundant Fog

Collection No.	Group No.	Location	Altitude	Habitat
194	159	1 mile East Turua (Hauraki plains)		Laxly-grazed cattle pasture, moderate fog
195	160	Grassland substation, Gore	230'	Pasture, ryegrass, Brown Top, Poa spp.
196		Moa Flat, South-West Otago	1230'	Pasture, sown 1958, Cocksfoot, ryegrass, goosegrass
197		Taupiri Road, Hodgkinsom's farm, Waikato	500'	Old pasture
198		Matakana, Prospect Bay, Kisslling's farm,	400'	Old pasture, (ryegrass, White Clover, Paspalum Kikuyu)
199		Puhatotara, land development block, Waipapa (Northland)	600'	Old pasture (ryegrass, White Clover, Yorkshire Fog, Paspalum, Cocksfoot)
200		Grassland division, Kaikara		Occasional Yorkshire Fog in 4 years old Ryegrass/White Clover, Paspalum/Ryegrass/White Clover, Kikuyu/Ryegrass.
201		Ohura		Roadside on town outskirts

LIST SYMBOL/MANDIS

DATE 01/17/78 TIME IS 14:34

SYSTEM/DUMPALL VERSION 2.9.170

LASTRECORD = 706
MAXRECSIZEIN = 14 BLOCKSIZEIN = 420
*** EBCDIC *** UNITS=WORDS

```
C*                                                                   0000100
C                                                                    0000200
      FILE  10 = FILE10, UNIT = READER                               0000300
      FILE  20 = FILE20, UNIT = DISK                                 0000400
      FILE  30 = FILE30, UNIT = DISK                                 0000500
$ SET AUTOBIND                                                       0000600
$ BIND = FROM CODON/=  ;                                             0000700
C*                                                                   0000800
C*                                                                   0000900
C*                 MANDIS              MANDIS              MANDIS    0001000
C*                                                                   0001100
C*    SUBPROGRAMS NEEDED: DARRAY, SMPRIN, DMPRIN, DMINV, PRBF, SIGNIF, 0001200
C*                        TRANSF, NROOT(EIGEN)                       0001300
C*     REFERENCE: COOLEY, W. W. & LOHNES, P. R. (1971)               0001400
C*                MULTIVARIATE DATA ANALYSIS.                        0001500
C*                                                                   0001600
      REAL MEAN, N1, N2, DIF                                         0001700
      INTEGER P, Q, TNP, TRANSX                                      0001800
      DOUBLE PRECISION DET, DETPI, DETW, DETB, DETD, M1, M2, BOXSM, XL, 0001900
     2                 YL, CPA, CPB, CPBA, CPC, CPCA, CPD, CPDA, T, Z 0002000
C*                                                                   0002100
      DIMENSION X(5,5,10,15), XIN(5,5,10,15), SUMT(15), SUM(15),     0002200
     1          MEAN(160,15), SD(15), CPA(15,15), CPB(15,15), CPCA(225),0002300
     2          CPC(15,15),CPD(15,15), T(15),U(15), V(15),W(15),S(15), 0002400
     3          Y(15), Z(15), KC(15,5), DATRAN(15,5), TRANSX(15,5),  0002500
     4         CV(15), KH(13), FMT1(16),MISS(15), LL(15), MM(15),    0002600
     5         CPBA(225), CPDA(225), CENT(160,15), FMTOUT(16)        0002700
      EQUIVALENCE (CPB,CPBA),(CPC,CPCA), (CPD,CPDA), (X,XIN)         0002800
C*                                                                   0002900
C*                                                                   0003000
C*     INPUTS                                                        0003100
C*                                                                   0003200
    1 READ (5,5) (KH(I), I = 1,13 )                                  0003300
    5 FORMAT (13A6)                                                  0003400
      IF(KH(1).IS.KH(2)) GO TO 46                                    0003500
C*                                                                   0003600
C*                                                                   0003700
C*                                                                   0003800
C*    IOPT    = THE ANALYSIS OPTION                                  0003900
C*    IOPT    = 1 FOR MANOVA AND DISCRIMINANT ANALYSIS               0004000
C*    IOPT    = 2 FOR MANOVA                                         0004100
C*    IOPT    = 3 FOR DISCRIMINANT ANALYSIS                          0004200
C*                                                                   0004300
C*    IGROUP = THE OPTION FOR GROUP SUMMARY                          0004400
C*    IGROUP = 1 FOR            PRINTING OF EACH GROUP SUMMARY       0004500
C*    IGROUP = 2 FOR SUPRESSING THE PRINTING OF EACH GROUP SUMMARY   0004600
C*                                                                   0004700
C*                                                                   0004800
C*    IMEAN   = THE OPTION FOR GROUP MEAN VECTOR                     0004900
C*            = 1  FOR SAVING ON FILE 20                             0005000
```

```
C*                = 2  FOR NOT SAVING ON FILE
C*
C*    ICENT       = THE OPTION FOR GROUP DISCRIMINANT SPACES CENTROID
C*                = 1  FOR PRINTING AND SAVING ON FILE 30
C*                = 2  FOR PRINTING ONLY
C*                = 3  FOR SAVING ON FILE 30 ONLY
C*                = 4  FOR NOT OBTAINING THE CENTROID
C*
      READ (5, 10) IOPT, IGROUP, IMEAN, ICENT
   10 FORMAT (8I5)
      GO TO (11, 11, 500), IOPT
   11 READ (5,10) NG,NR, NY, NP, NC, NCUT, NF
      IF (NCOT.EQ.0) NCOT = 1
      DO 15 P = 1, NC
      DO 15 JJ = 1, NCOT
      DATRAN(P,JJ) = '    '
   15 TRANSX(P,JJ) = 0
      READ (5,20) ((KC(P,J), J = 1, 5), (DATRAN(P,JJ), TRANSX(P,JJ),
     1            JJ = 1, 5), P = 1, NC)
   20 FORMAT (5A6, A4, I6, A4, I6, A4, I6, A4, I6, A4, I6)
      READ (5,25)(FMT1(I), I =1, NF)
      READ (5,25) (FMTOUT(I), I = 1,NF)
   25 FORMAT (16A5)
C*
C*
C*
C*     OUTPUTS-----TITLE AND CHARACTERS
C*
   46 PRINT 47, KH
   47 FORMAT (1H1,19X, 80(1H*)/20X, 1H*, 78X, 1H*/20X, 1H*,13A6, 1H*/20X
     1            , 1H*, 78X, 1H*/20X, 80(1H*))
      IF(KH(1).IS.KH(2)) GO TO 999
      PRINT 22
   22 FORMAT (//20X, 20(1H$), 4X, 33HMULTIVARIATE ANALYSIS OF VARIANCE ,
     1            4X, 20(1H$))
      PRINT 48
   48 FORMAT (//14X, 10HCHARACTER /)
      PRINT 49, (P,(KC(P,J ), J = 1,5), (DATRAN(P,JJ), TRANSX(P,JJ),
     1            JJ = 1, 5), P = 1, NC )
   49 FORMAT (/10X, I2, 2X, 5A6, 3X, A4, I6, 3X, A4, I6, 3X, A4, I6, 3X,
     1             A4, I6, 3X, A4, I6 / )
C*
C*     INITIALISED ACCUMULATOR
C*
C*
      DO 50 P = 1,NC
      SUMT(P) = 0.0
      DO 50 Q = 1,NC
      CPA(P,Q) = 0.0
      CPD(P,Q) = 0.0
   50 CPB(P,Q) = 0.0
      M2 = 0.0
      FA1 = 0.0
      FA2 = 0.0
      DETPI = 1.0
      TNP = 0
      NGA = NG
C*
C*
C*     RECYCLE POINT FOR ANALYSIS OF EACH GROUP
C*
```

```
C*
      DO 100 I = 1,NG
      NPA = 0
      MISSG = 0
C*
C*     DATA INPUT, TRANSFORMATION AND COUNTING OF MISSING DATA.
C*      (ALL NEGATIVE INPUT WILL BE TREATED AS MISSING DATA.)
C*
      DO 38 P = 1,NC
   38 MISS(P) = 0
      DO 44 J = 1, NR
      DO 44 K = 1, NY
      DO 44 L = 1, NP
      READ (10, FMT1) (XIN (J,K,L,P), P = 1, NC)
      DO 43 P = 1, NC
      IF (XIN(J,K,L,P).GE.0) GO TO 40
   39 MISS(P) = MISS(P) + 1
      X(J,K,L,P) = 0
      GO TO 43
   40 DO 42 JJ = 1, NCOT
      AUX = TRANSX(P,JJ)
      KK = TRANSX (P,JJ)
      IF (KK.LT.31.AND.KK.NE.0) AUX = XIN(J,K,L,KK)
      X(J,K,L,P) = XIN(J,K,L,P)
      IF (DATRAN(P,JJ).NE.'    ')
     1X(J,K,L,P) = TRANSF (DATRAN(P,JJ), XIN(J,K,L,P), AUX)
      IF (NCOT.GT.1.AND.JJ.LT.NCOT) XIN(J,K,L,P) = X(J,K,L,P)
   42 CONTINUE
   43 CONTINUE
   44 CONTINUE
      DO 45 P = 1, NC - 1
      IF (MISS(P).NE.MISS(P + 1)) GO TO 70
   45 CONTINUE
      MISSG = MISS(1)
      NPA = NP * NR * NY - MISSG
      TNP = TNP + NPA
C*
C*
C*     CORRECTION FACTOR AND MEAN FOR EACH CHARACTER
C*
      DO 55 P = 1,NC
      SUM(P) = 0.0
      DO 53 J = 1,NR
      DO 53 K = 1,NY
      DO 53 L = 1,NP
   53 SUM(P) = SUM(P) + X(J,K,L,P)
      SUMT(P) = SUMT(P) + SUM(P)
   55 MEAN(I,P) = SUM(P) / NPA
C*
C*     SUMS OF CROSSPRODUCTS AND STANDARD DEVIATION
C*
      DO 66 P = 1,NC
      DO 65 Q = 1,NC
      CPC(P,Q) = 0.0
      DO 60 J = 1,NR
      DO 60 K = 1,NY
      DO 60 L = 1,NP
   60 CPC(P,Q) = CPC(P,Q) + X(J,K,L,P) * X(J,K,L,Q)
      CPA(P,Q) = CPA(P,Q) + CPC(P,Q)
      CPC(P,Q) = CPC(P,Q) - SUM(P) * SUM(Q)/ NPA
      CPB(P,Q) = CPB(P,Q)+ CPC(P,Q)
```

```
   65 CPC(P,Q) = CPC(P,Q) / (NPA - 1)
      SD(P) = DSQRT(CPC(P,P))
   66 CV(P) = SD(P) / MEAN(I,P)
C*
C*     FIND THE DETERMINANT
C*
      CALL DARRAY (2, NC, NC, 15, 15, CPCA, CPC)
      CALL DMINV  (CPCA, NC, DET , LL, MM)
      IF(DET.EQ.0) DET = 0.0000000001
C*
C*     FIRST SET UP OF BOX'S M CRITERION
C*
      M2 = M2 + (NPA - 1) * DLOG(DET)
      FA1 = FA1 + 1.0 / (NPA - 1)
      FA2 = FA2 + 1.0 / (NPA - 1) ** 2.0
C*
C*     FIRST SET UP FOR SEAL'S TEST CRITERION
C*
      DETPI = DETPI * DET ** ((NPA - 1) / 2.0)
C*
C*     OUTPUT
C*
      GO TO (70, 100,100), IGROUP
   70 PAGE = MOD(I,2)
      IF (PAGE.NE.0 ) GO TO 74
      PRINT 73, I
   73 FORMAT (40X, 10(1H*), 5X, 6HGROUP , I3, 5X, 10(1H*)/)
      GO TO 76
   74 PRINT 75 , I
   75 FORMAT (1H1, 39X, 10(1H*), 5X, 6HGROUP , I3, 5X , 10(1H*)/)
   76 DO 78 P = 1, NC - 1
   78 IF (MISS(P).NE.MISS(P + 1)) GO TO 93
      PRINT 80
   80 FORMAT (14X, 9HCHARACTER, 26X, 4HMEAN, 7X, 4HS.D., 5X, 4HC.V. /)
      PRINT 85, (P, (KC(P,J), J = 1, 5), MEAN(I,P), SD(P), CV(P),
     1          P = 1,NC)
   85 FORMAT (10X, I2, 2X, 5A6, 3X, F8.4, 2X, F9.4, 2X, F7.4)
      PRINT 90, DET, MISSG, NPA
   90 FORMAT(//14X, 24HDISPERSION DETERMINANT = , F20.12,
     1          5X, 29HNO. OF MISSING OBSERVATION = , I3//57X,
     2          35HNO. OF OBSERVATION IN THIS GROUP = ,I3/)
      GO TO 98
   93 PRINT 94
   94 FORMAT (///20X, 10(1H*), 61HUNEQUAL NO. OF MISSING CHARACTER FROM
     1SAME EXPERIMENTAL UNIT , 10(1H*)// 20X, 20(1H*), 31HANALYSIS FOR T
     2HIS GROUP OMITTED , 20(1H*))
      NGA = NGA - 1
   98 PRINT 99
   99 FORMAT (/10X, 100(1H*)/)
  100 CONTINUE
      IF (NGA.LE.1) GO TO 900
C*
C*    SAVE MEAN VECTORS ON FILE 20
C*
      GO TO (102, 104), IMEAN
  102 DO 103 I = 1,NG
  103 WRITE (20, FMTOUT) (MEAN(I,P), P = 1,NC)
      LOCK 20
  104 CONTINUE
C*
C*    IF NUMBER OF GROUP AFTER ADJUSTMENT .LE., THEN TERMINATE
```

```
C*
      IF (NGA.LE.1) GO TO 900
C*
C*     SET UP THE TOTAL SSCP MATRIX, MEAN AND STANDARD DEVIATION OF EACH
C*     CHARACTER OVER ALL GROUPS.
C*
      DO 105 P = 1,NC
      DO 105 Q = 1,NC
      CPA(P,Q) = CPA(P,Q) - SUMT(P) * SUMT(Q) / TNP
      CPD(P,Q) = CPA(P,Q)
  105 CPC(P,Q) = CPB(P,Q) / (TNP - NGA)
C*
C*    CPA AND CPD ARE NOW TOTAL SSCP MATRICES OVER ALL GROUPS
C*    CPC IS NOW WITHIN GROUP MSCP MATRIX OVER ALL GROUPS
C*
      DO 110 P = 1,NC
      SUMT(P) = SUMT(P) / TNP
      SD(P) = DSQRT(CPC(P,P))
  110 CV(P) = SD(P) / SUMT(P)
C*
C*     OUTPUT
C*
      PRINT 115, NGA
  115 FORMAT (1H1, 35(1H*), 3X, 33HMEANS AND STANDARD DEVIATIONS FOR,
     1       I3, 7H GROUPS, 3X, 35(1H*)/ 39X, 43(1H*)/ 39X, 43(1H*))
      NMG = NG - NGA
      IF (NMG) 900, 119,116
  116 PRINT 117, NGA, NMG
  117 FORMAT (//35(1H*), 3X, 33HTHIS POOLED ANALYSIS IS BASED ON , I3,
     17H GROUPS, 3X, 35(1H*)/ 35(1H*),13X, I3, 21H GROUPS BEING OMITTED,
     213X, 35(1H*), 3X, 42HUNEQUAL MISSING CHARACTERS FROM SAME
     3 PLANT, 4X, 35(1H*) )
  119 PRINT 118, TNP
  118 FORMAT (//43X, 30H OVERALL NO. OF OBSERVATION = , I6, 5X//)
      PRINT 80
      PRINT 85, (P,(KC(P,J),J = 1,5), SUMT(P), SD (P), CV (P), P =1,NC)
      PRINT 120, NGA
  120 FORMAT (1H1, 40(1H*), 3X, 22HTOTAL SSCP MATRIX FOR ,I3, 7H GROUPS,
     1       3X, 40(1H*)/ 44X, 32(1H*)/ 44X, 32(1H*))
      CALL DMPRIN (CPA, NC, NC, 'T-SSCP', 15, 10)
      PRINT 99
C*
C*     OBTAIN THE AMONG-GROUP SSCP MATRIX BY SUBTRACTING WITHIN-GROUP
C*     SSCP MATRIX FROM TOTAL SSCP MATRIX.
C*
      DO 135 P = 1, NC
      DO 135 Q = 1, NC
  135 CPD(P,Q) = CPA(P,Q) - CPB(P,Q)
C*
C*    CPC IS NOW THE POOLED WITHIN GROUP MSCP MATRIX
C*    CPB IS NOW THE POOLED WITHIN GROUP SSCP MATRIX
C*    CPD IS NOW THE AMONG GROUP SSCP MATRIX
C*    CPA IS NOW THE TOTAL SSCP MATRIX
C*
C*     OUTPUT
C*
      PRINT 140, NGA
  140 FORMAT  (1H1, 35(1H*), 5X , 28HAMONG-GROUP SSCP MATRIX FOR , I3,
     1       7H GROUPS, 5X, 36(1H*)/ 41X, 38(1H*)/ 41X, 38(1H*))
      CALL DMPRIN (CPD, NC, NC, 'A-SSCP', 15, 10)
      PRINT 99
```

```
      PRINT 150, NGA
  150 FORMAT  (1H1, 35(1H*) , 5X, 29HWITHIN-GROUP SSCP MATRIX FOR , I3,
     1        7H GROUPS, 5X, 35(1H*)/ 41X, 39(1H*)/41X, 39(1H*))
      CALL DMPRIN (CPB, NC, NC, 'W-SSCP', 15, 10)
      PRINT 99
C*
C*     FIND THE DETERMINANT OF POOLED WITHIN-GROUP MSCP MATRIX
C*
      CALL DARRAY (2, NC, NC, 15, 15, CPCA, CPC)
      CALL DMINV  (CPCA, NC, DETW, LL, MM)
      IF (DETW.EQ.0) DETW = 0.0000000001
C*
C*
C*     SET UP BOX'S M CRITERION
C*
C*    TEST THE HOMOGENEITY OF GROUP DISPERSION BY BOX'S M
C*
C*
      M1 = (TNP - NGA) * DLOG (DETW)
      BOXSM = M1 - M2
      N1 = (NGA - 1.0) * NC * (NC + 1.0) / 2.0
      A1 = (FA1 - (1.0 / (TNP - NGA))) * (2.0 * (NC ** 2.0) + (3.0 * NC)
     1     - 1.0) / (6.0 * (NGA - 1.0) * (NC + 1.0))
      A2 = (FA2 - (1.0 / (TNP - NGA) ** 2.0)) * ((NC - 1.0) * (NC + 2.0)
     1     ) / (6.0 * (NGA - 1.0))
      DIF = A2 - A1  ** 2
      IF (DIF) 160, 160, 165
  160 N2 = (N1 + 2) / (A1 ** 2 - A2)
      B1 = N2 /   (1 - A1 +  (2 / N2))
      F = (N2 * BOXSM) / (N1 * (B1 - BOXSM))
      GO TO 170
  165 N2 = (N1 + 2) / (A2 - A1 ** 2)
      B1 = N1 /   (1 - A1 - (N1 / N2))
      F = BOXSM / B1
  170 DF1 = N1
      DF2 = N2
      PF = PRBF (DF1, DF2, F)
      SF = SIGNIF(PF)
C*
C*     SET UP SEAL'S TEST CRITERION
C*    TEST THE HOMOGENEITY OF GROUP DISPERSION BY CHI-SQUARE
C*
      DETPI = DLOG(DETPI / (DETW ** ((TNP - NGA) / 2.0)))
      CHI2 = -2 * ((1.0 - A1) * DETPI)
      DFCHI2 =   (NGA - 1.0) * NC * (NC + 1.0) / 2.0
      PCHI2 = PRBF(DFCHI2, 1000.0, CHI2 / DFCHI2)
      SCHI2 = SIGNIF(PCHI2)
C*
C*     OUTPUT
C*
      PRINT 173
  173 FORMAT (1H1, 35(1H*), 5X, 23HTEST OF HOMOGENEITY OF ,
     1        17HGROUP DISPERSIONS ,4X, 35(1H*)/41X,41(1H*)/
     2        41X, 41(1H*))
      PRINT 175, DETW
  175 FORMAT (//10X,50HDETERMINANT OF POOLED WITHIN-GROUPS MSCP MATRIX =
     1        , F20.12//)
      PRINT 180, BOXSM, F, DF1, DF2, PF, SF, CHI2, DFCHI2, PCHI2, SCHI2
  180 FORMAT (///10X, 19HBOX'S M CRITERION = , F20.12 //21X, 8HF-TEST =,
     1        F9.5, 5X, 26HWITH DEGREES OF FREEDOM = , I6, 5H AND , I6//
     2        12X, 17HAND PROBABILITY = , F7.4, 5X, A4////10X,
```

```
     3         23HSEAL'S TEST CRITERION = , F15.5// 15X,
     4         18HIS CHI-SQUARE WITH , I6, 19H DEGREES OF FREEDOM // 16X,
     5         17HAND PROBABILITY = , F7.4, 5X, A4///)
      PRINT 99
C*
C*      UNIVARIATE F-TEST
C*
      DF1 = NGA - 1
      DF2 = TNP - NGA
      PRINT 185 , DF1 , DF2
  185 FORMAT (1H1, 30(1H*), 4X, 31HUNIVARIATE F-RATIOS, WITH DF1 = , I4,
     1         10H AND DF2 = , I6, 4X, 30(1H*)/ 35X, 51(1H*)/35X, 51(1H*))
      PRINT 190
  190 FORMAT (///5X, 8HVARIABLE, 24X, 75HAMONG M.S.  WITHIN M.S.  F-RATI
     10       PROBABILITY  SIGNIFICANCE   ETA SQUARE /)
      DO 195 P = 1, NC
      AMS = CPD(P,P) / DF1
      WMS = CPB(P,P) / DF2
      F = AMS / WMS
      PF = PRBF(DF1, DF2, F)
      SF = SIGNIF(PF)
      ETASQ = CPD(P,P) /(CPD(P,P) + CPB(P,P))
  195 PRINT 200, (KC(P,J), J = 1 , 5), AMS, WMS ,F, PF, SF, ETASQ
  200 FORMAT (5X, 5A6, 2X,F10.5, 2X, F10.5, 3X, F10.5, 5X, F7.4, 6X, A4,
     1         10X, F8.5)
      PRINT 99
C*
C*      TEST OF EQUALITY OF CENTROIDS
C*
C*      FIND DETERMINANTS OF POOLED WITHIN GROUP SSCP MATRIX (DETB) AND
C*      TOTAL SSCP MATRIX (DETD)
C*
      DO 201 P = 1, NC
      DO 201 Q = 1, NC
      CPC(P,Q) = 0.0
  201 CPC(P,Q) = CPB(P,Q)
      CALL DARRAY (2, NC, NC, 15, 15, CPCA, CPC)
      CALL DMINV  (CPCA, NC, DETB, LL, MM)
      DO 202 P = 1, NC
      DO 202 Q = 1, NC
      CPC(P,Q) = 0.0
  202 CPC(P,Q) = CPA(P,Q)
      CALL DARRAY (2, NC, NC, 15, 15, CPCA, CPC)
      CALL DMINV  (CPCA, NC, DETD, LL, MM)
      IF(DETD.EQ.0)  DETD = 0.0000000001
C*
      XL = DETB / DETD
      YL = 1 - XL
      PRINT 203
  203 FORMAT (1H1, 35(1H*), 5X, 31H TEST OF EQUALITY OF CENTROIDS , 5X,
     1         35(1H*)/ 41X, 31(1H*)/ 41X, 31(1H*))
      PRINT 205, XL, YL
  205 FORMAT (/// 37X, 17H WILK'S LAMBDA = ,F20.12 / 10X, 44HGENERALISED
     1 CORRELATION RATIO, ETA SQUARE = , F20.12)
      IF(NC - 2) 210, 210 , 220
  210 IF (NGA - 3) 215, 215, 220
  215 YL = XL
      F1 = 2.0
      F2 = TNP - 3.0
      GO TO 225
  220 SL = SQRT(((NC ** 2.0) * ((NGA - 1.0) ** 2.0) - 4.0) /((NC ** 2.0)
```

```
     1      + ((NGA - 1.0) ** 2.0) - 5.0))
            YL = XL ** (1.0 / SL)
            PL = (TNP - 1.0) - ((NC + NGA) / 2.0)
            QL = -((NC * (NGA - 1.0)) - 2.0) / 4.0
            RL = (NC * (NGA - 1.0)) /2.0
            F1 = 2.0 * RL
            F2 = (PL * SL) + (2.0 * QL)
        225 DF1 = F1
            DF2 = F2
            F = ((1.0 - YL) / YL) *  (F2 / F1)
            PF = PRBF(DF1 , DF2, F)
            SF = SIGNIF(PF)
            PRINT 230, F, DF1, DF2 , PF, SF
        230 FORMAT (///10X, 35H F-RATIO, OVERALL DISCRIMINATION, = , F9.5,
     1             26H WITH DEGREES OF FREEDOM = , I6, 4H AND, I9//27X,
     2             19H AND PROBABILITY = , F7.4, 5X, A4///)
            PRINT 99
      C
      C     DISCRIMINATE ANALYSIS
      C
      C
            GO TO (527, 910, 500), IOPT
      C
        500 CONTINUE
            READ (5,10) NG, NC, TNP, NF
            NGA = NG
            READ (5,25) (FMT1(I), I = 1, NF)
            READ (5, 25) (FMTOUT(I), I = 1,NF)
      C
      C     READ IN UPPER HALF OF TSSCP
      C
      C
            DO 505 P = 1, NC
        505 READ (5, FMT1) (CPA(P,Q), Q = P, NC)
      C
      C
      C     READ IN UPPER HALF OF WSSCP
      C
            DO 515 P = 1, NC
        515 READ (5, FMT1) (CPB(P,Q), Q = P, NC)
      C
      C     READ IN GROUP MEAN VECTORS AND GRAND MEAN VECTOR
      C
            DO 517 I = 1, NG
        517 READ (5, FMT1) (MEAN(I,P), P = 1, NC)
            READ (5, FMT1) (SUMT(P), P = 1,NC)
      C
      C     COMPLETE THE OTHER HALF OF TSSCP & WSSCP
      C
            DO 520 P = 1, NC
            DO 520 Q = P, NC
            CPA(Q,P) = CPA(P,Q)
        520 CPB(Q,P) = CPB(P,Q)
      C
      C     COMPUTE A-SSCP
      C
            DO 525 P = 1, NC
            DO 525 Q = 1, NC
        525 CPD(P,Q) = CPA(P,Q) - CPB(P,Q)
      C
      C     OUTPUTS-----TITLES
```

```
C
      PRINT 47, KH
      PRINT 526
  526 FORMAT (//20X, 25(1H$), 5X, 21HDISCRIMINANT ANALYSIS, 5X, 25(1H$))
      GO TO 529
  527 PRINT 528
  528 FORMAT (1H1)
      PRINT 526
  529 CONTINUE
      PRINT 531, NGA, NC, TNP
  531 FORMAT (////20X, 'NO. OF GROUP = ', I3, 3X, 'NO. OF CHARACTER = ',
     1       I3, 3X, 'TOTAL NO. OF OBSERVATION = ', I6 )
C
      IF (NC - NGA) 530, 535, 535
  530 NCA = NC
      GO TO 540
  535 NCA = NGA - 1
  540 CONTINUE
      CALL DARRAY (2, NC, NC, 15, 15, CPBA, CPB)
      CALL DARRAY (2, NC, NC, 15, 15, CPCA, CPC)
      CALL DARRAY (2, NC, NC, 15, 15, CPDA, CPD)
      CALL NROOT  (NC, CPDA, CPBA, T, CPCA)
      CALL DARRAY (1, NC, NC, 15, 15, CPBA, CPB)
      CALL DARRAY (1, NC, NC, 15, 15, CPCA, CPC)
      CALL DARRAY (1, NC, NC, 15, 15, CPDA, CPD)
C
C     CPA NOW CONTAINS T-SSCP MATRIX
C
C
C     CPC NOW CONTAINS EIGENVECTORS OF W-SSCP INVERSE * A-SSCP
C      T  NOW CONTAINS EIGENVALUES  OF W-SSCP INVERSE * A-SSCP
C
C
      XL = 1.0
      TRACE = 0.0
      DO 545 P = 1, NCA
      U(P) = T(P) / (1.0 + T(P))
      V(P) = SQRT(U(P))
      W(P) = 1.0 / (1.0 + T(P))
      XL = XL * W(P)
  545 TRACE = TRACE + T(P)
      DO 550 P = 1, NCA
      Z(P) = 0.0
  550 Z(P) = 100.0 * (T(P) /TRACE)
C
      GO TO (575, 910, 553), IOPT
C
C     TEST THE EQUALITY OF CENTROIDS, VIA DIFFERENT APPROACH
C
  553 IF (NC - 2) 555, 555, 565
  555 IF (NGA - 3) 560, 560, 565
  560 YL = XL
      F1 = 2.0
      F2 = TNP - 3.0
      GO TO 570
  565 SL = SQRT(((NC * NC) * ((NGA - 1.0) ** 2) - 4.0) / ((NC * NC) +
     1((NGA - 1.0) ** 2) - 5.0))
      YL = XL ** (1.0 / SL)
      PL = (TNP - 1.0) - (( NC + NGA) / 2.0)
      QL = -((NC * (NGA - 1.0)) - 2.0) / 4.0
      RL = (NC * (NGA - 1.0)) / 2.0
```

```
      F1 = 2.0 * RL
      F2 = (PL * SL) + (2.0 * QL)
  570 DF1 = F1
      DF2 = F2
      F = ((1.0 - YL) / YL) * (F2 / F1)
      PF = PRBF(DF1, DF2 , F)
      SF = SIGNIF (PF)
      YL = 1.0 - XL
      PRINT 203
      PRINT 205, XL, YL
      PRINT 230, F, DF1, DF2, PF, SF
      PRINT 99
  575 PRINT 577
  577 FORMAT (//// 30(1H*), 5X, 45HCHI-SQUARE TEST WITH SUCCESSIVE ROOTS
     1 REMOVED, 5X, 30(1H*)/35X, 45(1H*)/35X, 45(1H*))
      PRINT 578
  578 FORMAT (// 5X, ' ROOTS     CANONICAL     R SQUARED      EIGEN-     CHI
     1-       D.F.     PROBA-    SIGNI-    LAMBDA   PERCENT' / 5X, 'REMOVED
     2R (ETA)     (ETA SQUARE)    VALUES    SQUARE              BILITY  FICANCE
     3        TRACE' /)
C
C     SET UP THE REVERSE W(P)PI IN S (P)
C
      P = NCA
      S(P+1) = 1.0
  580 S(P) = S(P+1) * W(P)
      P = P - 1
      IF (P) 585, 585, 580
  585 CONTINUE
      PL = (TNP - 1.0) - ((NC + NGA) / 2.0)
      DO 600 P = 1, NCA
      PT = P - 1
      NDF = (NC - PT) * (NGA - PT - 1.0)
      Y(P) = -PL * ALOG (S(P))
      PCHI2 = PRBF(NDF, 1000.0, Y(P)/NDF)
      SCHI2 = SIGNIF(PCHI2)
  600 PRINT 605, PT,V(P), U(P), T(P), Y(P), NDF, PCHI2, SCHI2,S(P), Z(P)
  605 FORMAT (/7X, I3, 5X, F7.4, 6X, F7.4, 2X, F11.5,F10.4, 1X, I4, 2X,
     1          F8.6, 4X, A4, 2X, F9.6, 1X, F8.4)
C
C
      DO610  P = 1, NCA
      DO610  Q = 1, NC
      CPD(P,Q) = 0.0
      DO 610 M = 1, NC
  610 CPD(P,Q) = CPD(P,Q) + CPC(M,P) * CPA(M,Q) / (TNP - 1.0)
      DO 617 P = 1, NCA
      DO 615 Q = 1, NCA
      CPB(P,Q) = 0.0
      DO 615 M = 1, NC
  615 CPB(P,Q) = CPB(P,Q) + CPD(P,M) * CPC(M,Q)
  617 T(P) = DSQRT(CPB(P,P))
C
C     CPD NOW CONTAINS (EIGENVECTORS OF W **(-1)* A) * T-MSCP
C     CPB NOW CONTAINS (EIGENVELTORS OF W **(-1)* A) ** 2 * T-MSCP
C
      DO 620 P = 1, NC
      DO 620 Q = 1, NCA
  620 CPC(P,Q) = CPC(P,Q) / T(Q)
C
      PRINT 625
```

```
  625 FORMAT (1H1, 39(1H*), 5X, 'DISCRIMINANT FUNCTIONS VECTORS', 5X,     00060000
     1        39(1H*),/ 45X, 30(1H*)/45X, 30(1H*))                         00060100
      PRINT 628                                                            00060200
  628 FORMAT (/28X, 'THE COEFFICIENTS FOR PRODUCING STANDARDIZED DISCRIM00060300
     2INANT SCORES',/33X, 'FROM GROUP DEVIATION VECTORS ARE LISTED AS CO00060400
     2LUMNS')                                                              00060500
      CALL DMPRIN (CPC, NC, NC, 'CHARA.', 15, 10)                          00060600
      DO 630 P = 1,NC                                                      00060700
  630 Z(P) = DSQRT(CPA(P,P) / (TNP - 1.0))                                 00060800
C                                                                          00060900
C     Z NOW CONTAINS TOTAL SAMPLE STANDARD DEVIATION                       00061000
      DO 635 P = 1, NC                                                     00061100
      DO 635 Q = 1, NC                                                     00061200
  635 CPA(P,Q) = CPA(P,Q) / (TNP * Z(P) * Z(Q))                            00061300
C                                                                          00061400
C     CPA NOW CONTAINS TOTAL SAMPLE CORRELATION MATRIX                     00061500
C                                                                          00061600
      DO 640 P = 1, NC                                                     00061700
      DO 640 Q = 1, NCA                                                    00061800
  640 CPB(P,Q) = CPC(P,Q) * Z(P)                                           00061900
      DO 645 P = 1, NC                                                     00062000
      DO 645 Q = 1, NCA                                                    00062100
      CPD(P,Q) = 0.0                                                       00062200
      DO 645 L = 1, NC                                                     00062300
  645 CPD(P,Q) =  CPD(P ,Q) + CPA(P,L) * CPB(L ,Q)                         00062400
C                                                                          00062500
C                                                                          00062600
      PRINT 650                                                            00062700
  650 FORMAT(1H1, 32(1H*),5X, 45H FACTOR STRUCTURE FOR DISCRIMINANT FUNC00062800
     1TIONS , 5X, 32(1H*)/ 38X, 45(1H*)/ 38X, 45(1H*))                     00062900
      PRINT 653                                                            00063000
  653 FORMAT (/17X, 'THE CORRELATIONS BETWEEN DISCRIMINANT SCORES (IN CO00063100
     1LUMNS) AND ORIGINAL SCORES (IN ROWS)')                               00063200
      CALL DMPRIN (CPD, NC, NC, 'CHARA.', 15, 10)                          00063300
C                                                                          00063400
      DO 655 Q = 1 ,NC                                                     00063500
  655 Z(Q) = 0.0                                                           00063600
      DO 665 P = 1, NCA                                                    00063700
      Y(P) = 0.0                                                           00063800
      DO 660 Q = 1, NC                                                     00063900
      Z(Q) = Z(Q) + CPD(Q,P) * CPD(Q,P)                                    00064000
  660 Y(P) = Y(P) + CPD(Q,P) * CPD(Q,P)                                    00064100
  665 Y(P) = 100.0 * (Y(P) / NC)                                           00064200
      PRINT 670, NCA                                                       00064300
  670 FORMAT (1H1, 33(1H*), 5X, 17HCOMMUNALITIES FOR, I5, 21H DISCRIMINA00064400
     1NT FACTORS, 5X, 31(1H*)/39X, 43(1H*), /39X, 43(1H*))                 00064500
      PRINT 673                                                            00064600
  673 FORMAT (/35X,'THE SUM OF SQUARES OF ROWS OF THE STRUCTURE MATRIX')00064700
      CALL DMPRIN (Z, NC, 1, 'CHARA.', 15, 10)                             00064800
      PRINT 99                                                             00064900
      PRINT 675                                                            00065000
  675 FORMAT (////1X, 15(1H*), 5X, 'PERCENTAGE OF TRACE OF TOTAL CORRELA00065100
     1TION MATRIX ACCOUNTED FOR BY EACH FUNCTION', 5X, 15(1H*) /21X,      00065200
     278(1H*), /21X, 78(1H*))                                              00065300
      PRINT 678                                                            00065400
  678 FORMAT (/31X, 'THE SUM OF SQUARES OF EACH COLUMN OF THE STRUCTURE 00065500
     1MATRIX',/36X, 'DIVIDED BY THE TRACE OF TOTAL CORRELATION MATRIX') 00065600
      CALL SMPRIN (Y, NC, 1, 'FUNCT.', 15, 10)                             00065700
      PRINT 99                                                             00065800
      PRINT 679                                                            00065900
  679 FORMAT (////1X, 19(1H*), 5X, 'STANDARD DEVIATION (BASED ON TOTAL',00066000
```

```
     1            ' MEAN SQUARE) FOR DISCRIMINANT SCORES', 5X, 19(1H*)/25X,
     2            71(1H*)/25X, 71(1H*))
      CALL DMPRIN (T, NC, 1, ' S.D. ', 15, 10)
C
C
C     OBTAIN GROUP CENTROID IN DISCRIMINANT SPACES
C
C
      GO TO (680, 680, 680, 730), ICENT
C
  680 CONTINUE
      DO 700 I = 1,NG
      DO 685 P = 1, NCA
      CENT(I,P) = 0.0
      DO 685 Q = 1, NC
      CENT (I,P) = CENT (I,P) + (MEAN(I,Q) - SUMT(Q)) * CPC(Q,P)
  685 CONTINUE
  700 CONTINUE
      GO TO (701, 701, 705, 730), ICENT
  701 PRINT 702, NG, NCA
  702 FORMAT  (1H1, 25(1H*), 3X, 12HCENTROIDS OF, I4, 10H GROUPS IN, I3,
     1         31H DIMENSIONAL DISCRIMINANT SPACE , 3X, 25(1H*)/ 29X,
     2         60(1H*)/ 29X, 60(1H*)/)
      CALL SMPRIN (CENT, NG, NCA, 'GROUP', 160, 10)
      GO TO (705, 730, 705, 730), ICENT
C
C     SAVE THE DISCRIMINANT CENTROID IN FILE 30 FOR *SIMMAT*
C
  705 CONTINUE
      DO 710 I = 1,NG
  710 WRITE (30, FMTOUT) (CENT(I,P), P = 1,NC)
      LOCK 30
  730 CONTINUE
C
C     TAIL PIECES
C
      GO TO 910
  900 PRINT 905
  905 FORMAT (/// 30X, 60(1H*)/30X, 60(1H*)/37X, 47HTOO MANY MISSING GRO
     1UP, POOLED ANALYSIS OMITTED/30X, 60(1H*)/30X, 60(1H*))
  910 PRINT 915
  915 FORMAT(/// 30X, 60(1H*)/30X, 60(1H*)//40X, 41H PROGRAM MANDIS BY
     1S.H.TEOW, AUGUST 1976 // 30X, 60(1H*)/30X, 60(1H*))
  920 CONTINUE
      GO TO 1
  999 CALL EXIT
      END
```

```
LIST SYMBOL/TRANSF

DATE 01/17/78          TIME IS 14:35

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =      233
MAXRECSIZEIN =    14     BLOCKSIZEIN =   420
*** EBCDIC ***           UNITS=WORDS

C
C       FUNCTION TRANSF
C       TRANSFORMATIONS AND MANIPULATIONS OF DATA
C       CHOICE SIGNALLED BY DATRAN
C
C
C       AVAILABLE TRANSFORMATIONS
C
C         1  LOGA   LOG10(X)
C         2  LOGN   LOGE(X)
C         3  SQRT   SQRT(X)
C         4  SQT+   SQRT(X+0.5)
C         5  ACSN   SIN**(-1) (SQRT(X))
C         6  RCIP   1/X
C         7  LGN+   LOGE(X+1)
C         8  CORR   0.5*(LOGE((1+X)/(1-X)))
C         9  EMPI   LOGE(X/(1-X))
C        10  LGIT   LOGE((X/Y0)/(1-(X/Y0)))
C        11  GOMP   B**X
C        12  CODE   X/K
C        13  CTR-   X-K
C        14  LOG+   LOG10(X+1)
C        15  RANK   NORMAL SCORES
C        16  CTR+   X+K
C        17  MULT   X * K
C        18  CMPL   K - X    (COMPLEMENT)
C
C       AVAILABLE MANIPULATIONS
C
C       501  MRAT   X/Y
C       502  MMDF   (X-Y)/2
C       503  MDRX   (X-Y)/X
C       504  MDRY   (X-Y)/Y
C       505  MMSM   (X+Y)/2
C       506  MSUM   X+Y
C       507  MDIF   X-Y
C       508  MMGO   (X*Y)**0.5
C       509  MDCL   LOGE(X-Y+25)
C       510  MPRD   X * Y
C
C       AVAILABLE DATA HANDLING
C
C       900  WIPE   REPLACE X WITH ZERO
C       901  RECV   SHIFT Y INTO X
C
C
C
C       CONTROL CARDS
C
C       DATRAN (4 COLUMNS) = MNEMONIC
```

```
C                 WHERE MNEMONIC= ONE OF THE PRECEDING LIST, ASSUMED TO BE A4.
C           TRANSX = SECOND INPUT FOR FUNCTION, OR BLANK
C                 THUS: (1)TRANSX .LT. 31 = INDEX OF X VARIATE USED AS
C                                          SECOND INPUT;
C                       (2)TRANSX.GT.30 = A CONSTANT FOR CODING, ETC,
C                                          ASSUMED TO BE IN DIGIT FORM
C                                          OF 6 INTEGERS, CORRESPONDING TO
C                                          REAL NUMBER OF FORMAT F7.4
C     ******************************************************************
C
      FUNCTION TRANSF(DATRAN,XX,XY)
      DIMENSION RANKIT(20,10)
      DATA((RANKIT(I,J),J=1,10),I=2,20)/0.564 ,0,0,0,0,0,0,0,0,
     1   0.864,0,0,0,0,0,0,0,0,  1.029,0.297,0,0,0,0,0 ,0,0,0,
     2 1.163,0.495,0,0,0,0 ,0 ,0,0,0,1.267,0.642,0.202,0,0,0,0,0 ,0,
     3 1.352,0.757,0.353,0,0,0 ,0,0,0,0,1.424,0.852,0.473,0.153,0 ,0,0,
     4 0,0,0 ,1.485,0.932,0.572,0.275,0,0 ,0,0 ,0,0,1.539,1.001 ,0.656,
     5 0.376,0.123,0 ,0,0,0,0 ,1.586,1.062 ,0.729 ,0.462,0.225,0,0,0,0,
     6 0,1.629,1.116,0.793 ,0.537,0.312,0.103 ,0,0 ,0 ,0,1.668,1.164,
     7 0.850 ,0.603,0.388,0.191,0,0,0,0 ,1.703,1.208,0.901,0.662,0.456,
     8 0.267,0.088,0,0,0,1.736,1.248,0.948,0.715,0.516,0.335,0.165,0 ,
     9 0 ,0,1.766,1.285,0.990,0.763,0.570,0.396,0.234,0.077,0 ,0 ,
     9 1.794,1.319,1.029,0.807,0.619,0.451,0.295,0.146 ,0,0,1.820,1.350,
     9 1.066,0.848,0.665,0.502,0.351,0.208,0.069,0 ,1.844,1.380,1.099 ,
     9 0.886,0.707,0.548,0.402,0.264,0.131,0,1.867 ,1.408,1.131,0.921,
     9 0.745,0.590,0.448,0.315,0.187,0.062 /
C
      IF(XX.EQ.0.0)XX=0.000001
C
C                                              TRANSFORMATIONS
C
      IF(DATRAN.EQ.'LOGA')GO TO 1
      GO TO 2
    1 TRANSF=ALOG10(XX)
      RETURN
    2 IF(DATRAN.EQ.'LOGN')GO TO 3
      GO TO 4
    3 TRANSF=ALOG(XX)
      RETURN
    4 IF(DATRAN.EQ.'SQRT')GO TO 5
      GO TO 6
    5 TRANSF=SQRT(XX)
      RETURN
    6 IF(DATRAN.EQ.'SQT+')GO TO 7
      GO TO 8
    7 TRANSF=SQRT(XX+0.5)
      RETURN
    8 IF(DATRAN.EQ.'ACSN')GO TO 9
      GO TO 10
    9 TRANSF=ARSIN(SQRT(XX))
      RETURN
   10 IF(DATRAN.EQ.'RCIP')GO TO 11
      GO TO 12
   11 TRANSF=1.0/XX
      RETURN
   12 IF(DATRAN.EQ.'LGN+')GO TO 13
      GO TO 14
   13 TRANSF=ALOG(XX+1.0)
      RETURN
   14 IF(DATRAN.EQ.'CORR')GO TO 15
      GO TO 16
```

```
   15 TRANSF=0.5*(ALOG((1.0+XX)/(1.0-XX)))
      RETURN
   16 IF(DATRAN.EQ.'EMPI')GO TO 17
      GO TO 18
   17 TRANSF=ALOG(XX/(1.0-XX))
      RETURN
   18 IF(DATRAN.EQ.'LGIT')GO TO 19
      GO TO 20
   19 XY=XY/10000
      TRANSF=ALOG((XX/XY)/(1.0-(XX/XY)))
      RETURN
   20 IF(DATRAN.EQ.'GOMP')GO TO 21
      GO TO 22
   21 XY=XY/10000
      TRANSF=XY**XX
      RETURN
   22 IF(DATRAN.EQ.'RANK')GO TO 23
      GO TO 24
   23 I=XY/10000
      IF(I.LT.2.OR.I.GT.20)GO TO 999
      J=XX
      IF(J.LT.1)J=1
      IF(J-I/2)901,901,902
  901 TRANSF=3.0-RANKIT(I,J)
      RETURN
  902 JJ=I-J+1
      IF(JJ.LT.1)JJ=1
      IF(JJ.GT.I)JJ=I
      TRANSF=RANKIT(I,JJ)+3.0
      RETURN
   24 IF(DATRAN.EQ.'CODE')GO TO 25
      GO TO 26
   25 XY=XY/10000
      TRANSF=XX/XY
      RETURN
   26 IF(DATRAN.EQ.'CTR-')GO TO 27
      GO TO 28
   27 XY=XY/10000
      TRANSF=XX-XY
      RETURN
   28 IF(DATRAN.EQ.'LOG+') GO TO 29
      GO TO 30
   29 TRANSF=ALOG10(XX+1 )
      RETURN
   30 IF(DATRAN.EQ.'CTR+')GO TO 31
      GO TO 32
   31 XY=XY/10000
      TRANSF=XX+XY
      RETURN
   32 IF(DATRAN.IS.'MULT')GO TO 33
      GO TO 34
   33 XY=XY/10000
      TRANSF=XX*XY
      RETURN
   34 IF(DATRAN.IS.'CMPL')GO TO 35
      GO TO 500
   35 XY=XY/10000
      TRANSF=XY-XX
      RETURN
C
C                                              MANIPULATIONS
```

```
      C
  500 IF(DATRAN.EQ.'MRAT')GO TO 501
      GO TO 502
  501 IF(XY.EQ.0.0)GO TO 999
      TRANSF=XX/XY
      RETURN
  502 IF(DATRAN.EQ.'MMDF')GO TO 503
      GO TO 504
  503 TRANSF=(XX-XY)/2
      RETURN
  504 IF(DATRAN.EQ.'MDRX')GO TO 505
      GO TO 506
  505 TRANSF=(XX-XY)/XX
      RETURN
  506 IF(DATRAN.EQ.'MDRY')GO TO 507
      GO TO 508
  507 IF(XY.EQ.0.0)GO TO 999
      TRANSF=(XX-XY)/XY
      RETURN
  508 IF(DATRAN.EQ.'MMSM')GO TO 509
      GO TO 510
  509 TRANSF=(XX+XY)/2
      RETURN
  510 IF(DATRAN.EQ.'MSUM')GO TO 511
      GO TO 512
  511 TRANSF=XX+XY
      RETURN
  512 IF(DATRAN.EQ.'MDIF')GO TO 513
      GO TO 514
  513 TRANSF=XX-XY
      RETURN
  514 IF(DATRAN.EQ.'MMGO')GO TO 515
      GO TO 516
  515 IF(XY.EQ.0.0)XY=0.000001
      TRANSF=SQRT(XX*XY)
      RETURN
  516 IF(DATRAN.EQ.'MDCL')GO TO 517
      GO TO 518
  517 TRANSF=ALOG(XX-XY+25.0)
      RETURN
  518 IF(DATRAN.IS.'MPRD')GO TO 519
      GO TO 800
  519 TRANSF=XX*XY
      RETURN
      C
      C                                     DATA SHIFTING
      C
  800 IF(DATRAN.EQ.'WIPE')GO TO 801
      GO TO 802
  801 TRANSF=0.0
      RETURN
  802 IF(DATRAN.EQ.'RECV')GO TO 803
      GO TO 999
  803 TRANSF=XY
      RETURN
      C
      C                                     UNTRANSFORMED DATA
      C
  999 TRANSF=XX
      DATRAN='ORIG'
      RETURN
```

2340000

END

234

```
LIST SYMBOL/SIGNIF
DATE 01/17/78          TIME IS 14:35
SYSTEM/DUMPALL VERSION  2.9.170
LASTRECORD =        21
MAXRECSIZEIN =    14    BLOCKSIZEIN =   420
*** EBCDIC ***          UNITS=WORDS

      FUNCTION SIGNIF(PROB)
C
C     SIGNIFICANCE SYMBOLS FOR PROBABILITY TESTS
C
      IF(PROB.LE.0.1)GO TO 5
      SIGNIF=' NS '
      GO TO 30
    5 IF(PROB.LE.0.05)GO TO 10
      SIGNIF='(NS)'
      GO TO 30
   10 IF(PROB.LE.0.01)GO TO 15
      SIGNIF=' *  '
      GO TO 30
   15 IF(PROB.LE.0.005)GO TO 20
      SIGNIF=' ** '
      GO TO 30
   20 IF(PROB.LE.0.001)GO TO 25
      SIGNIF=' ***'
      GO TO 30
   25 SIGNIF='****'
   30 RETURN
      END
```

```
LIST SYMBOL/DARRAY

DATE 01/17/78          TIME IS 14:35

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =       84
MAXRECSIZEIN =     14     BLOCKSIZEIN =   420
*** EBCDIC ***            UNITS=WORDS
```

```
C
C     ..................................................................
C
C        SUBROUTINE DARRAY
C
C        PURPOSE
C           CONVERT DATA ARRAY FROM SINGLE TO DOUBLE DIMENSION OR VICE
C           VERSA.  THIS SUBROUTINE IS USED TO LINK THE USER PROGRAM
C           WHICH HAS DOUBLE DIMENSION ARRAYS AND THE SSP SUBROUTINES
C           WHICH OPERATE ON ARRAYS OF DATA IN A VECTOR FASHION.
C
C        USAGE
C        CALL DARRAY (MODE, I, J, N, M, S, D)
C
C        DESCRIPTION OF PARAMETERS
C           MODE - CODE INDICATING TYPE OF CONVERSION
C                    1 - FROM SINGLE TO DOUBLE DIMENSION
C                    2 - FROM DOUBLE TO SINGLE DIMENSION
C           I    - NUMBER OF ROWS IN ACTUAL DATA MATRIX
C           J    - NUMBER OF COLUMNS IN ACTUAL DATA MATRIX
C           N    - NUMBER OF ROWS SPECIFIED FOR THE MATRIX D IN
C                  DIMENSION STATEMENT
C           M    - NUMBER OF COLUMNS SPECIFIED FOR THE MATRIX D IN
C                  DIMENSION STATEMENT
C           S    - IF MODE=1, THIS VECTOR CONTAINS, AS INPUT, A DATA
C                  MATRIX OF SIZE I BY J IN CONSECUTIVE LOCATIONS
C                  COLUMN-WISE.  IF MODE=2, IT CONTAINS A DATA MATRIX
C                  OF THE SAME SIZE AS OUTPUT.  THE LENGTH OF VECTOR S
C                  IS IJ, WHERE IJ=I*J.
C           D    - IF MODE=1, THIS MATRIX (N BY M) CONTAINS, AS OUTPUT,
C                  A DATA MATRIX OF SIZE I BY J IN FIRST I ROWS AND
C                  J COLUMNS.  IF MODE=2, IT CONTAINS A DATA MATRIX OF
C                  THE SAME SIZE AS INPUT.
C
C        REMARKS
C           VECTOR S CAN BE IN THE SAME LOCATION AS MATRIX D.  VECTOR S
C           IS REFERRED AS A MATRIX IN OTHER SSP ROUTINES, SINCE IT
C           CONTAINS A DATA MATRIX.
C           THIS SUBROUTINE CONVERTS ONLY GENERAL DATA MATRICES (STORAGE
C           MODE OF 0).
C
C        SUBROUTINES AND FUNCTION SUBROUTINES REQUIRED
C           NONE
C
C        METHOD
C           REFER TO THE DISCUSSION ON VARIABLE DATA SIZE IN THE SECTION
C           DESCRIBING OVERALL RULES FOR USAGE IN THIS MANUAL.
C
C     ..................................................................
C
```

```
      SUBROUTINE DARRAY (MODE, I, J, N, M, S, D)
      DOUBLE PRECISION S, D
      DIMENSION S(1), D(1)
C
      NI=N-I
C
C         TEST TYPE OF CONVERSION
C
      IF(MODE-1) 100, 100, 120
C
C         CONVERT FROM SINGLE TO DOUBLE DIMENSION
C
  100 IJ=I*J+1
      NM=N*J+1
      DO 110 K=1,J
      NM=NM-NI
      DO 110 L=1,I
      IJ=IJ-1
      NM=NM-1
  110 D(NM)=S(IJ)
      GO TO 140
C
C         CONVERT FROM DOUBLE TO SINGLE DIMENSION
C
  120 IJ=0
      NM=0
      DO 130 K=1,J
      DO 125 L=1,I
      IJ=IJ+1
      NM=NM+1
  125 S(IJ)=D(NM)
  130 NM=NM+NI
C
  140 RETURN
      END
```

```
LIST SYMBOL/PRBF
DATE 01/17/78          TIME IS 14:35
SYSTEM/DUMPALL VERSION  2.9.170
LASTRECORD =       22
MAXRECSIZEIN =     14    BLOCKSIZEIN =   420
*** EBCDIC ***           UNITS=WORDS

      FUNCTION PRBF(DA,DB,FR)
C
C     PROBABILITY OF F, CHI, T, AND Z
C
      PRBF=1.0
      IF(DA*DB*FR.EQ.0.0)RETURN
      IF(FR.LT.1.0)GO TO 5
      A=DA
      B=DB
      FRA=FR
      GO TO 10
    5 A=DB
      B=DA
      FRA=1.0/FR
   10 AA=2.0/(9.0*A)
      BB=2.0/(9.0*B)
      ZA=ABS(((1.0-BB)*FRA**0.333333-1.0+AA)/SQRT(BB*FRA**0.666667+AA))
      IF(B.LT.4.0)ZA=ZA*(1.0+0.08*ZA**4/B**3)
     0PRBF=0.5/(1.0+ZA*(0.196854+ZA*(0.115194+ZA*(0.000344+ZA*0.019527
     1))))**4
      IF(FR.LT.1.0)PRBF=1.0-PRBF
      RETURN
      END
```

LIST SYMBOL/SMPRIN

DATE 01/17/78   TIME IS 14:35

SYSTEM/DUMPALL VERSION 2.9.170

```
LASTRECORD =       47
MAXRECSIZEIN =    14    BLOCKSIZEIN =  420
*** EBCDIC ***          UNITS=WORDS
```

```
C
      SUBROUTINE SMPRIN  (X, N, M, KH, ND, L)
      DOUBLE PRECISION FMT, T
      DIMENSION X(ND, M), FMT(2) , T( 9)
      DATA T/ ' 10I11) ',' 10F11.1) ',' 10F11.2) ',' 10F11.3) ',
     1        ' 10F11.4) ',' 10F11.5) ',' 10F11.6) ',' 10F11.7) ',
     2        ' 1P10E11.3) '/, FMT(1) /' (1H+, 10X,'/
C
      L = L + 1
      IF (L-9)8, 8, 2
    2 CONTINUE
      XM = ABS(X(1,1))
      DO 3 I = 1,ND
      DO 3 J = 1,M
      XM = AMAX1( ABS(X(I,J)), XM)
    3 CONTINUE
      L = 1
      S = 10000000
    4 CONTINUE
      IF (XM.GE.S) GO TO 8
      S = S / 10
      L = L+ 1
      GO TO (4, 4, 4 , 4, 4, 4, 4, 6, 8), L
    6 CONTINUE
      S = S / 100
      GO TO 4
    8 CONTINUE
      FMT (2) = T(L)
      IF (M.GT.1) GO TO 20
      PRINT 15
      DO 10 I = 1, N, 10
      J = MINO (I + 9, N)
      PRINT 5, KH, (K, K = 1,J)
    5 FORMAT (/A7, 10I11)
      PRINT 15
   15 FORMAT (10X)
   10 PRINT FMT,(X(K,1) , K = I, J)
      RETURN
   20 DO 25 K = 1,M, 10
      PRINT 5
      L = MINO (K+9, M)
      PRINT 5, KH, (J, J = K, L)
      DO 25 I = 1,N
      PRINT 30, I
   30 FORMAT (/I6, 4X)
   25 PRINT FMT, (X(I,J), J = K,L)
      RETURN
      END
```

LIST SYMBOL/DMPRIN

DATE 01/17/78   TIME IS 14:35

SYSTEM/DUMPALL VERSION 2.9.170

LASTRECORD =   47
MAXRECSIZEIN =   14   BLOCKSIZEIN =   420
*** EBCDIC ***   UNITS=WORDS

```
C
      SUBROUTINE DMPRIN (X, N, M, KH, ND, L)
      DOUBLE PRECISION X, FMT, XM, T
      DIMENSION X(ND, M), FMT(2) , T( 9)
      DATA T/ ' 10I11) ',' 10F11.1) ',' 10F11.2) ',' 10F11.3) ',
     1        ' 10F11.4) ',' 10F11.5) ',' 10F11.6) ',' 10F11.7) ',
     2        ' 1P10E11.3) '/, FMT(1) /' (1H+, 10X,'/
C
      L = L + 1
      IF (L-9)8, 8, 2
    2 CONTINUE
       XM = DABS(X(1,1))
      DO 3 I = 1,ND
      DO 3 J = 1,M
       XM = DMAX1( DABS(X(I,J)), XM)
    3 CONTINUE
      L = 1
      S = 10000000
    4 CONTINUE
      IF (XM.GE.S) GO TO 8
      S = S / 10
      L = L+ 1
      GO TO (4, 4, 4 , 4, 4, 4, 4, 6, 8), L
    6 CONTINUE
      S = S / 100
      GO TO 4
    8 CONTINUE
      FMT (2) = T(L)
      IF (M.GT.1) GO TO 20
      PRINT 15
      DO 10 I = 1, N, 10
      J = MIN0 (I + 9, N)
      PRINT 5, KH, (K, K = I,J)
    5 FORMAT (/A7, 10I11)
      PRINT 15
   15 FORMAT (10X)
   10 PRINT FMT,(X(K,1) , K = I, J)
      RETURN
   20 DO 25 K = 1,M, 10
      PRINT 5
      L = MIN0 (K+9, M)
      PRINT 5, KH, (J, J = K, L)
      DO 25 I = 1,N
      PRINT 30, I
   30 FORMAT (/I6, 4X)
   25 PRINT FMT, (X(I,J), J = K,L)
      RETURN
      END
```

```
LIST SYMBOL/EIGEN

DATE 01/17/78          TIME IS 14:35

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =       176
MAXRECSIZEIN =     14    BLOCKSIZEIN =   420
*** EBCDIC ***           UNITS=WORDS
```

```
C
C     ...............................................................
C
C        SUBROUTINE EIGEN
C
C           PURPOSE
C              COMPUTE EIGENVALUES AND EIGENVECTORS OF A REAL SYMMETRIC
C              MATRIX
C
C           USAGE
C              CALL EIGEN(A,R,N,MV)
C
C           DESCRIPTION OF PARAMETERS
C              A - ORIGINAL MATRIX (SYMMETRIC), DESTROYED IN COMPUTATION.
C                  RESULTANT EIGENVALUES ARE DEVELOPED IN DIAGONAL OF
C                  MATRIX A IN DESCENDING ORDER.
C              R - RESULTANT MATRIX OF EIGENVECTORS (STORED COLUMNWISE,
C                  IN SAME SEQUENCE AS EIGENVALUES)
C              N - ORDER OF MATRICES A AND R
C              MV- INPUT CODE
C                      0   COMPUTE EIGENVALUES AND EIGENVECTORS
C                      1   COMPUTE EIGENVALUES ONLY (R NEED NOT BE
C                          DIMENSIONED BUT MUST STILL APPEAR IN CALLING
C                          SEQUENCE)
C
C           REMARKS
C              ORIGINAL MATRIX A MUST BE REAL SYMMETRIC (STORAGE MODE=1)
C              MATRIX A CANNOT BE IN THE SAME LOCATION AS MATRIX R
C
C           SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED
C              NONE
C
C           METHOD
C              DIAGONALIZATION METHOD ORIGINATED BY JACOBI AND ADAPTED
C              BY VON NEUMANN FOR LARGE COMPUTERS AS FOUND IN 'MATHEMATICAL
C              METHODS FOR DIGITAL COMPUTERS', EDITED BY A. RALSTON AND
C              H.S. WILF, JOHN WILEY AND SONS, NEW YORK, 1962, CHAPTER 7
C
C     ...............................................................
C
      SUBROUTINE EIGEN(A,R,N,MV)
      DIMENSION A(1),R(1)
      DOUBLE PRECISION A, R, ANORM, ANRMX, THR, X, Y, SINX, SINX2, COSX,
     1                 COSX2, SINCS, RANGE
C
C        GENERATE IDENTITY MATRIX
C
    5 RANGE = 1.0D-12
      IF(MV-1) 10,25,10
```

```
   10 IQ=-N
      DO 20 J=1,N
      IQ=IQ+N
      DO 20 I=1,N
      IJ=IQ+I
      R(IJ)=0.0
      IF(I-J) 20,15,20
   15 R(IJ)=1.0
   20 CONTINUE
C
C        COMPUTE INITIAL AND FINAL NORMS (ANORM AND ANORMX)
C
   25 ANORM=0.0
      DO 35 I=1,N
      DO 35 J=I,N
      IF(I-J) 30,35,30
   30 IA=I+(J*J-J)/2
      ANORM=ANORM+A(IA)*A(IA)
   35 CONTINUE
      IF(ANORM) 165,165,40
   40 ANORM =DSQRT(2 * ANORM)
      ANRMX=ANORM*RANGE/FLOAT(N)
C
C        INITIALIZE INDICATORS AND COMPUTE THRESHOLD, THR
C
      IND=0
      THR=ANORM
   45 THR=THR/FLOAT(N)
   50 L=1
   55 M=L+1
C
C        COMPUTE SIN AND COS
C
   60 MQ=(M*M-M)/2
      LQ=(L*L-L)/2
      LM=L+MQ
   62 IF (DABS(A(LM)) - THR) 130, 65, 65
   65 IND=1
      LL=L+LQ
      MM=M+MQ
      X=0.5*(A(LL)-A(MM))
   68 Y = -A(LM) / DSQRT(A(LM) * A(LM) + X * X)
      IF(X) 70,75,75
   70 Y=-Y
   75 SINX = Y / DSQRT(2.0 * (1.0 + (DSQRT(1.0 - Y * Y))))
      SINX2=SINX*SINX
   78 COSX = DSQRT (1.0 - SINX2)
      COSX2=COSX*COSX
      SINCS =SINX*COSX
C
C        ROTATE L AND M COLUMNS
C
      ILQ=N*(L-1)
      IMQ=N*(M-1)
      DO 125 I=1,N
      IQ=(I*I-I)/2
      IF(I-L) 80,115,80
   80 IF(I-M) 85,115,90
   85 IM=I+MQ
      GO TO 95
   90 IM=M+IQ
```

```
   95 IF(I-L) 100,105,105
  100 IL=I+LQ
      GO TO 110
  105 IL=L+IQ
  110 X=A(IL)*COSX-A(IM)*SINX
      A(IM)=A(IL)*SINX+A(IM)*COSX
      A(IL)=X
  115 IF(MV-1) 120,125,120
  120 ILR=ILQ+I
      IMR=IMQ+I
      X=R(ILR)*COSX-R(IMR)*SINX
      R(IMR)=R(ILR)*SINX+R(IMR)*COSX
      R(ILR)=X
  125 CONTINUE
      X=2.0*A(LM)*SINCS
      Y=A(LL)*COSX2+A(MM)*SINX2-X
      X=A(LL)*SINX2+A(MM)*COSX2+X
      A(LM)=(A(LL)-A(MM))*SINCS+A(LM)*(COSX2-SINX2)
      A(LL)=Y
      A(MM)=X
C
C         TESTS FOR COMPLETION
C
C         TEST FOR M = LAST COLUMN
C
  130 IF(M-N) 135,140,135
  135 M=M+1
      GO TO 60
C
C         TEST FOR L = SECOND FROM LAST COLUMN
C
  140 IF(L-(N-1)) 145,150,145
  145 L=L+1
      GO TO 55
  150 IF(IND-1) 160,155,160
  155 IND=0
      GO TO 50
C
C         COMPARE THRESHOLD WITH FINAL NORM
C
  160 IF(THR-ANRMX) 165,165,45
C
C         SORT EIGENVALUES AND EIGENVECTORS
C
  165 IQ=-N
      DO 185 I=1,N
      IQ=IQ+N
      LL=I+(I*I-I)/2
      JQ=N*(I-2)
      DO 185 J=I,N
      JQ=JQ+N
      MM=J+(J*J-J)/2
      IF(A(LL)-A(MM)) 170,185,185
  170 X=A(LL)
      A(LL)=A(MM)
      A(MM)=X
      IF(MV-1) 175,185,175
  175 DO 180 K=1,N
      ILR=IQ+K
      IMR=JQ+K
      X=R(ILR)
```

```
      R(ILR)=R(IMR)
  180 R(IMR)=X
  185 CONTINUE
      RETURN
      END
```

LIST SYMBOL/NROOT

DATE 01/17/78 TIME IS 14:36

SYSTEM/DUMPALL VERSION 2.9.170

```
LASTRECORD =      124
MAXRECSIZEIN =    14     BLOCKSIZEIN =   420
*** EBCDIC ***           UNITS=WORDS
```

```
C
C     ...............................................................
C
C        SUBROUTINE NROOT
C
C           PURPOSE
C              COMPUTE EIGENVALUES AND EIGENVECTORS OF A REAL NONSYMMETRIC
C              MATRIX OF THE FORM B-INVERSE TIMES A.  THIS SUBROUTINE IS
C              NORMALLY CALLED BY SUBROUTINE CANOR IN PERFORMING A
C              CANONICAL CORRELATION ANALYSIS.
C
C           USAGE
C              CALL NROOT (M,A,B,XL,X)
C
C           DESCRIPTION OF PARAMETERS
C              M   - ORDER OF SQUARE MATRICES A, B, AND X.
C              A   - INPUT MATRIX (M X M).
C              B   - INPUT MATRIX (M X M).
C              XL  - OUTPUT VECTOR OF LENGTH M CONTAINING EIGENVALUES OF
C                    B-INVERSE TIMES A.
C              X   - OUTPUT MATRIX (M X M) CONTAINING EIGENVECTORS COLUMN-
C                    WISE.
C
C           REMARKS
C              NONE
C
C           SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED
C              EIGEN
C
C           METHOD
C              REFER TO W. W. COOLEY AND P. R. LOHNES, 'MULTIVARIATE PRO-
C              CEDURES FOR THE BEHAVIORAL SCIENCES', JOHN WILEY AND SONS,
C              1962, CHAPTER 3.
C
C     ...............................................................
C
      SUBROUTINE NROOT (M, A, B , XL, X)
      DIMENSION A(1), B(1), XL(1), X(1)
      DOUBLE PRECISION A, B, X , XL, SUMV
C
C        COMPUTE EIGENVALUES AND EIGENVECTORS OF B
C
      K=1
      DO 100 J=2,M
      L=M*(J-1)
      DO 100 I=1,J
      L=L+1
      K=K+1
  100 B(K)=B(L)
C
```

```
C        THE MATRIX B IS A REAL SYMMETRIC MATRIX.
C
      MV = 0
      CALL  EIGEN (B, X, M , MV)
C
C     FORM RECIPROCALS OF SQUARE ROOT OF EIGENVALUES.  THE RESULTS
C     ARE PREMULTIPLIED BY THE ASSOCIATED EIGENVECTORS.
C
      L=0
      DO 110 J=1,M
      L=L+J
  110 XL(J) = 1.0 / DSQRT(DABS(B(L)))
      K=0
      DO 115 J=1,M
      DO 115 I=1,M
      K=K+1
  115 B(K)=X(K)*XL(J)
C
C     FORM (B**(-1/2))PRIME * A * (B**(-1/2))
C
      DO 120 I=1,M
      N2=0
      DO 120 J=1,M
      N1=M*(I-1)
      L=M*(J-1)+I
      X(L)=0.0
      DO 120 K=1,M
      N1=N1+1
      N2=N2+1
  120 X(L)=X(L)+B(N1)*A(N2)
      L=0
      DO 130 J=1,M
      DO 130 I=1,J
      N1=I-M
      N2=M*(J-1)
      L=L+1
      A(L)=0.0
      DO 130 K=1,M
      N1=N1+M
      N2=N2+1
  130 A(L)=A(L)+X(N1)*B(N2)
C
C     COMPUTE EIGENVALUES AND EIGENVECTORS OF A
C
      CALL  EIGEN (A, X, M , MV)
      L=0
      DO 140 I=1,M
      L=L+I
  140 XL(I)=A(L)
C
C     COMPUTE THE NORMALIZED EIGENVECTORS
C
      DO 150 I=1,M
      N2=0
      DO 150 J=1,M
      N1=I-M
      L=M*(J-1)+I
      A(L)=0.0
      DO 150 K=1,M
      N1=N1+M
      N2=N2+1
```

```
  150 A(L)=A(L)+B(N1)*X(N2)                                           0011200
      L=0                                                             0011300
      K=0                                                             0011400
      DO 180 J=1,M                                                    0011500
      SUMV=0.0                                                        0011600
      DO 170 I=1,M                                                    0011700
      L=L+1                                                           0011800
  170 SUMV=SUMV+A(L)*A(L)                                             0011900
  175 SUMV = DSQRT (SUMV)                                             0012000
      DO 180 I=1,M                                                    0012100
      K=K+1                                                           0012200
  180 X(K)=A(K)/SUMV                                                  0012300
      RETURN                                                          0012400
      END                                                             0012500
```

LIST SYMBOL/DMINV

DATE 01/17/78   TIME IS 14:36

SYSTEM/DUMPALL VERSION 2.9.170

```
LASTRECORD =      234
MAXRECSIZEIN =    14    BLOCKSIZEIN =   420
*** EBCDIC ***          UNITS=WORDS
```

```
C
C     -----------------------------------------------------------------
C
C     SUBROUTINE DMINV
C            ( IMPROVED VERSION )
C        PURPOSE
C           INVERT A MATRIX
C
C        USAGE
C           CALL DMINV (A, N, D, L, M)
C
C        DESCRIPTION OF PARAMETERS
C           A - INPUT MATRIX, DESTROYED IN COMPUTATION AND REPLACED
C               BY RESULTANT INVERSE.
C           N - ORDER OF MATRIX A
C           D - RESULTANT DETERMINANT
C           L - WORK VECTOR OF LENGTH N
C           M - WORK VECTOR OF LENGTH N
C
C        REMARKS
C           MATRIX A MUST BE A GENERAL MATRIX  (N BY N)
C
C        SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED
C           DSOLVE   (SUPPLIED WITH DMINV)
C
C        METHOD
C           THE INVERSION IS DONE USING GAUSSIAN ELIMINATION WITH
C           PARTIAL PIVOTING. THE SOLUTION IS ADJUSTED, TO GIVE
C           MAXIMUM POSSIBLE ACCURACY, USING ITERATIVE IMPROVEMENT.
C           A DETERMINANT OF ZERO INDICATES THAT THE MATRIX IS
C           SINGULAR.
C
C     -----------------------------------------------------------------
C
      SUBROUTINE DSOLVE (N, UL, B, X, L)
         DIMENSION UL(255,255),B(1),X(1),L(1)
C
      DOUBLE PRECISION UL, X, SUM, B
C
         NP1  = N+1
         IP   = L(1)
         X(1) = B(IP)
C
         DO 2 I=2,N
            IP = L(I)
            IM1 = I-1
            SUM = 0.0
C
            DO 1 J=1,IM1
    1          SUM = SUM + UL(IP,J) * X(J)
```

```
      2             X(I) = B(IP) - SUM
C
          IP = L(N)
          X(N) = X(N) / UL(IP,N)
C
          DO 4 IBACK=2,N
               I = NP1 - IBACK
C
C              I  GOES (N-1), ... ,1
C
               IP = L(I)
               IP1 = I + 1
               SUM = 0.0
C
               DO 3 J=IP1,N
      3             SUM = SUM + UL(IP,J) * X(J)
      4        X(I) = (X(I)-SUM)/UL(IP,I)
C
          RETURN
      END
C
C     -----------------------------------------------------------------
C
      SUBROUTINE DMINV (A, N, D, L, M)
          DIMENSION A(1),L(1),M(1),AA(255,255),UL(255,255),
     *              SCALES(255),X(255),R(255),DX(255),B(255),DBX(255)
          DOUBLE PRECISION SUM,DPIVOT,DEPS,DBXNOR,DBX,DBDXNO,DT
     *                    ,DBUL,DAAIJ,DXJ
     *   ,STATE,A, X, AA, UL, SCALES,ROWNRM,BIG,SIZE,PIVOT,R,DX,DIGITS,
     *    XNORM, DXNORM, T, EM, B, D, DET
C
               DET = 1.0
C
C         COPY MATRIX
               DO 1 J=1,N
                    JJ = (J-1)*N
                    DO 1 I=1,N
                         IJ = JJ+I
      1                  AA(I,J) = A(IJ)
C
C         DECOMPOSE MATRIX
C
C              INITIALIZE  L, UL, AND SCALES
C
                    DO 9 I=1,N
                         L(I) = I
                         ROWNRM = 0.0
C
                         DO 6 J=1,N
                              UL(I,J) = AA(I,J)
      4                            DBUL = UL(I,J)
                                   IF(ROWNRM-DABS(DBUL)) 5, 6, 6
      5                                 ROWNRM = DABS(DBUL)
      6                       CONTINUE
C
                         IF(ROWNRM) 7, 8, 7
      7                       SCALES(I) = 1.0 / ROWNRM
                              GO TO 9
      8                       SCALES(I) = 0.0
      9                  CONTINUE
C
```

```
C                   GAUSSIAN ELIMINATION WITH PARTIAL PIVOTING
C
                    INTCH = 0
                    NM1 = N-1
                    DO 19 K=1,NM1
                         BIG = 0.0
C
                         DO 14 I=K,N
                              IP = L(I)
      11                           DBUL = UL(IP,K)
                                   SIZE = DABS(DBUL) * SCALES(IP)
      12                      IF(SIZE-BIG) 14, 14, 13
      13                           BIG = SIZE
                                   IDXPIV = I
      14                      CONTINUE
                         IF(BIG) 15, 20, 15
      15                      IF(IDXPIV-K) 16, 17, 16
      16                           J = L(K)
                                   L(K) = L(IDXPIV)
                                   L(IDXPIV) = J
                                   INTCH = INTCH + 1
      17                      KP = L(K)
                              PIVOT = UL(KP,K)
                              KP1 = K+1
C
                              DO 18 I=KP1,N
                                   IP = L(I)
                                   EM = -UL(IP,K) / PIVOT
                                   UL(IP,K) = -EM
C
                                   DO 18 J=KP1,N
                                        UL(IP,J) = UL(IP,J)+EM*UL(KP,J)
      18                           CONTINUE
      19                 CONTINUE
C
C          EVALUATE DETERMINANT
C
                    KP = L(N)
                    IF(UL(KP,N)) 21, 20, 21
      20                 DET = 0.0
                         GO TO 52
      21   CONTINUE
      24            DO 25 K=1,N
                         KP = L(K)
                         D = UL(KP,K)
               DET = DET * D
      25                 CONTINUE
      26            IF( (INTCH/2)*2-INTCH ) 27, 28, 28
      27                 DET = -DET
C
C          SOLVE AND IMPROVE INVERSION
C
      28   DO 51 J=1,N
                JJ = (J-1)*N
C
                DO 31 I=1,N
C
C                   GENERATE IDENTITY VECTOR
C
                         IF(I-J) 30, 29, 30
      29                      B(I) = 1.0
```

```
                  GO TO 31
   30             B(I) = 0.0
   31       CONTINUE
C
      CALL DSOLVE (N,UL,B,X,L)
C      ITERATIVELY IMPROVE SOLUTION
C
C              DOUBLE PRECISION
   35             DEPS = 1.0D-23
                  ITMAX = 46
                  DBXNOR = 0.0
C
               DO 36 I=1,N
                  DBX(I) = X(I)
                  DBXNOR = DMAX1(DBXNOR,DABS(DBX(I)))
   36             X(I) = DBX(I)
               IF(DBXNOR) 38, 37, 38
   37             DIGITS = -DLOG10(DEPS)
                  GO TO 49
C
   38       DO 48 ITER=1,ITMAX
C
               DO 40 I=1,N
                  SUM = 0.0
C
                  DO 39 J2=1,N
                     DAAIJ = AA(I,J2)
                     DXJ = X(J2)
   39                SUM = SUM + DAAIJ * DXJ
C
                  SUM = B(I) - SUM
   40             R(I) = SUM
              CALL DSOLVE (N,UL,R,DX,L)
               DXNORM = 0.0
C               DOUBLE PRECISION
C
   44             DO 45 I=1,N
                     DT = X(I)
                     X(I) = X(I) + DX(I)
                     DBX(I) = X(I)
                     DBDXNO = DMAX1(DBDXNO,DABS
     *                                          (DBX(I)-DT))
                     X(I) = DBX(I)
   45                CONTINUE
C
                  STATE = DBDXNO - DEPS * DBXNOR
                  IF(ITER-1) 47, 46, 47
   46                DIGITS = -DLOG10(DMAX1(DBDXNO/DBXNOR,
     *                                                  DEPS))
   47          IF(STATE) 49, 49, 48
   48          CONTINUE
C
C      TRANSFER SOLUTION VECTOR TO SOLUTION MATRIX
C
   49       DO 50 I=1,N
               IJ = JJ+I
   50          A(IJ) = X(I)
C
   51       CONTINUE
C
   52   D = DET
```

```
      RETURN                                                            2340000
      END                                                               2350000
```

LIST SYMBOL/CONDIF

DATE 01/17/78   TIME IS 14:36

SYSTEM/DUMPALL VERSION 2.9.170

```
LASTRECORD =       34
MAXRECSIZEIN =    14     BLOCKSIZEIN =   420
*** EBCDIC ***           UNITS=WORDS
```

```
$ SET AUTOBIND
$ BIND = FROM CODON/=
      REAL MEAN
      DIMENSION MEAN(160), X(160,11), SD(11), KARHDG(3), KTYPE(3),
     1          KC(3,11)
C
C     INPUT
C
      READ 10, NG, NC, NP
   10 FORMAT (8I5)
      DO 15 I = 1,NG
   15 READ 20,(X(I,J) , J = 1,NC)
   20 FORMAT (3X, 11F7.4)
      READ 20,(SD(J), J = 1, NC)
      DO 25 J = 1, NC
   25 READ 30,(KC(K,J), K = 1,3)
      READ 30,(KTYPE(K), K = 1,3)
   30 FORMAT (3A6)
C
C     MAIN LOOP
C
      DEF = (NP - 1) * NG
      DO 100 J = 1,NC
      SEDD = SD(J) * SQRT(2.0 / NP)
      DO 35 K = 1,3
   35 KARHDG(K) = KC(K,J)
      DO 40 I = 1, NG
   40 MEAN(I) = X(I,J)
      PRINT 50, KARHDG
   50 FORMAT (1H1, 37(1H*), 5X, 17HINPUT VECTOR FOR ,3A6,5X, 37(1H*),
     1        /42X, 35(1H*)/ 42X, 35(1H*))
      CALL SMPRIN (MEAN, NG, 1, 'GROUP ', 160, 4)
      CALL DIFFS (MEAN, NG, DEF, SEDD, 0.01, J, KARHDG, 1, KTYPE, 1)
  100 CONTINUE
      END
```

LIST SYMBOL/DIFFS

DATE 01/17/78 TIME IS 14:36

SYSTEM/DUMPALL VERSION 2.9.170

LASTRECORD = 174
MAXRECSIZEIN = 14 BLOCKSIZEIN = 420
*** EBCDIC *** UNITS=WORDS

```
      SUBROUTINE DIFFS(MEAN,NMNS,DFE,SEDD,FPROB,IANAL,KARHDG,KEXPT,
     1                KTYPE,ISIGDC)
C
C     RANKS MEANS, FINDS SIGNIFICANT DIFFERENCES AND UNDERLINES GROUPS
C
C
C     SUBPROGRAMS NEEDED: TEE, DUNCAN
C
      REAL MEAN, LSD, MNRANK
      INTEGER DFE, SIGEND
      DIMENSION MEAN(300),ILIST(300),MNRANK(300),LSD(2),LEVEL(2),
     1          KARHDG(5),KTYPE(3),NOTE(4),SIGEND(300),NPOSTN(300),
     2          BUFF( 20)
C
C     RANK MEANS FROM GREATEST TO LEAST
C
      DO 15 I=1,NMNS
      BIG=0.0
      DO 10 M=1,NMNS
      IF(MEAN(M)-BIG)10,5,5
    5 JLIST=M
      BIG=MEAN(M)
   10 CONTINUE
      IF(JLIST.LT.1)JLIST=1
      MNRANK(I)=BIG
      ILIST(I)=JLIST
   15 MEAN(JLIST)=-100.0
C
C     ESTIMATE LSD(.05), LSD(.01)
C
      T05=TEE(DFE,1)
      T01=TEE(DFE,2)
      LSD(1)=SEDD*T05
      LSD(2)=SEDD*T01
      LEVEL(1)='5%'
      LEVEL(2)='1%'
C
C     SEEK SIGNIFICANCE GROUPS, AND INDICATE(P=.05,.01)
C
      IF(FPROB.IS.' NS '.OR.FPROB.IS.'(NS)')GO TO 180
      IF(FPROB.GT.0.05)GO TO 180
      IPROB=2
      IF(FPROB.IS.' *  ')IPROB=1
      IF(FPROB.LE.0.05.AND.FPROB.GT.0.01)IPROB=1
      DO 120 I=1,IPROB
C
C     PRINT HEADINGS
C
      IL=MIN0(NMNS,20)
      IF(ISIGDC-1)180,72,73
```

```
   72 NOTE(1)='SINGLE'
      NOTE(2)=' LEAST'
      NOTE(3)=' SIGNI'
      NOTE(4)='F DIFF'
      GO TO 74
   73 NOTE(1)='DUNCAN'
      NOTE(2)='S  MUL'
      NOTE(3)='TIPLE '
      NOTE(4)='RANGES'
   74 PRINT 75,IANAL,(KARHDG(J),J=1,5),KEXPT,(KTYPE(J),J=1,3),LEVEL(I),
     1      LSD(I),(NOTE(M),M=1,4)
   75 FORMAT(1H1,4X,10HCHARACTER ,I2,3X,5A6,7X,12HENVIRONMENT ,I2,10X,
     1 19HTHESE ARE MEANS OF ,3A6//26X,67HSIGNIFICANT DIFFERENCES AMONG
     2ST MEANS RANKED FROM GREATEST TO LEAST///16X,21HSIGNIFICANCE LEVEL
     3 = ,A2,10X,5HLSD =,F10.4,10X,4A6,5H USED///5X,
     4  114HRANKED MEANS IDENTIFICATION CODES ARE ON THE FIRST LINE OF E
     5ACH SUBSET BLOCK, WHILE IDENTIFICATIONS OF MEANS WHICH/9X, 52HTERM
     6INATE SIGNIFICANCE GROUPS ARE ON THE SECOND LINE)
      IF(NMNS.GT.20)PRINT 76
   76 FORMAT(1H+,60X,51H, WITH THEIR RANK POSITIONS INDICATED ON LINE TH
     1REE)
      PRINT 77,(K,K=1,IL)
   77 FORMAT(//2X,12HSUBSET/ORDER,20I5//)
C
C     FIND ENDS OF SIGNIFICANCE GROUPS
C
      DO 82 J=1,NMNS
      SIGEND(J)=ILIST(J)
      NPOSTN(J)=J
      IF(J.EQ.NMNS)GO TO 82
      DO 84 K=J+1,NMNS
      DUNC=1.0
      IF(ISIGDC.EQ.2)DUNC=DUNCAN(DFE,K-J,I)
      IF(MNRANK(J)-MNRANK(K)-LSD(I)*DUNC)80,80,82
   80 SIGEND(J)=ILIST(K)
   84 NPOSTN(J)=K
   82 CONTINUE
C
C     SET UP THE PRINTING INTO BLOCKS OF 20
C
      IM=1
      DO 55 M=1,NMNS,20
      L=MINO(M+19,NMNS)
C
C     PRINT RANKED MEANS IDENTIFICATION CODES
C
      PRINT 40, IM,(ILIST(J),J=M,L)
   40 FORMAT(/6X,I2,6X,20I5)
C
C     PRINT SIGNIFICANCE GROUPS
C
      WRITE(BUFF,90)(SIGEND(J),J=M,L)
   90 FORMAT(20I6)
      DO 91 J=M,L
      IF(J.EQ.1)GO TO 91
      IF(SIGEND(J).EQ.SIGEND(J-1))BUFF(J-(IM-1)*20)='          '
   91 CONTINUE
      PRINT 45,(BUFF(J-(IM-1)*20),J=M,L)
   45 FORMAT(//14X,20C5)
C
C     PRINT GROUP-END POSITIONS
```

```
C
      WRITE(BUFF,90)(NPOSTN(J),J=M,L)
      DO 92 J=M,L
      IF(NMNS.LT.21)BUFF(J-(IM-1)*20)='          '
      IF(J.EQ.1)GO TO 92
      IF(SIGEND(J).EQ.SIGEND(J-1))BUFF(J-(IM-1)*20)='          '
   92 CONTINUE
      PRINT 46,(BUFF(J-(IM-1)*20), J=M,L)
   46 FORMAT(///14X,20C5//////////)
   55 IM=IM+1
      IF(IPROB.EQ.1)PRINT 117
  117 FORMAT(/////11X,97HONLY SIGNIFICANT DIFFERENCES AT P=0.05 WERE OBT
     1AINED, IN VIEW OF SIGNIFICANCE OF ESTIMATED F-TEST)
      PRINT 116
  116 FORMAT(//////30X,60(1H*))
C
C     END OF LOOP FOR A GIVEN SIGNIFICANCE LEVEL
C
  120 CONTINUE
C
C     PRINT MEAN RANKS
C
  125 IF(NMNS.LT.21)GO TO 127
      IF(NMNS.LT.61.AND.IPROB.EQ.0)GO TO 128
      PRINT 129,IANAL,(KARHDG(J),J=1,5)
  129 FORMAT(1H1,4X,10HCHARACTER ,I2,3X,5A6//)
      GO TO 131
  128 PRINT 126
  126 FORMAT(////)
  131 CALL PRLA20(ILIST,KTYPE,'RANKED POSITIONS  ',NMNS,2)
      PRINT 116
  127 CONTINUE
C
C     RECONSTITUTE MEANS VECTOR
C
      DO 130 I=1,NMNS
  130 MEAN(ILIST(I))=MNRANK(I)
      RETURN
C
C     OUTPUT FOR NO SIGNIFICANT DIFFERENCES
C
C     PRINT HEADINGS
C
  180 IPROB=0
      IL=MINO(NMNS,20)
      PRINT 185,IANAL,(KARHDG(J),J=1,5),KEXPT,(KTYPE(J),J=1,3),
     1          (K,K=1,IL)
  185 FORMAT(1H1,4X,10HCHARACTER ,I2,3X,5A6,7X,12HENVIRONMENT ,I2,7X,
     1 19HTHESE ARE MEANS OF ,3A6//5X,83HMEANS RANKED FROM GREATEST TO L
     2EAST, BUT THERE ARE NO SIGNIFICANT DIFFERENCES .....//5X,34HRANKED
     3 MEANS IDENTIFICATION CODES:/5X,9HROW/ORDER,20I5///)
C
C     PRINT RANKED MEANS IDENTIFICATION CODES, WITH NO SIGNIFICANCES
C
      IM=1
      DO 195 M=1,NMNS,20
      L=MINO(M+19,NMNS)
      PRINT 190,IM,(ILIST(J),J=M,L)
  190 FORMAT(1H0,5X,I2,6X,20I5/)
  195 IM=IM+1
      PRINT 200
```

```
      200 FORMAT(////30X,60(1H*))                                        1730000
          GO TO 125                                                      1740000
          END                                                            1750000
```

```
LIST SYMBOL/SIMMAT

DATE 01/17/78           TIME IS 14:36

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =       77
MAXRECSIZEIN =     14     BLOCKSIZEIN =   420
*** EBCDIC ***            UNITS=WORDS

      FILE  5 = FILE5, UNIT = READER
      FILE  6 = FILE6, UNIT = PRINTER
      FILE  30 = FILE30, UNIT = DISK
      FILE  35 = FILE35 , UNIT = DISK
       $ SET AUTOBIND
       $ BIND = FROM CODON/=
      C
      C        PROGRAME SIMMAT  (SIMILARITY MATRIX)
      C
      C
      C           NG    = NUMBER OF GROUP
      C           NC    = NUMBER OF CHARACTER
      C           NTIN  = FILE UNIT FOR INPUT DATA
      C           NTOUT= FILE UNIT FOR SAVING THE DISTANCE VECTOR OR
      C                  SQUARED DISTANCE VECTOR
      C
      C        IPRINT= OPTION FOR PRINTING OF DISTANCE MATRIX
      C              = 1   FOR PRINT THE DISTANCE MATRIX
      C              = 2   FOR NOT PRINTING THE DISTANCE MATRIX
      C
      C         IDIST= OPTION FOR SAVING THE DISTANCE MATRIX ON FILE NTOUT
      C              = 1   FOR SAVING THE SQUARED DISTANCE ONLY
      C              = 2   FOR SAVING THE DISTANCE ONLY
      C              = 3   FOR NOT SAVING ANY
      C
      C         FMTIN= FORMAT FOR INPUT DATA
      C        FMTOUT= FORMAT FOR OUTPUT DATA, I.E. FOR SAVING ON FILE NTOUT
      C
      C
            REAL MEAN
            DIMENSION MEAN(160,15), D(160,160), SD(15), FMTIN(16), FMTOUT(16)
           1        , R(12720)
            READ (5,5) NG, NC, NTIN, NTOUT, IPRINT, IDIST
          5 FORMAT (10I5)
            READ (5,10) (FMTIN (I), I = 1,16)
            READ (5,10) (FMTOUT(I), I = 1,16)
         10 FORMAT  (16A5)
            DO 20 I = 1 ,NG
         20 READ (NTIN , FMTIN) (MEAN(I,P) , P = 1,NC)
            READ (5     , 25    ) (SD(P), P = 1, NC)
         25 FORMAT(4E20.13)
            CALL DISTA (MEAN, NG, NC, D, SD, 160, 15)
      C
      C     OPTION FOR PRINTING THE D AND D-SQUARE MATRIX
      C
            GO TO (30 , 40) , IPRINT
         30 CONTINUE
            WRITE (6,35) NG
         35 FORMAT(1H1, 44(1H*), 5X, 'D AND D-SQUARE MATRIX', 5X, 44(1H*)/50X
           1        21(1H*)/50X, 21 (1H*)//, 29X, 'DISTANCES (IN LOWER HALF) '
```

```
         2          'AND DISTANCES SQUARE (IN UPPER HALF)',/37X,'BETWEEN' I4,
         3           ' STANDARDIZED MEAN VECTORS'/)
          CALL SMPRIN (D, NG, NG, 'D&D-SQ', 160,10)
       40 CONTINUE
C
C
C         OPTION FOR SAVING DISTANCE OR DISTANCES SQUARE ON TAPE
C
          GO TO (50, 70, 100), IDIST
       50 CONTINUE
          DO 60 I = 1, NG-1
          DO 60 J = I+1, NG
          K = ((J - 1) * (J - 2)) / 2 + I
          R(K) = D(I,J)
       60 CONTINUE
          GO TO 90
       70 CONTINUE
          DO 80 I = 1, NG-1
          DO 80 J = I+1, NG
          K = ((J - 1) * (J - 2)) / 2 + I
          R(K) = D(J,I)
       80 CONTINUE
       90 CONTINUE
          NOS = NG * (NG - 1) / 2
          WRITE(NTOUT, FMTOUT) (R(K), K = 1,NOS)
          LOCK NTOUT
      100 CALL EXIT
          END
```

```
LIST SYMBOL/DISTA
DATE 01/17/78          TIME IS 14:36
SYSTEM/DUMPALL VERSION  2.9.170
LASTRECORD =       64
MAXRECSIZEIN =    14    BLOCKSIZEIN =   420
*** EBCDIC ***          UNITS=WORDS
```

```
C
C     ==================================================================
C
C     SUBROUTINE DISTA (X, NG, NC, D, SD, NGD, NCD)
C
C     DESCRIPTION OF PARAMETERS
C
C         X = INPUT MATRIX OF NG * NC.
C
C        NG = NO. OF GROUP (ROW) OF INPUT MATRIX.
C
C        NC = NO. OF CHARACTER (COLUMN) OF INPUT MATRIX.
C
C         D = OUTPUT MATRIX OF NG * NG , CONTAINING  D-SQUARE IN UPPER
C             HALF, AND D IN LOWER HALF. IT SHOULD BE DIMENSIONED AS A
C             SQUARE MATRIX OF NGD * NGD IN THE CALLING PROGRAM.
C
C        SD = VECTOR OF NC ELEMENTS, FOR WEIGHING X ,E.G. THE STANDARD
C             DEVIATION OF EACH CHARACTER.
C             SET ALL ELEMENTS OF SD = 1 IF NO WEIGHING IS NEEDED.
C
C       NGD = NO. OF ROW OF INPUT MATRIX DIMENSIONED FOR IN THE CALLING
C             PROGRAM.
C
C       NCD = NO. OF COLUMN OF INPUT MATRIX  AND NO. OF ELEMENT OF
C             VECTOR SD DIMENSIONED FOR IN THE CALLING PROGRAM.
C
C     SUBROUTINE NEEDED
C        NONE
C
C     PROGRAMMED BY
C
C        S. H. TEOW (OCTOBER, 1976)
C
C     ==================================================================
C
      SUBROUTINE DISTA(X, NG, NC, D, SD, NGD, NCD)
      DIMENSION X(NGD, NCD), D(NGD,NGD), SD(NCD)
C
C     STANDARDIZATION
C
      DO 10 I = 1, NG
      DO 10 K = 1, NC
      IF(SD(K).LE.0.0) SD(K) = 1.0
      X(I,K) = X(I,K) / SD(K)
   10 CONTINUE
C
C     FIND D SQUARE
C
      DO 20 I = 1,NG
```

```
      DO 20 J = 1, NG                                                   0005100
      D(I,J) = 0.0                                                      0005200
      DO 15 K = 1,NC                                                    0005300
      D(I,J) = D(I,J) + (X(I,K) - X(J,K)) ** 2                          0005400
   15 CONTINUE                                                          0005500
   20 CONTINUE                                                          0005600
C                                                                       0005700
C     FIND D                                                            0005800
C                                                                       0005900
      DO 25 I = 1, NG                                                   0006000
      DO 25 L = I, NG                                                   0006100
      D(L,I) = SQRT (D(L,I))                                            0006200
   25 CONTINUE                                                          0006300
      RETURN                                                            0006400
      END                                                               0006500
```

```
LIST SYMBOL/CLUSAN

DATE 01/17/78          TIME IS 14:36

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =        20
MAXRECSIZEIN =    14    BLOCKSIZEIN =   420
*** EBCDIC ***          UNITS=WORDS

      FILE  5 = FILE5, UNIT = READER                                  0000100
      FILE  6 = FILE6, UNIT = PRINTER                                 0000200
      FILE  20 = NTIN, UNIT = DISK                                    0000300
      FILE  30 = NTSV, UNIT = DISK                                    0000400
       $ SET AUTOBIND                                                 0000500
       $ BIND = FROM CODON/=                                          0000600
      C                                                               0000700
      C     PROGRAM  CLUSAN                                           0000800
      C                                                               0000900
      C     CLUSTER ANALYSIS.                                         0001000
      C                                                               0001100
            DIMENSION X(40000)                                        0001200
            LIMIT = 40000                                             0001300
      C                                                               0001400
            CALL CONTRO(X, LIMIT)                                     0001500
      C                                                               0001600
            WRITE (6,20)                                              0001700
         20 FORMAT (1H1// 42X, 37(1H*)/ 42X, 37(1H*)/// 42X,          0001800
           1         37HPROGRAM CLUSAN BY S.H. TEOW JAN. 1977 ///     0001900
           2         42X, 37(1H*)/ 42X, 37(1H*))                      0002000
            END                                                       0002100
```

LIST SYMBOL/CONTRO

DATE 01/17/78   TIME IS 14:36

SYSTEM/DUMPALL VERSION 2.9.170

LASTRECORD =   172
MAXRECSIZEIN =   14   BLOCKSIZEIN =   420
*** EBCDIC ***   UNITS=WORDS

```
      SUBROUTINE CONTRO (X, LIMIT)                                      0000100
C                                                                       0000200
C     THIS SUBROUTINE ALLOCATES STORAGE, READS INPUT AND CONTROLS       0000300
C     EXECUTION FOR A HIERARCHICAL CLUSTERING JOB BASED ON PROVIDED     0000400
C     SIMILARITY MATRIX.                                                0000500
C                                                                       0000600
C-----------------------------------------------------------------------0000700
C                                                                       0000800
C     INPUT SPECIFICATIONS                                              0000900
C                                                                       0001000
C     CARD 1  TITLE CARD FOR THIS RUN                                   0001100
C        TO ENABLE MULTIPLE RUN (E.G. FOR DIFFERENT METHOD OF CLUSTERING0001200
C        ON SAME SET OF DATA), THE FIRST 5 COLUMNS (COLS 1 TO 5) OF THE 0001300
C        TITLE MUST NOT CONTAIN THE SAME WORD AS THE NEXT 5 COLUMNS     0001400
C        (COLS 6 TO 10)  OR ELSE IT WILL BE READ AS END CARD            0001500
C                                                                       0001600
C     CARD 2  INFORMATION FOR SUBROUTINES CLSTR AND TREE                0001700
C        COLS 1-5   NE = NUMBER OF ENTITIES (DATA UNITS OR VARIABLES) TO 0001800
C                        BE CLUSTERED                                   0001900
C        COLS 6-10 ISIGN = OPTION FOR SIMILARITY FUNCTION               0002000
C                        = +1, FOR DISTANCE MEASURE                     0002100
C                        = -1, FOR CORRELATION MEASURE                  0002200
C        COLS11-15 NTSV = TAPE UNIT ON WHICH CLSTR RESULT ARE SAVE      0002300
C                       = 7, PUNCH RESULT ON CARDS                      0002400
C                    .LE.0, DO NOT SAVE RESULT                          0002500
C        COLS16-20 NTIN = TAPE UNIT FROM WHICH SIMILARITY MATRIX IS READ0002600
C                       = 5, CARD READER                                0002700
C        COLS21-25 KOUT = OUTPUT OPTION                                 0002800
C                       = +1 PRINT ONLY TITLE AND *IS* ARRAY FROM       0002900
C                            SUBROUTINE *TREE*                          0003000
C                       = +2, FOR STANDARD OUTPUT                       0003100
C        COLS26-30 METHOD = THE METHOD OF CLUSTERING USED IN THE        0003200
C                           SUBROUTINE *CLSTR*                          0003300
C                         = 1 FOR CALLING METH1                         0003400
C                         = 2 FOR CALLING METH2                         0003500
C                         = 3 FOR CALLING METH3                         0003600
C                         = 4 FOR CALLING METH4                         0003700
C                         = 5 FOR CALLING METH5                         0003800
C                         = 6 FOR CALLING METH6                         0003900
C                         = 7 FOR CALLING METH7                         0004000
C                                                                       0004100
C                                                                       0004200
C ****ANY PREPOSITIONING OF THE I/O UNITS NTSV AND NTIN MUST BE         0004300
C     ACCOMPLISHED IN THE MAIN PROGRAM OR THROUGH USE OF CONTROL CARDS. 0004400
C                                                                       0004500
C                                                                       0004600
C     CARD 3  INPUT FORMAT FOR SIMILARITY MATRIX(16A5 FORMAT)           0004700
C                                                                       0004800
C     CARD(S) 4 SIMILARITY MATRIX                                       0004900
C ****INCLUDE CARDS 4 ONLY IF THE SIMILARITY MATRIX IS ON CARDS         0005000
```

```
C
C     CARD(S) 5 LABEL CARDS FOR ENTITIES.    THERE ARE TWO OPTIONS.
C         1.   INCLUDE 1 CARD WITH THE 5 CHARACTERS *NOLAB* IN COLUMNS 1-5
C              UNDER THIS OPTION LABELS ARE NOT PRINTED ON THE TREE OUTPUT
C
C         2.   INCLUDE NE CADRS, COLUMNS 1 TO 10 CONTAINING A LABEL FOR
C              ONE ENTITY. ORDER THE LABEL CARDS IN THE SAME SEQUENCE AS
C              THE ENTITIES ARE REPRESENTED IN THE SIMILARITY MATRIX.
C
C ****IF MULTIPLE RUN OF THIS PROGRAM IS REQUIRED THEN REPEAT CARD 1 TO
C     CARD(S) 5 AGAIN
C
C     CARD 6  END CARD . COLUMNS 1 TO 5 MUST CONTAIN THE SAME WORD AS
C               COLUMNS 6 TO 10  OR LEAVE COLUMNS 1 TO 10 BLANK.
C               THIS WILL SIGNAL THE PROGRAM TO END .
C
C-----------------------------------------------------------------------
C
C     DECK SETUP SPECIFICATIONS FOR THE MAIN PROGRAME
C
C
C     THE USER PROVIDES MAIN PROGRAM WHICH PERFORMS THE FOLLOWING TASKS.
C
C         1.    ASSIGNS INPUT/OUTPUT UNITS
C         2.    ESTABLISHES THE DIMENSION OF THE X ARRAY AND SETS THIS
C               DIMENSION EQUAL TO LIMIT.
C         3.    CALLS SUBROUTINE CONTRO .
C
C
C-----------------------------------------------------------------------
C
C-----------------------------------------------------------------------
C
C
C
      INTEGER FIRST
      DIMENSION X(1), FMT(16), TITLE(16), EPS(25)
   10 READ (5, 1000) TITLE
      WRITE (6, 2500) TITLE
      IF (TITLE(1).IS.TITLE(2)) GO TO 100
      READ (5,1100) NE, ISIGN, NTSV, NTIN, KOUT, METHOD
      WRITE(6, 2200) NE, ISIGN, NTSV, NTIN, KOUT, METHOD
C
C
C     PARTITION THE STORAGE ARRAY
C
      N1 = 1
      N2 = N1 + NE
      N3 = N2 + NE
      N4 = N3 + NE
      N5 = N4 + NE
      N6 = N5 + NE
      N7 = N6 + NE
      M2 = N7 + (NE * (NE - 1)) / 2
      M3 = M2 + NE
      M4 = M3 + NE
      M5 = M4 + NE
      M6 = M5 + NE
      M7 = M6 + NE
      M8 = M7 + NE - 1
      L2 = N7 + 25 * NE
```

```
      L3 = L2 + 2 * NE
      L4 = L3 + NE
      L5 = L4 + NE
      L6 = L5 + NE
      L7 = L6 + NE  - 1
C
C     CHECK FOR SUFFICIENT STORAGE
C
      MAX = M8
      IF(L7.GT.MAX) MAX = L7
      WRITE (6,2300) MAX, LIMIT
      IF(MAX.GT.LIMIT) STOP
C
C     READ THE SIMILARITY MATRIX
C
      READ (5, 1000) FMT
      WRITE (6,2100) FMT
      FIRST = N7
      LAST = M2 - 1
      READ (NTIN, FMT) (X(I), I = FIRST, LAST )
C
C     READY TO CLUSTER
C
      CALL CLSTR(X(N1), X(N2), X(N3), X(N4), X(N5), X(N6), X(N7),
     1           X(M2), X(M3), X(M4), X(M5), X(M6), X(M7), TITLE, NE,
     2           ISIGN, NTSV, METHOD)
C
C     READ THE LABLE
      FIRST = L2
      LAST = L2 + 1
      READ (5, 1000) (X(I), I = FIRST, LAST)
      IF (X(FIRST).IS.5HNOLAB) GO TO 80
C
C     READ REMAINING LABELS
C
      DO 70 K = 2, NE
      FIRST = LAST + 1
      LAST = LAST + 2
   70 READ (5,1000) (X(I), I = FIRST, LAST)
C
C     DRAW THE TREE CORRESPONDING TO THE CLUSTERING
C
   80 MERGES = NE - 1
      CALL TREE (X(N1), X(N2), X(N3), X(N4), X(N5), X(N6), X(N7),
     1           X(L2), X(L3), X(L4), X(L5), X(L6), EPS, TITLE, MERGES,
     2           1, 6, 1, KOUT, NE)
      INQUIRE (NTIN, KIND = KND)
      IF (NTIN.NE.5.AND.KND.NE.9) REWIND (NTIN)
   90 GO TO 10
  100 RETURN
 1000 FORMAT (16A5)
 1100 FORMAT (8I5)
 2100 FORMAT (10X, 16H INPUT FORMAT : , 16A5//)
 2200 FORMAT (10X, 5H NE = ,I8,//, 10X, 8H ISIGN = , I5,//, 10X,
     1           7H NTSV = , I6,// 10X, 7H NTIN = , I6, // 10X,
     2           7H KOUT = , I6, //10X, 10H METHOD = , I3/)
 2300 FORMAT(10X, 19H REQUIRED STORAGE = , I7, 6H WORDS,//
     1          10X, 19H ALLOTTED STORAGE = , I7, 6H WORDS,/)
 2500 FORMAT (1H1, /20X, 32(1H$), 18H CLUSTER ANALYSIS , 32(1H$), /
     1           20X, 1H$, 80X, 1H$/ 20X, 1H$, 80X, 1H$/20X, 1H$, 16A5,1H$/
     2           20X, 1H$, 80X, 1H$/ 20X, 1H$, 80X, 1H$/20X, 82(1H$)///)
```

```
173      END                                                   0017300
```

```
LIST SYMBOL/CLSTR

DATE 01/17/78          TIME IS 14:36

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =     252
MAXRECSIZEIN =   14     BLOCKSIZEIN =  420
*** EBCDIC ***          UNITS=WORDS

      SUBROUTINE CLSTR (II, JJ, SS, IL, JL, NEXT, S, LAST, NEAR, SREF,  0000100
     1                  LIST, A, B, TITLE, N, ISIGN, NT, METHOD)        0000200
C                                                                       0000300
C     INPUT PARAMETERS                                                  0000400
C                                                                       0000500
C     N = NUMBER OF OBJECTS TO BE CLUSTERED                             0000600
C                                                                       0000700
C     S = THE ONE DIMENSIONAL ARRAY WHICH STORED THE LOWER TRIANGULAR   0000800
C         PORTION OF THE SIMILARITY MATRIX BY ROWS.                     0000900
C                                                                       0001000
C     ISIGN = OPTION SPECIFYING TYPE OF SIMILARITY FUNCTION USED        0001100
C     ISIGN = +1 FOR DISTANCE MEASURE (DECREASING FUNCTION OF SIMILARITY0001200
C     ISIGN = -1 FOR CORRELATION MEASURE (INCREASING FUNCTION OF        0001300
C                SIMILARITY)                                            0001400
C                                                                       0001500
C     NT = TAPE UNIT ON WHICH THE RESULT ARE SAVED                      0001600
C     NT.LE.0 FOR NOT SAVING RESULTS ON TAPE                            0001700
C                                                                       0001800
C     TITLE = IDENTIFYING TITLE FOR THIS RUN                            0001900
C                                                                       0002000
C     METHOD  = THE METHOD USED FOR CLUSTERING                          0002100
C             = 1 FOR CALLING METH1                                     0002200
C             = 2 FOR CALLING METH2                                     0002300
C             = 3 FOR CALLING METH3                                     0002400
C             = 4 FOR CALLING METH4                                     0002500
C             = 5 FOR CALLING METH5                                     0002600
C             = 6 FOR CALLING METH6                                     0002700
C             = 7 FOR CALLING METH7                                     0002800
C                                                                       0002900
C                                                                       0003000
C     OUTPUT                                                            0003100
C                                                                       0003200
C                                                                       0003300
C        THE RESULTS ARE READY FOR SUBROUTINE TREE.                     0003400
C                                                                       0003500
C     II = ARRAY OF K ELEMENTS CONTAINING LOWER GROUP IDENTIFICATION    0003600
C          NUMBER , MERGED AT STAGE K                                   0003700
C                                                                       0003800
C     JJ = ARRAY OF K ELEMENTS CONTAININGUPPERR GROUP IDENTIFICATION    0003900
C          NUMBER , MERGED AT STAGE K                                   0004000
C                                                                       0004100
C     SS = ARRAY OF K ELEMENTS CONTAINING VALUE OF SIMILARITY FUNCTION  0004200
C          ASSOCIATED WITH MERGE AT STAGE K                             0004300
C                                                                       0004400
C     IL = ARRAY OF K ELEMENTS CONTAINING STAGE NUMBER AT WHICH II(K)   0004500
C          WAS LAST IN A MERGE (0 FOR FIRST MERGE)                      0004600
C                                                                       0004700
C     JL = ARRAY OF K ELEMENTS CONTAINING STAGE NUMBER AT WHICH JJ(K)   0004800
C          WAS LAST IN A MERGE (0 FOR FIRST MERGE)                      0004900
C                                                                       0005000
```

```
C     NEXT = ARRAY OF K ELEMENTS CONTAINING NUMBER OF NEXT STAGE AT
C            WHICH II(K) IS IN A MERGE
C
C     LAST = ARRAY OF K ELEMENTS CONTAINING NUMBER OF PREVIOUS STAGE AT
C            WHICH II(K) WAS IN A MERGE
C
C     NEAR = ARRAY OF N ELEMENTS CONTAINING IDENTIFICATION NUMBER OF
C            EXTREME ELEMENT IN ROW I OF THE LOWER TRIANGULAR SIMILARITY
C            MATRIX.
C
C     SREF = ARRAY OF N ELEMENTS CONTAINING THE VALUE OF SIMILARITY
C            FUNCTION FOR THE PAIR (I, NEAR(I))
C
C     LIST = ARRAY OF N ELEMENTS CONTAINING I-TH GROUP IDENTIFICATION
C            NUMBER IN SEQUENTIAL LIST OF CURRENT CLUSTERS
C
C     A = ARRAY OF N ELEMENTS AS WORKING SPACE FOR SUBROUTINE METHOD
C
C     B = ARRAY OF N ELEMENTS AS WORKING SPACE FOR SUBROUTINE METHOD
C
C
      DIMENSION S(1), II(1), JJ(1), SS(1), IL(1), JL(1), NEXT(1),
     1          NEAR(1), SREF(1), LIST(1), LAST(1), A(1), B(1)
      DIMENSION TITLE (16)
C
C     PRINT HEADING
C
      WRITE (6, 1000)
 1000 FORMAT (/     11X, 'THE METHOD USED IN THIS ANALYSIS IS : ' )
C
C
C     INITIALIZE VARIABLES AND SET CONSTANTS
C
      NCL = N
      K = 1
      SIGN = ISIGN
      BIG = SIGN * 1.0E50
      GO TO (201, 202, 203, 204, 205, 206, 207), METHOD
  201 CALL METH1 (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1            NREF, 1)
      GO TO 300
  202 CALL METH2 (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1            NREF, 1)
      GO TO 300
  203 CALL METH3 (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1            NREF, 1)
      GO TO 300
  204 CALL METH4 (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1            NREF, 1)
      GO TO 300
  205 CALL METH5 (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1            NREF, 1)
      GO TO 300
  206 CALL METH6 (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1            NREF, 1)
      GO TO 300
  207 CALL METH7 (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1            NREF, 1)
  300 CONTINUE
C
C     INITIALIZE ARRAYS
```

```
C
      DO 10 J = 1,N
      LAST(J) = 0
      NEXT(J) = 0
      LIST(J) = J
      SREF(J) = BIG
   10 CONTINUE
C
C     FIND EXTREM ENTRY IN EACH ROW
C
      L = 0
      DO 30 I = 2,N
      I1 = I - 1
      DO 30 J = 1,I1
      L = L + 1
C
C     IN EFFECT S(L) = S(I,J)
C
      IF (((S(L) - SREF(I)) * SIGN).GT.0) GO TO 30
      NEAR(I) = J
      SREF(I) = S(L)
   30 CONTINUE
C
C      MAIN LOOP. FIND EXTREM VALUE IN SREF ARRAY
C
   40 SREFX = BIG
      DO 50 I = 2,NCL
      LISTI = LIST(I)
      IF (((SREF(LISTI) - SREFX) * SIGN).GT.0) GO TO 50
      IREF = I
      LREF = LISTI
      SREFX = SREF(LISTI)
   50 CONTINUE
C
C     LREF IS THE ROW NUMBER CONTAINING THE EXTREME ENTRY IN THE S ARRAY
C     . IF THERE ARE TIE,  THEN LREF IS THE HIGHEST NUMBERED ROW WITH
C     THIS EXTREME VALUE. HENCE LREF .GT. NEAR(LREF). IREF IDENTIFIES
C     THE PLACEMENT OF LREF IN THE LIST ARRAY.
C
      NREF = NEAR(LREF)
      GO TO (401, 401, 401, 401, 401, 401, 400), METHOD
  400 CALL METH7  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1              NREF, 2)
  401 CONTINUE
C
C     GENERATE MERGE DATA NEEDED FOR SUBROUTINE TREE
C
      II(K) = NREF
      JJ(K) = LREF
      SS(K) = SREFX
      IL(K) = LAST(NREF)
      JL(K) = LAST(LREF)
      LAST (NREF) = K
      IF (IL(K).EQ.0) GO TO 60
      ILK = IL(K)
      NEXT(ILK) = K
   60 IF (JL(K).EQ.0) GO TO 70
      JLK = JL(K)
      NEXT (JLK) = K
   70 K = K + 1
C     TERMINATE IF N-1 MERGES HAVE OCCURED
```

```
      IF (K.EQ.N) GO TO 140
C
C     UPDATE FOR THE NEXT CYCLE
C
      NCL = NCL- 1
      IF (IREF.GT.NCL) GO TO 90
C
C     UPDATE LIST ARRAY BY REMOVING LREF AND PUSHING DOWN THE LIST
C
      DO 80 I = IREF, NCL
   80 LIST(I) = LIST(I+1)
C
C     UPDATE FOR NEXT CYCLE
C
   90 CONTINUE
      GO TO (601, 602, 603, 604, 605, 606, 607), METHOD
  601 CALL METH1  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1             NREF, 3)
      GO TO 700
  602 CALL METH2  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1             NREF, 3)
      GO TO 700
  603 CALL METH3  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1             NREF, 3)
      GO TO 700
  604 CALL METH4  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1             NREF, 3)
      GO TO 700
  605 CALL METH5  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1             NREF, 3)
      GO TO 700
  606 CALL METH6  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1             NREF, 3)
      GO TO 700
  607 CALL METH7  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL, LREF,
     1             NREF, 3)
  700 CONTINUE
      GO TO 40
C
C     CLUSTERING FINISHED AND ALL ANCILLARY INFORMATION GENERATED.
C     SAVE RESULTS AS DESIRED.
C
  140 K = K - 1
  160 IF (NT.LE.0) RETURN
      WRITE (NT, 2300) TITLE
      WRITE (NT, 2100)
      DO 170 I = 1, K
  170 WRITE(NT, 2200) I, II(I), JJ(I), SS(I), IL(I), JL(I), NEXT(I)
      LOCK NT
      RETURN
 2100 FORMAT (/6X, 'STAGE', 5X, 'LOWER', 5X, 'HIGHER', 5X, 'VALUE OF ',
     1            'CRITERION', 4X, 'STAGE WHERE', 4X, 'STAGE WHERE', 4X,
     2            'STAGE WHERE',
     3           /6X, 'OF', 8X, 'CLUSTER', 3X, 'CLUSTER', 4X, 'ASSOCIATED',
     4            ' WITH', 7X, 'I WAS LAST', 5X, 'J WAS LAST', 5X, 'I IS ',
     5            'NEXT',
     6           /6X, 'MERGE', 5X, 'ID. NO.', 3X, 'ID. NO.', 4X,'THE MERGE'
     7            , 13X,'IN A MERGE', 5X, 'IN A MERGE', 5X, 'IN A MERGE'/
     8           /8X, 'K', 9X, 'I', 9X, 'J', 14X, 'S', 20X, 'IL', 13X, 'JL'
     9            , 12X, 'NEXT')
 2200 FORMAT ( /3I10, 5X, E16.8, 3I15)
```

```
 2300 FORMAT (/ 10X, 10(1H*), 16A5, 10(1H*))
      END
      FUNCTION LFIND(I,J)
C
C     IF THE LOWER TRIANGULAR PORTION OF A SYMMETRIC MATRIX IS STORED BY
C     ROWS IN A ONE-DIMENSIONAL ARRAY, THEN THE ELEMENT (I,J) IN THE
C     FULL MATRIX IS ELEMENT LFIND(I,J) IN THE LINEAR ARRAY.
C
      IF (I.GT.J) GO TO 10
C
C     ROW J, COLUMN I
C
      LFIND = ((J - 1) * (J - 2))/ 2 + I
      RETURN
C
C     ROW I, COLUMN J
C
   10 LFIND = ((I - 1) * (I - 2))/2 + J
      RETURN
      END
```

```
LIST SYMBOL/METH1

DATE 01/17/78            TIME IS 14:37

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =        75
MAXRECSIZEIN =      14    BLOCKSIZEIN =   420
*** EBCDIC ***            UNITS=WORDS

      SUBROUTINE METH1 (S, NEAR, SREF, LIST, NUMBER, SUM, SREFX, SIGN,
     1                  N, NCL, LREF, NREF, JOB)
C
C     HIERARCHICAL CLUSTERING BY CENTROID SORTING
C
C     THE ALGORITHM USED IS DESCRIBED IN :
C     LANCE, G.N. AND W.T. WILLIAMS, (1967) A GENERAL THEORY OF
C        CLASSIFICATIORY SORTING STRATEGIES, 1. HIERARCHICAL SYSTEM, THE
C            COMPUTER JOURNAL, VOLUMN 9, NUMBER 4, FEBRUARY 1967,
C        PP373-380.
C
      DIMENSION S(1), NEAR(1), SREF(1), LIST(1), NUMBER(1), SUM(1)
      GO TO (10, 25, 30) , JOB
C
C     JOB = 1, INITIALIZE.
C     NUMBER(I) = NUMBER OF ENTITIES CURRENTLY IN THE I-TH CLUSTER
C     SUM(I) = DUMMY
C
   10 WRITE (6, 2000)
 2000 FORMAT (1H+, 49X, 42H CENTROID CLUSTERING.  BEWARE OF REVERSALS ,
     1        / 51X, 41(1H=))
      DO 20 J = 1,N
   20 NUMBER(J) = 1
      RETURN
C
C     JOB = 2, DUMMY ENTRY.
C
   25 RETURN
C
C     JOB = 3, UPDATE FOR NEXT ROUND.
C     UP DATE THE NEW CLUSTER
C
   30 NTOT = NUMBER(NREF) + NUMBER(LREF)
      TOT = NTOT
      ALL = NUMBER(LREF)/TOT
      ALN = NUMBER(NREF) / TOT
      PROD = ALN * ALL
      LBET = LFIND(LREF,NREF)
      DO 40 J = 1,NCL
      I = LIST(J)
      IF (I.EQ.NREF) GO TO 40
C
C     RECALL THAT LREF HAS BEEN REMOVED FROM LIST AND THEREFORE I NEED
C     NOT BE TESTED FOR EQUALITY WITH LREF.
C
      LL = LFIND(I,LREF)
      LN = LFIND(I,NREF)
      S(LN) = ALL * S(LL) + ALN * S(LN) - PROD * S(LBET)
   40 CONTINUE
      NUMBER(NREF) = NTOT
```

```
      C                                                                 0005100
      C     UPDATE THE NEAR AND SREF ARRAY.  IF THE EXTREME ELEMENT IN ROW I  0005200
      C     WAS EITHER LREF OR NREF , THEN IT IS NECESSARY TO FIND A NEW      0005300
      C     EXTREME ELEMENT.  ROW PRIOR TO NREF NEED NOT BE CONSIDERED.       0005400
      C                                                                 0005500
            BIG = SIGN * 1.0E50                                         0005600
            DO 50 J = 1,NCL                                             0005700
            I = LIST(J)                                                 0005800
            IF (I.EQ.NREF) GO TO 55                                     0005900
         50 CONTINUE                                                    0006000
         55 IF (J.EQ.1) GO TO 80                                        0006100
         60 SREF(I) = BIG                                               0006200
            J1 = J - 1                                                  0006300
            DO 70 L = 1, J1                                             0006400
            LISTL = LIST(L)                                             0006500
            LL = LFIND(I,LISTL)                                         0006600
            IF (((S(LL)-SREF(I))*SIGN).GE.0.0) GO TO 70                 0006700
            NEAR(I) = LISTL                                             0006800
            SREF(I) = S(LL)                                             0006900
         70 CONTINUE                                                    0007000
         80 J = J + 1                                                   0007100
            IF (J.GT.NCL) RETURN                                        0007200
            I = LIST(J)                                                 0007300
            IF (NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60            0007400
            GO TO 80                                                    0007500
            END                                                         0007600
```

LIST SYMBOL/METH2

DATE 01/17/78   TIME IS 14:37

SYSTEM/DUMPALL VERSION 2.9.170

```
LASTRECORD =        63
MAXRECSIZEIN =      14    BLOCKSIZEIN =   420
*** EBCDIC ***            UNITS=WORDS
```

```
      SUBROUTINE METH2 (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL,
     1                  LREF, NREF, JOB)
C
C     HIERARCHICAL CLUSTERING BY COMPLETE LINKAGE.
C     THE ALGORITHM IS DERIVED FROM:
C     JOHNSON, S.C., (1967) HIERARCHICAL CLUSTERING SCHEMES,
C     PSYCHOMATRIKA, VOLUME 32, NUMBER 3, SEPTEMBER 1967, PP 241-254.
C
C
      DIMENSION S(1), NEAR(1), SREF(1), LIST(1), A(1), B(1)
      GO TO (10, 15, 20), JOB
C
C     JOB = 1, INITIALIZATION.
C
   10 WRITE(6, 2000)
 2000 FORMAT (1H+, 49X, 28H COMPLETE LINKAGE CLUSTERING ,/51X, 28(1H=))
      RETURN
C
C     JOB = 2, DUMMY ENTRY.
C
C
   15 RETURN
C
C     JOB = 3, UPDATE FOR NEXT ROUND.
C     UPDATE THE NEW CLUSTER
C
   20 DO 30 J = 1, NCL
      I = LIST(J)
      IF (I.EQ.NREF) GO TO 30
C
C     RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C     TESTED FOR EQUALITY WITH LREF.
C
      LL = LFIND(I,LREF)
      LN = LFIND(I,NREF)
      IF (((S(LL) - S(LN))* SIGN).LE.0.0) GO TO 30
      S(LN) = S(LL)
   30 CONTINUE
C
C     UPDATE THE NEAR AND SREF ARRAYS. IF THE EXTREME ELEMENT IN ROW I
C     WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW
C     EXTREME ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
C
      BIG = SIGN * 1.0E50
   40 DO 50 J = 1, NCL
      I = LIST(J)
      IF (I.EQ.NREF) GO TO 55
   50 CONTINUE
   55 IF (J.EQ.1) GO TO 80
   60 SREF(I) = BIG
```

```
      J1 = J - 1                                                      0005100
      DO 70 L = 1, J1                                                 0005200
      LISTL = LIST(L)                                                 0005300
      LL = LFIND(I, LISTL)                                            0005400
      IF (((S(LL) - SREF(I))* SIGN).GE.0.0) GO TO 70                  0005500
      NEAR (I) = LISTL                                                0005600
      SREF(I) = S(LL)                                                 0005700
   70 CONTINUE                                                        0005800
   80 J = J + 1                                                       0005900
      IF (J.GT.NCL) RETURN                                            0006000
      I = LIST(J)                                                     0006100
      IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60                 0006200
      GO TO 80                                                        0006300
      END                                                             0006400
```

```
LIST SYMBOL/METH3

DATE 01/17/78          TIME IS 14:37

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =          77
MAXRECSIZEIN =        14     BLOCKSIZEIN =   420
*** EBCDIC ***               UNITS=WORDS

      SUBROUTINE METH3 (S, NEAR, SREF, LIST, NUMBER, SUM, SREFX, SIGN,
     1                  N, NCL, LREF, NREF, JOB)
C
C     HIERARCHICAL CLUSTERING BY MINIMIZING YHE AVERAGE DISTANCE OR
C     MAXIMIZING THE AVERAGE CORRELATION WITHIN THE NEW GROUP. THAT IS,
C     FOR EACH POTENTIAL MERGE THE AVERAGE OF ALL LINKAGES WITHIN THE
C     NEW GROUP IS CALCULATED.
C
      DIMENSION S(1), NEAR(1), SREF(1), LIST(1), NUMBER(1), SUM(1)
      GO TO (10, 25, 30), JOB
C
C     JOB = 1, INITIALIZE.
C     NUMBER(I) = NUMBER OF ENTITIES CURRENTLY IN THE I-TH CLUSTER
C     SUM(I) = SUM OF ALL PAIRWISE SIMILARITIES AMONG ENTITIES IN THE
C              I-TH CLUSTER
C
   10 WRITE (6, 2000)
 2000 FORMAT (1H+, 49X, 37H AVERAGE LINKAGE WITHIN THE NEW GROUP ,
     1        / 51X, 37(1H=))
      DO 20 J = 1, N
      NUMBER(J) = 1
   20 SUM(J) = 0.0
      RETURN
C
C     JOB = 2, DUMMY ENTRY.
C
   25 RETURN
C
C     JOB = 3, UPDATE FOR NEXT ROUND.
C     UPDATE THE NEW CLUSTER
   30 NUMBER(NREF) = NUMBER (NREF) + NUMBER(LREF)
      LN = LFIND(LREF,NREF)
      SUM(NREF) = SUM(NREF) + SUM(LREF) + S(LN)
C
C     UPDATE ENTITIES IN THE REDUCED SIMILARITY MATRIX.   THE ENTITIES
C     ARE THE SUM TOTAL OF SIMILARITY VALUES ASSOCIATED WITH ALL
C     PAIRWISE LINKS BETWEEN THE ELEMENTS OF THE TWO CLUSTERS.
C
      DO 40  J = 1, NCL
      I = LIST(J)
      IF(I.EQ.NREF) GO TO 40
C
C     RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C     TESTED FOR EQUALITY WITH LREF.
C
      LL = LFIND(I,LREF)
      LN = LFIND(I,NREF)
      S(LN) = S(LN) + S(LL)
   40 CONTINUE
C
```

```
C     UPDATE THE NEAR AND SREF ARRAYS. IF THE EXTREME ELEMENT IN ROW I
C     WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW
C     EXTREME ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
C
      BIG = SIGN * 1.0E50
      DO 50 J = 1, NCL
      I = LIST(J)
      IF (I.EQ.NREF) GO TO 55
   50 CONTINUE
   55 IF (J.EQ.1) GO TO 80
   60 SREF(I) = BIG
      J1 = J - 1
      DO 70 L = 1, J1
      LISTL = LIST(L)
      LL = LFIND(I, LISTL)
      NTOT = NUMBER(I) + NUMBER(LISTL)
      NTOT = (NTOT * (NTOT - 1))/2
      SREFX = (SUM(I) + SUM(LISTL) + S(LL)) / NTOT
      IF (((SREFX - SREF(I)) * SIGN).GE.0.0) GO TO 70
      NEAR (I) = LISTL
      SREF(I) = SREFX
   70 CONTINUE
   80 J = J + 1
      IF (J.GT.NCL) RETURN
      I = LIST(J)
      IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60
      GO TO 80
      END
```

LIST SYMBOL/METH4

DATE 01/17/78 TIME IS 14:37

SYSTEM/DUMPALL VERSION 2.9.170

```
LASTRECORD =      66
MAXRECSIZEIN =    14    BLOCKSIZEIN =   420
*** EBCDIC ***          UNITS=WORDS
```

```
      SUBROUTINE METH4  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL,
     1                   LREF, NREF, JOB)
C
C
C     HIERARCHICAL CLUATERING BY SINGLE LINKAGE.
C     THE ALGORITHM IS DERIVED FROM:
C     JOHNSON, S.C.(1967) HIERARCHICAL CLUSTERING SCHEMES, PSYCHOMETRIKA
C     , VOLUMN 32, NUMBER 3, SEPTEMBER 1967,  PP241-254.
C
C
      DIMENSION S(1), NEAR(1), SREF(1), LIST(1), A(1), B(1)
      GO TO (10, 15, 20) ,JOB
C
C     JOB = 1, INITIALIZATION.
C
   10 WRITE(6,3000)
 3000 FORMAT (1H+, 49X, 26H SINGLE LINKAGE CLUSTERING , / 51X, 26(1H=))
      RETURN
C
C     JOB = 2 DUMMY ENTRY.
C
   15 RETURN
C
C     JOB = 3, UP DATE FOR NEXT ROUND.
C
   20 CONTINUE
      DO 50 J = 1, NCL
C
C     UPDATE ENTRIES IN S ARRAY ASSOCIATED WITH NREF
C
      I = LIST(J)
      IF(I.EQ.NREF) GO TO 50
C
C     RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C     TESTED FOR EQUALITY WITH LREF
      LL = LFIND(I,LREF)
      LN = LFIND(I,NREF)
      IF(((S(LL) - S(LN)) * SIGN).GE.0.0) GO TO 35
      S(LN) = S(LL)
      IF(I.GT.NREF) GO TO 30
C
C     IF I.LT.NREF
C     CHECK WHETHER S(LN) HAS A BETTER VALUE THAN SREF(NREF)
      IF(((S(LN) - SREF(NREF)) * SIGN).GT.0.0) GO TO 50
      NEAR(NREF) = I
      SREF(NREF) = S(LN)
      GO TO 50
   30 IF(I.GT.LREF) GO TO 40
C
C     IF I.GT.NREF.AND.I.LT.LREF
```

```
      C       CHECK WHETHER S(LN) HAS A BETTER VALUE THAN SREF(I)           0005100
      C                                                                     0005200
              IF(((S(LN) - SREF(I)) * SIGN).GE.0.0) GO TO 50                0005300
              SREF(I) = S(LN)                                               0005400
              NEAR(I) = NREF                                                0005500
              GO TO 50                                                      0005600
           35 IF(I.LT.LREF) GO TO 50                                        0005700
      C                                                                     0005800
      C       IF I.GT.LREF                                                  0005900
      C       UPDATE NEAR ARRAY FOR THOSE ROWS WHOSE EXTREME ELEMENT WAS LREF  0006000
      C                                                                     0006100
           40 IF(NEAR(I).NE.LREF) GO TO 50                                  0006200
              NEAR(I) = NREF                                                0006300
              SREF(I) = S(LN)                                               0006400
           50 CONTINUE                                                      0006500
              RETURN                                                        0006600
              END                                                           0006700
```

```
LIST SYMBOL/METH5

DATE 01/17/78          TIME IS 14:37

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =         75
MAXRECSIZEIN =       14    BLOCKSIZEIN =   420
*** EBCDIC ***             UNITS=WORDS

      SUBROUTINE METH5 (S, NEAR, SREF, LIST, NUMBER, SUM, SREFX, SIGN,
     1                  N, NCL, LREF, NREF, JOB)
C
C     HIERARCHICAL CLUSTERING BY MINIMIZING THE AVERAGE DISTANCE OR
C     MAXIMIZING THE AVERAGE CORRELATION BETWEEN THE MERGED GROUPS.
C
C     THE ALGORITHM IS DERIVED FROM THE *GROUP AVERAGE* METHOD DESCRIBED
C     IN :
C     LANCE, G.N. AND W.T. WILLIAMS, (1967) A GENERAL THEORY OF
C     CLASSIFICATORY SORTING STRATEGIES, 1. HIERARCHICAL SYSTEM, THE
C     COMPUTER JOURNAL, VOLUME 9, NUMBER 4, FEBRUARY 1967, PP 373-380.
C
      DIMENSION S(1), NEAR(1), SREF(1), LIST(1), NUMBER(1), SUM(1)
      GO TO (10, 25, 30), JOB
C
C     JOB = 1, INITIALIZE.
C     NUMBER(I) = NUMBER OF ENTITIES CURRENTLY IN THE I-TH CLUSTER
C
   10 WRITE (6, 2000)
 2000 FORMAT (1H+, 49X, 42H AVERAGE LINKAGE BETWEEN THE MERGED GROUPS ,
     1        / 51X, 42(1H=))
      DO 20 J = 1, N
   20 NUMBER(J) = 1
      RETURN
C
C     JOB = 2, DUMMY ENTRY.
C
   25 RETURN
C
C     JOB = 3, UPDATE FOR NEXT ROUND.
C     UPDATE THE NEW CLUSTER
C
   30 NUMBER(NREF) = NUMBER (NREF) + NUMBER(LREF)
C
C     UPDATE ENTITIES IN THE REDUCED SIMILARITY MATRIX.  THE ENTITIES
C     ARE THE SUM TOTAL OF SIMILARITY VALUES ASSOCIATED WITH ALL
C     PAIRWISE LINKS BETWEEN THE ELEMENTS OF THE TWO CLUSTERS.
C
      DO 40  J = 1, NCL
      I = LIST(J)
      IF(I.EQ.NREF) GO TO 40
C
C     RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C     TESTED FOR EQUALITY WITH LREF.
C
      LL = LFIND(I,LREF)
      LN = LFIND(I,NREF)
      S(LN) = S(LN) + S(LL)
   40 CONTINUE
C
```

```
C       UPDATE THE NEAR AND SREF ARRAYS. IF THE EXTREME ELEMENT IN ROW I
C       WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW
C       EXTREME ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
C
        BIG = SIGN * 1.0E50
        DO 50 J = 1, NCL
        I = LIST(J)
        IF (I.EQ.NREF) GO TO 55
     50 CONTINUE
     55 IF (J.EQ.1) GO TO 80
     60 SREF(I) = BIG
        J1 = J - 1
        DO 70 L = 1, J1
        LISTL = LIST(L)
        LL = LFIND(I, LISTL)
        SREFX = S(LL)/ (NUMBER(I) * NUMBER(LISTL))
        IF (((SREFX - SREF(I)) * SIGN).GE.0.0) GO TO 70
        NEAR (I) = LISTL
        SREF(I) = SREFX
     70 CONTINUE
     80 J = J + 1
        IF (J.GT.NCL) RETURN
        I = LIST(J)
        IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60
        GO TO 80
        END
```

LIST SYMBOL/METH6

DATE 01/17/78   TIME IS 14:37

SYSTEM/DUMPALL VERSION 2.9.170

LASTRECORD =   70
MAXRECSIZEIN =   14   BLOCKSIZEIN =   420
*** EBCDIC ***   UNITS=WORDS

```
      SUBROUTINE METH6  (S, NEAR, SREF, LIST, A, B, SREFX, SIGN, N, NCL,
     1                    LREF, NREF, JOB)
C
C     HIERARCHICAL CLUSTERING BY THE MEDIAN METHOD
C     THE ALGORITHM IS DERIVED FROM :
C     GOWER, J.C., (1967) A COMPARISION OF SOME METHODS OF CLUSTER
C     ANALYSIS, BIOMETRICS, VOLUMN 23, NUMBER 4, DECEMBER 1967,
C     PP623-637.
C
C
      DIMENSION S(1), NEAR(1), SREF(1), LIST(1), A(1), B(1)
      GO TO (10, 15, 20) ,JOB
C
C     JOB = 1, INITIALIZATION.
C
   10 WRITE (6, 2000)
 2000 FORMAT (1H+, 49X, 44H MEDIAN METHOD OF GOWER, BEWARE OF REVERSALS,
     1        / 51X, 44(1H=))
      RETURN
C
C     JOB = 2, DUMMY ENTRY.
C
   15 RETURN
C
C     JOB = 3, UPDATE FOR NEXT ROUND.
   20 LBET = LFIND(LREF, NREF)
      DO 30 J = 1, NCL
      I = LIST(J)
      IF (I.EQ.NREF) GO TO 30
C     RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C     TESTED FOR EQUALITY WITH LREF.
C
      LL = LFIND(I,LREF)
      LN = LFIND(I,NREF)
      IF(SIGN) 25, 25, 27
C
C     IF S IS AN INCREASING FUNCTION OF SIMILARITY (E.G.CORRELATION)THEN
C
   25 S(LN) = (S(LN) + S(LL))/2 + (1.0 - S(LBET))/4
      GO TO 30
C
C     IF S IS A DECREASING FUNCTION OF SIMILARITY (E.G. DISTANCE) THEN
C
   27 S(LN) = (S(LN) + S(LL))/2 - S(LBET) / 4
   30 CONTINUE
C
C     UPDATE THE NEAR AND SREF ARRAYS. IF THE EXTREME ELEMENT IN ROW I
C     WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW
C     EXTREME ELEMENT. ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
C
```

```
      BIG = SIGN * 1.0E50                                      0005100
   40 DO 50 J = 1, NCL                                         0005200
      I = LIST(J)                                              0005300
      IF (I.EQ.NREF) GO TO 55                                  0005400
   50 CONTINUE                                                 0005500
   55 IF(J.EQ.1) GO TO 80                                      0005600
   60 SREF(I) = BIG                                            0005700
      J1 = J-1                                                 0005800
      DO 70 L = 1, J1                                          0005900
      LISTL = LIST(L)                                          0006000
      LL = LFIND(I,LISTL)                                      0006100
      IF(((S(LL) - SREF(I))*SIGN).GE.0.0) GO TO 70             0006200
      NEAR(I) = LISTL                                          0006300
      SREF(I) = S(LL)                                          0006400
   70 CONTINUE                                                 0006500
   80 J = J + 1                                                0006600
      IF(J.GT.NCL) RETURN                                      0006700
      I = LIST(J)                                              0006800
      IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60          0006900
      GO TO 80                                                 0007000
      END                                                      0007100
```

```
LIST SYMBOL/METH7

DATE 01/17/78          TIME IS 14:37

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =        74
MAXRECSIZEIN =      14     BLOCKSIZEIN =   420
*** EBCDIC ***             UNITS=WORDS

      SUBROUTINE METH7  (S, NEAR, SREF, LIST, NUMBER, SUM, SREFX, SIGN, 0000100
     1                  N, NCL, LREF, NREF, JOB)                        0000200
C                                                                       0000300
C     HIERARCHICAL CLUSTREING BY THE METHOD OF  :                       0000400
C     WARD, J.H.,JR, (1963) HIERARCHICAL GROUPING TO OPTIMISE AN        0000500
C     OBJECTIVE FUNCTION, JOURNAL OF THE AMERICAN STATISTICAL           0000600
C     ASSOCIATION, VOLUME 58, 1963, PP 236-244.                         0000700
C                                                                       0000800
C                                                                       0000900
C     THE PARTICULAR ALGORITHM USED HERE IS DESCRIBED IN                0001000
C     WISHART, D., (1969) AN ALGORITHM FOR HIERARCHICAL CLASSIFICATION, 0001100
C     BIOMETRICS, VOLUME 22, NUMBER 1, MARCH 1969, PP 165-170.          0001200
C                                                                       0001300
C                                                                       0001400
      DIMENSION S(1), NEAR(1), SREF(1), LIST(1), NUMBER(1), SUM(1)      0001500
      GO TO (10, 25, 30), JOB                                           0001600
C     JOB = 1, INITIALIZE.                                              0001700
C     NUMBER (I) = NUMBER ENTITIES CURRENTLY IN THE I-TH CLUSTER        0001800
   10 WRITE (6, 2000)                                                   0001900
 2000 FORMAT (1H+, 49X, 44H HIERARCHICAL GROUPING BY THE METHOD OF WARD,0002000
     1       / 51X, 41(1H=))                                            0002100
      DO 20 J = 1, N                                                    0002200
   20 NUMBER(J) = 1                                                     0002300
      SUM(1) = 0.0                                                      0002400
      RETURN                                                            0002500
C                                                                       0002600
C     JOB = 2, CALAULATE OBJECTIVE FUNCTION VALUE                       0002700
C                                                                       0002800
   25 SUM(1) = SUM(1) + SREFX / 2.0                                     0002900
      SREFX = SUM(1)                                                    0003000
      RETURN                                                            0003100
C                                                                       0003200
C     JOB = 3, UPDATE FOR NEXT ROUND.                                   0003300
C                                                                       0003400
   30 LBET = LFIND(LREF,NREF)                                           0003500
      NTOT = NUMBER(LREF) + NUMBER(NREF)                                0003600
      DO 40 J = 1, NCL                                                  0003700
      I = LIST (J)                                                      0003800
      IF (I.EQ.NREF) GO TO 40                                           0003900
C                                                                       0004000
C                                                                       0004100
C     RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE      0004200
C     TESTED FOR EQUALITY WITH LREF.                                    0004300
      LL = LFIND(I, LREF)                                               0004400
      LN = LFIND(I, NREF)                                               0004500
      S(LN) = (S(LN) * (NUMBER(I) + NUMBER(NREF)) + S(LL) * (NUMBER(I)  0004600
     1   + NUMBER(LREF)) - S(LBET) * NUMBER(I)) / (NTOT+NUMBER(I))0004700
   40 CONTINUE                                                          0004800
      NUMBER(NREF) = NTOT                                               0004900
C                                                                       0005000
```

```
C       UPDATE THE NEAR AND SREF ARRAY. IF THE EXTREME ELEMENT IN ROW I
C       WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW
C       EXTREME ELEMENT. ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
C
        BIG = SIGN * 1.0E50
        DO 50 J = 1,NCL
        I = LIST(J)
        IF(I.EQ.NREF) GO TO 55
     50 CONTINUE
     55 IF (J.EQ.1) GO TO 80
     60 SREF(I) = BIG
        J1 = J - 1
        DO 70 L = 1, J1
        LISTL = LIST(L)
        LL = LFIND(I, LISTL)
        IF (((S(LL) - SREF(I)) * SIGN).GE.0.0) GO TO 70
        NEAR(I) = LISTL
        SREF(I) = S(LL)
     70 CONTINUE
     80 J = J + 1
        IF (J.GT.NCL) RETURN
        I = LIST (J)
        IF (NEAR(I).EQ.LREF.OR. NEAR(I).EQ.NREF) GO TO 60
        GO TO 80
        END
```

LIST SYMBOL/TREE

DATE 01/17/78   TIME IS 14:37

SYSTEM/DUMPALL VERSION 2.9.170

LASTRECORD =   264
MAXRECSIZEIN =   14   BLOCKSIZEIN = 420
*** EBCDIC ***   UNITS=WORDS

```
C
C     SUBROUTINE TREE
C
        SUBROUTINE TREE (I, J, S, IL, JL, NEXT, A, LABEL, LCLNO, LINE,
     1                   IS, LAST, EPS, TITLE, N, KBEG, NT, INTRV, IPRNT,
     2                   MAXIN)
      DIMENSION I(N), J(N), S(N), IS(N), IL(N), JL(N), NEXT(N),
     1          LCLNO(N), LAST(N), A(25,N)
      DIMENSION LINE(MAXIN), LABEL(2,MAXIN), EPS(25), TITLE(16)
      DATA BARI, BLNKI, BARS, BLANK /4H---I, 4H   I, 4H----, 4H    /
C
C     DEFAULT VALUES
C
      IF (KBEG.LT.1) KBEG = 1
      IF (INTRV.LT.1.OR.INTRV.GT.3) INTRV = 1
      IF (NT.LE.0) NT = 6
C
C     INITIALIZE ARRAYS
C
      NOBJ = N + 1
      DO 10 K = 1,NOBJ
      LINE(K) = 0
      LCLNO(K) = 0
      LAST(K) = 0
      DO 10 L = 1, 25
      A(L,K) = BLANK
   10 CONTINUE
C
C     SEGMENT THE S ARRAY
C
      GO TO (20, 40, 120), INTRV
C
C     CONSTRUCT INTRVALS OF EQUAL LENGTH
C
   20 RANGE = S(N) - S(KBEG)
      DELTA = RANGE / 25
      EPS(1) = S(KBEG) + DELTA
      DO 30 K = 2, 24
   30 EPS(K) = EPS(K-1) + DELTA
      EPS(25) = S(N)
C
C     CONSTRUCT THE IS ARRAY
C
   40 IF (EPS(1).GT.EPS(2)) GO TO 70
C
C     S INCREASE WITH DISSIMILARITY (AS DOES A DISTANCE)
C
      KK = 1
      DO 60 K = 1,N
   50 IF (S(K).LE.EPS(KK)) GO TO 60
```

200

```
      IF (KK.EQ.25) GO TO 60                                            0005100
      KK = KK + 1                                                       0005200
      GO TO 50                                                          0005300
   60 IS(K) = KK                                                        0005400
      GO TO 120                                                         0005500
C                                                                       0005600
C     S DECREASE WITH DISSIMILARITY (AS DOES A CORRELATION)             0005700
C                                                                       0005800
   70 KK = 24                                                           0005900
      KKK = 25                                                          0006000
      NN = N + 1                                                        0006100
      DO 90 K = 1,N                                                     0006200
      KCOMP = NN - K                                                    0006300
   80 IF (S(KCOMP).LT.EPS(KK)) GO TO 90                                 0006400
      KKK = KK                                                          0006500
      KK = KK - 1                                                       0006600
      IF (KK.EQ.0) GO TO 100                                            0006700
      GO TO 80                                                          0006800
   90 IS(KCOMP) = KKK                                                   0006900
  100 DO 110 K = 1,KCOMP                                                0007000
  110 IS(K) = 1                                                         0007100
C                                                                       0007200
C     PRINT INPUT TO TREE                                               0007300
C                                                                       0007400
  120 WRITE (NT,2000) TITLE                                             0007500
      WRITE (NT,2100) KBEG, N                                           0007600
      WRITE(NT, 2200)                                                   0007700
      WRITE(NT, 2300)                                                   0007800
      M = 1                                                             0007900
      WRITE(NT, 2400) M, S(KBEG), EPS(M)                                0008000
      DO 130 M = 2, 25                                                  0008100
      MM = M - 1                                                        0008200
  130 WRITE(NT, 2400) M, EPS(MM), EPS(M)                                0008300
      IF (IABS(IPRNT).EQ.1) GO TO 150                                   0008400
C                                                                       0008500
C     PRINT THE CLUSTER MERGE DATA                                      0008600
C                                                                       0008700
      WRITE(NT,2000) TITLE                                              0008800
      WRITE (NT, 2500)                                                  0008900
      DO 140 K = KBEG, N                                                0009000
      WRITE (NT, 2600) K, I(K), J(K), S(K), IS(K), IL(K), JL(K), NEXT(K)0009100
  140 CONTINUE                                                          0009200
C                                                                       0009300
C     START TREE WITH THE MOST SIMILAR PAIR                             0009400
C                                                                       0009500
  150 K = KBEG                                                          0009600
      LNO = 0                                                           0009700
C                                                                       0009800
C     MERGE CLUSTERS I(K) AND J(K)                                      0009900
C                                                                       0010000
  160 IK = I(K)                                                         0010100
      JK = J(K)                                                         0010200
C                                                                       0010300
C     SET LINE NUMBERS FOR OUTPUT                                       0010400
C                                                                       0010500
      IF (IL(K).NE.0) GO TO 170                                         0010600
      LNO = LNO + 1                                                     0010700
      LINE(IK) = LNO                                                    0010800
      LCLNO(LNO) = IK                                                   0010900
  170 IF (JL(K).NE.0)GO TO 180                                          0011000
      LNO = LNO + 1                                                     0011100
```

```
      LINE(JK) = LNO
      LCLNO(LNO) = JK
C
C     FILL IN THE PRINT LINES
C
  180 ISK = IS(K)
      KT = 0
      ITEM = IK
  190 LITEM = LINE(ITEM)
      IF (ISK - LAST(LITEM) - 1)  225, 200, 210
C
C     ADD ONLY ONE MORE SEGMENT FOR LINE(ITEM)
C
  200 A(ISK,LITEM) = BARI
      LAST(LITEM) = ISK
      GO TO 225
C
C     ADD MORE THAN ONE SEGMENT
C
  210 LBEG = LAST(LITEM) + 1
      LEND = ISK - 1
      DO 220 L = LBEG, LEND
  220 A(L, LITEM) = BARS
      GO TO 200
C
C     REPEAT FOR CLUSTER J(K)
C
  225 KT = KT + 1
      IF (KT.NE.1) GO TO 230
      ITEM = JK
      GO TO 190
C
C     TAKE CARE OF ANY LINES BETWEEN I(K) AND J(K)
C
  230 LIK = LINE(IK)
      LJK = LINE(JK)
      IF (LIK.GT.LJK) GO TO 240
      LBOT = LJK
      LTOP = LIK
      GO TO 250
  240 LBOT = LIK
      LTOP = LJK
  250 IF (LBOT.EQ.(LTOP+1)) GO TO 270
C
C     MUST FILL IN SOME VERTICAL CONNECTIONS
C
      LBEG = LTOP + 1
      LEND = LBOT - 1
      DO 260 L = LBEG, LEND
      IF (A(ISK,L).EQ.BARI) GO TO 260
      A(ISK,L) = BLNKI
      LAST(L) = ISK
  260 CONTINUE
C
C     UPDATE LINE NUMBER FOR NEW CLUSTER
C
  270 LINE(IK) = (LINE(IK) + LINE(JK)) / 2
C
C     MERGE COMPLETE. FIND NEXT STAGE
C
      KLAST = K
```

```
      K = NEXT(K)
      IF(K.GT.N.OR.K.LT.KBEG) GO TO 400
      IF(IL(K).LE.0) GO TO 280
      IF(JL(K).LE.0) GO TO 290
      GO TO 300
  280 IL(K) = -IL(K)
      GO TO 160
  290 JL(K) = -JL(K)
      GO TO 160
C
C     THIS MERGE INVOLVES CLUSTER THAT EACH HAVE MORE THAN ONE MEMBER.
C     BACKTRACK TO THE ROOT OF THE TREE ALONG THE UNEXPLORED BRANCH.
C
  300 IF(IL(K).EQ.KLAST) GO TO 310
C
C     GO DOWN IL(K) BRANCH. SET JL(K) SO WE KNOW NOT TO GO DOWN THAT
C     BRANCH AGAIN.
C
      JL(K) = -JL(K)
      K = IL(K)
      GO TO 320
C
C     GO DOWN JL(K)BRANCH. SET IL(K) SO WE KNOW NOT TO GO DOWN THAT
C     BRANCH AGAIN.
C
  310 IL(K) = -IL(K)
      K = JL(K)
  320 IF (K.LT.1.OR.K.GT.N) GO TO 600
C
C     TEST TO SEE IF THE END HAS BEEN REACHED. IL(K)=JL(K) IF BOTH ZERO.
C
      IF (IL(K)- JL(K)) 330, 160, 350
  330 IF (IL(K).EQ.0) GO TO 360
  340 K = IL(K)
      GO TO 320
  350 IF(JL(K).EQ.0) GO TO 340
  360 K = JL(K)
      GO TO 320
C
C     PRINT THE TREE
C
  400 WRITE (NT, 2000) TITLE
      WRITE (NT, 3000) (K, K =1, 25)
      IF (LABEL(1,1).EQ.5HNOLAB) GO TO 420
      DO 410 L = 1,LNO
      LL = LCLNO(L)
  410 WRITE (NT, 3100) (LABEL(K,LL), K = 1, 2), LL, (A(K,L), K = 1, 25)
      GO TO 440
C
C     LEAVE LABLE SPACE BLANK
C
  420 DO 430 L = 1, LNO
      LL = LCLNO(L)
  430 WRITE (NT, 3200) LL, (A(K,L), K = 1, 25)
C
C     TREE COMPLETE
C
  440 WRITE (NT, 3000) (K, K = 1, 25)
      ENDFILE NT
      RETURN
C
```

```
C       ERROR. PRINT AS MUCH OF THE TREE AS HAS BEEN CONSTRUCTED
C
  600 WRITE (NT, 6000) KLAST, K
      GO TO 400
 2000 FORMAT (1H1, 10X, 10(1H*), 16A5, 10(1H*)/)
 2100 FORMAT(/5X, 65H THIS RUN DEPICTS THE PORTION OF THE TREE GENERATED
     1 BETWEEN STAGE, I5, 10H AND STAGE, I5, 19H OF THE CLUSTERING./)
 2200 FORMAT(/5X, 63H THE CRITERION VALUES ARE SEGMENTED INTO THE FOLLOW
     1ING CLASSES./)
 2300 FORMAT (10X, 6H CLASS, 4X, 11HLOWER BOUND, 5X, 11HUPPER BOUND)
 2400 FORMAT(/ 10X, I5, 1X, 2E16.8)
 2500 FORMAT (/6X, 'STAGE', 5X, 'LOWER', 5X, 'HIGHER', 5X, 'VALUE OF ',
     1         'CRITERION', 4X, 'CLASS WHERE', 4X, 'STAGE WHERE', 4X,
     2         'STAGE WHERE', 4X, 'STAGE WHERE',
     3         /6X, 'OF', 8X, 'CLUSTER', 3X, 'CLUSTER', 4X, 'ASSOCIATED',
     4         ' WITH', 7X, 'MERGE OCCURE', 3X, 'I WAS LAST', 5X, 'J WAS'
     5         ' LAST',  5X, 'I IS NEXT',
     6         /6X, 'MERGE', 5X, 'ID. NO.', 3X, 'ID. NO.', 4X,'THE MERGE'
     7         , 13X, '(1--25)', 8X, 'IN A MERGE', 5X, 'IN A MERGE', 5X,
     8         'IN A MERGE', /
     8         /8X, 'K', 9X, 'I', 9X, 'J', 14X, 'S', 20X, 'IS', 13X, 'IL'
     9         , 13X, 'JL', 12X, 'NEXT')
 2600 FORMAT (/ 3I10, 5X, E16.8, 4I15)
 3000 FORMAT (10H0ITEM NAME,  2X, 5HID NO, 2X, 25I4)
C
C
C
 3100 FORMAT (1H0, 2A5, I6, 2X, 25A4)
 3200 FORMAT (1H0, 10X, I6, 2X, 25A4)
 6000 FORMAT (37H0ERROR. WHILE BACKTRACKING FROM KLAST, I6,
     1         26H K WAS FOUND OUT OF RANGE., /1X, 3HK = , I20)
      END
```

```
LIST SYMBOL/SEFWIG

DATE 01/17/78           TIME IS 14:38

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =      250
MAXRECSIZEIN =    14     BLOCKSIZEIN =   420
*** EBCDIC ***           UNITS=WORDS

C
FILE  20 = NTIN, UNIT = READER
FILE   6 = FILE6, UNIT = PRINTER
FILE  16 = FILE16, UNIT = PRINTER
FILE  30 = NTMERG , UNIT = DISK
 $ SET AUTOBIND
 $ BIND = FROM CODON/=
C      PROGRAM SEFWIG  (SELETED ERROR FOR WITHIN GROUP )
C
C      THIS PROGRAM IS TO AID IN ANALYZING THE WITHIN CLUSTER VARIANCE
C      FOR SELECTED VARIABLES.
C
C
C      -----------------------------------------------------------------
C
C
C      INPUT SPECIFICATIONS
C
C      CARD  1 TITLE CARD
C         TO ENABLE MULTIPLE RUN, THE FIRST 5 COLUMNS (COLS 1TO 5) OF THI
C         MUST NOT CONTAIN THE SAME WORD AS THE NEXT 5 COLUMNS (COLS 6 TO
C         OR ELSE IT WILL BE READ AS THE END CARD. SEE CARD 7
C
C      CARD  2 PARAMETER CARD
C         COLS   1- 5  NTIN    = INPUT UNIT FOR DATA
C         COLS   6-10  NTMERG  = INPUT UNIT FOR MERGE SPECIFICATIONS
C         COLS  11-15  NE      = NUMBER OF ENTITIES (DATA UNITS)  (MAX 160)
C         COLS  16-20  NC      = NUMBER OF VARIABLES USED IN THIS RUN
C                                (MAX 15)
C         COLS  21-30  PRMAX   = PROBABILITY LEVEL FOR OVERALL F-VALUE
C                                WHERE PRINT BEGIN
C                                READ IN FORMAT IS F10.7
C                                DEFAULT VALUE IS 0.10000
C         COLS  31-40  PRMIN   = PROBABILITY LEVEL FOR OVERALL F-VALUE
C                                WHERE PRINT STOP
C                                READ IN FORMAT IS F10.7
C                                DEFAULT VALUE IS 0.00000
C
C      CARD  3 FORMAT CARD CONTAINING FMTIN FOR READING DATA
C
C      CARD(S) 4 LABEL CARD  (THERE MUAT BE NC OF THESE CARDS)
C
C         COLS   1- 5  LABEL(I) = 5 CHARACTER LABEL FOR THE I-TH VARIABLE
C
C      CARD(S) 5 DATA CARD   ORIGINAL DATA SET UP ACCORDING TO FMTIN
C
C ****INCLUDE CARD 5 ONLY IF ORIGINAL DATA IS ON CARDS  E.G. NTIN = 5
C
C      CARD(S) 6 MERGE DATA CARD CONTAINING MERGE DATA AS GENERATED BY
C                             PROGRAM  *CLUSAN*
```

205

```
C           COLS  1-10  K  = STAGE OF CLUSTERING
C           COLS 11-20  II = LOWER NUMBERED CLUSTER MERGED AT STAGE K
C           COLS 21-30  JJ = UPPER NUMBERED CLUSTER MERGED AT STAGE K
C           COLS 31-46  CR = VALUE OF CRITERION ASSOCIATED WITH THIS MERGE
C ****INCLUDE CARD 6 ONLY IF MERGE DATA IS ON CARDS  E.G. NTMERG = 5
C
C        CARD  7 END CARD
C           COLUMN 1 TO 5 MUST CONTAIN THE SAME WORD AS COLUMN 6 TO 10,
C           SUCH AS LEAVING COLUMN 1 TO 10 BLANK.  THIS WILL SIGNAL THE PRO
C           TO TERMINATE.
C
C
C        ----------------------------------------------------------------
C
C        OTHER VARIABLES IN THE PROGRAM
C
C           GSS(I)     = UNADJUSTED TOTAL SUM OF SQUARES FOR VARIABLE I
C           TSS(I)     =   ADJUSTED TOTAL SUM OF SQUARES FOR VARIABLE I
C           WSS(I)     = WITHIN CLUSTER SUM OF SQUARES FOR VARIABLE I AT
C                        CURRENT STAGE
C           SUM(I)     = SUM OF VALUES FOR VARIABLE I
C           NUMBER(I) = NUMBER OF ENTITIES CURRENTLY IN CLUSTER J
C           UNEXP(I)   = UNEXPLAINED PORTION OF VARIANCE FOR VARIABLE I
C                      = WSS(I) / TSS(I)
C           NCL        = NUMBER OF CLUSTER AT CURRENT STAGE
C
C
C
C        ----------------------------------------------------------------
C
      DIMENSION NUMBER(160), DATA(2400), LABEL(15), TSS(15), WSS(15),
     1          GSS(15), SUM(15), UNEXP(15), FMTIN(16), TITLE(16)
      DIMENSION ASS(15), F(15), PF(15)
      INTEGER FIRST
C
C
   10 READ (5, 1100) TITLE
      WRITE (16, 1200)
      IF (TITLE(1).IS.TITLE(2)) GO TO 110
      WRITE (6, 2200) TITLE
      READ (5,1000) NTIN,NTMERG, NE, NC, PRMAX, PRMIN
      WRITE (6, 2000) NTIN, NTMERG, NE, NC
      READ (5,1100) FMTIN
      WRITE (6, 2100) FMTIN
      READ (5, 1300) (LABEL(I), I = 1, NC)
C
C        INITIALIZE  AND SET DEFAULT VALUES
C
      IF (PRMAX.LE.0.OR.PRMAX.GT.1) PRMAX = 0.10000
      IF (PRMIN.LT.0.OR.PRMIN.GE.1) PRMIN = 0.00000
      NCL = NE - 1
      DO 20 I = 1,NC
      GSS(I) = 0.0
   20 SUM(I) = 0.0
      LAST = 0
      K = 0
      DO 30 I = 1,NE
      NUMBER(I) = 1
      FIRST = LAST + 1
      LAST = LAST + NC
      READ (NTIN, FMTIN) (DATA(J), J = FIRST,LAST)
```

```
      DO 30 J = 1, NC
      K = K + 1
      GSS(J) = GSS(J) + DATA(K) * DATA(K)
   30 SUM(J) = SUM(J) + DATA(K)
      DO 40 J = 1,NC
      TSS(J) = GSS(J) - SUM(J) * SUM(J) / NE
      WSS(J) = 0.0
   40 UNEXP(J) = 0.0
C
C     WRITE OUT TSS ARRAY
C
      WRITE (6,3000)
      DO 50 J = 1,NC
   50 WRITE (6, 3100) LABEL(J), TSS(J)
C
C     WRITE PAGE HEADINGS
C
      WRITE (6, 2700) TITLE
      WRITE (6,2400)
      WRITE (6, 2500) (LABEL(I), I = 1,NC)
      WRITE (16, 2700) TITLE
      WRITE (16, 2310) PRMIN, PRMAX
      WRITE (16, 2410)
      WRITE (16, 2510) (LABEL(I), I = 1, NC)
C
C     READ MERGE DATA
C
C     IF MERGE DATA IS ON DISK AS GENERATED BY *CLUSAN* THEN READ WILL
C     BEGIN FROM THE 8TH RECORD, SO AS TO SKIP THE HEADING OF THE MERGE
C     DATA FILE AND FROM THEN ON ONLY ALTERNATE RECORD WILL BE READ AS
C     SET UP BY THE WRITE FORMAT OF *CLUSAN FOR THE MERGE DATA FILE.
C
C     IF THE MERGE DATA ARE ON CARDS AND ARE READ IN BY CARD READER THEN
C     READING WILL COMMENCE ON THE FIRST CARD AND EVERY FOLLOWING CARDS.
C
      INQUIRE (NTMERG, KIND = KND)
      ICON = 0
      ICOE = 1
      IF (KND.NE.9) ICON = 8
      IF (KND.NE.9) ICOE = 2
      NSTGS = NE - 1
C
C     READ LAST CRITERION AS TOTAL SUM OF SQUARE (TCR)
C
      LASREC = (ICOE * NSTGS) + ICON
      READ (NTMERG = LASREC, 4000) K, II, JJ, TCR
      REWIND(NTMERG)
      DO 90 I = 1,NSTGS
      IREC = (ICOE * I) + ICON
      READ (NTMERG = IREC, 4000) K, II, JJ, CR
C
C     TEST FOR PROPER SEQUENCE OF MERGE INSTRUCTIONS.
C
      IF (K.NE.I) GO TO 100
C
C     UPDATE CLUSTER INFORMATION
C
      IWRD = (II - 1) * NC
      JWRD = (JJ - 1) * NC
      NTOT = NUMBER(II) + NUMBER(JJ)
      DO 60 J = 1,NC
```

```
      IWRDJ = IWRD + J
      JWRDJ = JWRD + J
      DIFF = NUMBER(II) * DATA(JWRDJ) - NUMBER(JJ) * DATA(IWRDJ)
      WSS(J) = WSS(J) + DIFF * DIFF/ (NUMBER(II) * NUMBER(JJ) * NTOT)
      UNEXP(J) = WSS(J)/ TSS(J)
   60 DATA(IWRDJ) = DATA(IWRDJ) + DATA(JWRDJ)
      NUMBER(II) = NTOT
   70 WRITE (6, 2600) K, NCL, II, JJ, CR, (UNEXP(J), J = 1, NC)
C
C     CALCULATE F-VALUES AND PROBABILITIES
C
      DFA = NCL - 1
      IF (DFA.LE.0) DFA = 1
      DFW = K
      ACR = TCR - CR
      FCR = (ACR / CR) * (DFW / DFA)
      PCR = PRBF(DFA, DFW, FCR)
      IF (PCR.GT.PRMAX.OR.PCR.LT.PRMIN) GO TO 90
      DO 210 J = 1,NC
      IF (WSS(J).EQ.0)WSS(J) = 0.00000000001
      ASS(J) = TSS(J) - WSS(J)
      F(J) = (ASS(J) / WSS(J)) * (DFW / DFA)
      PF(J) = PRBF(DFA, DFW, F(J))
  210 CONTINUE
      WRITE (16, 2600) K, NCL, II, JJ, FCR, (F(J), J = 1,NC)
      WRITE (16, 2610) PCR,(PF(J), J = 1, NC)
   90 NCL = NCL - 1
      WRITE (6, 2500) (LABEL(I), I = 1,NC)
      WRITE (16, 2510) (LABEL(I), I = 1, NC)
      GO TO 10
C
C     ERROR IN SEQUENCE OF MERGE INSTRUCTIONS
C
  100 WRITE (6,5000) I,K
      I = I - 1
      WRITE (6, 5100) I, (UNEXP(J), J = 1,NC)
      GO TO 10
  110 CONTINUE
      WRITE (16, 2200) TITLE
      WRITE (16, 4500)
      STOP
C
 1000 FORMAT (4I5, 2F10.7)
 1100 FORMAT (16A5)
 1200 FORMAT (1X)
 1300 FORMAT (A5)
C
 2000 FORMAT (/10X, 9HNTIN     = , I6/10X, 9HNTMERG = , I6/10X,
     1        9HNE       = , I6, /10X, 9HNC       = , I6, /)
 2100 FORMAT (/10X, 15HINPUT FORMAT : , 16A5//)
 2200 FORMAT (1H1, /20X, 32(1H$), 18H  PROGRAM SEFWIG  ,32(1H$),/
     1        20X, 1H$, 80X, 1H$/20X, 1H$, 80X, 1H$/20X, 1H$, 16A5, 1H$/
     2        20X, 1H$, 80X, 1H$/20X, 1H$, 80X, 1H$/20X, 82(1H$)///)
 2310 FORMAT (//20X, 'F-VALUES AND PROBABILITIES AT DIFFERENT STAGES',
     1        ' OF CLUSTERING.'//20X, 'OVERALL PROBABILITIES BETWEEN',
     2        F10.7,' AND', F10.7,' ONLY WILL BE LISTED ')
 2400 FORMAT (//30X, 40HPROPORTION OF VARIANCE NOT EXPLAINED BY   ,
     1        20HCLUSTERING (WSS/TSS)/ )
 2410 FORMAT (//20X, 'F-VALUES (AMS/WMS) ON FIRST LINE, AND ',
     1        'PROBABILITIES ON SECOND LINE.' / 20X,'THE OVERALL COLUMN'
     2        ,' WILL BE MEANINGLESS, UNLESS CLUSTER IS PERFORMED BY ', 
```

```
     3         'METH7 OF *CLUSAN*'/)
 2500 FORMAT(/29H    K NCL  II  JJ   CRITERION , 15(1X, A5))
 2510 FORMAT(/29H    K NCL  II  JJ    OVERALL  , 15(1X, A5))
 2600 FORMAT (/I4, 3I4, E13.6, 15F6.3)
 2610 FORMAT (2X, 'PROBABILITIES', 1X, E13.5, 15F6.3)
 2700 FORMAT (1H1, 4X, 10(1H*), 5X, 16A5, 5X, 10(1H*))
 3000 FORMAT (//24X, 9HVARIABLE ,3X, 19HTOTAL SUM OF SQUARE)
 3100 FORMAT (/25X, A5, 6X, E16.8)
 4000 FORMAT (3I10, 5X, E16.8)
 4500 FORMAT (1H1, //42X, 37(1H*)/ 42X, 37(1H*) ///42X,
     1         37HPROGRAM SEFWIG BY S.H. TEOW JUNE 1977 ///
     2         42X, 37(1H*)/ 42X, 37(1H*))
C
 5000 FORMAT (//10(1H*) ,24H ERROR IN MERGE SEQUENCE,/10X, 5H I = , I10/
     1         5H K = , I10)
 5100 FORMAT (//1X, 28HRESULTS FOR PRECEDING STAGE , /11X, I5, 14X,
     1         15F6.3)
      END
```

```
LIST SYMBOL/POSTCA

DATE 01/17/78          TIME IS 14:38

SYSTEM/DUMPALL VERSION  2.9.170

LASTRECORD =       232
MAXRECSIZEIN =     14     BLOCKSIZEIN =   420
*** EBCDIC ***            UNITS=WORDS

FILE  20 = NTIN, UNIT = DISK                                          0000100
 $ SET AUTOBIND                                                       0000200
 $ BIND = FROM CODON/=                                                0000300
C                                                                     0000400
C       PROGRAM POSTCA                                                0000500
C                                                                     0000600
C       ==============================================================0000700
C                                                                     0000800
C       THIS PROGRAM IS DESIGNED TO ASSIST IN THE INTERPRETATION OF   0000900
C       CLUSTERED DATA UNITS.  ORGINAL DATA IS PERMUTED TO THE SEQUENCE 0001000
C       APPEARING IN THE HIERARCHICAL TREE (OR ANY SEQUENCE THE USER  0001100
C       WISHES TO SPECIFY).                                           0001200
C       CLUSTERS ARE IDENTIFIED BY STATING THE NUMBER OF DATA UNITS IN 0001300
C       EACH CLUSTER, SAY N1, N2, ETC.  THEN THE FIRST N1 UNITS IN THE 0001400
C       SEQUENCE LIST ARE IN THE FIRST CLUSTER, THE NEXT N2 UNITS IN THE 0001500
C       SECOND CLUSTER AND SO FORTH.                                  0001600
C       EACH CLUSTER IS DESCRIBED BY A LISTING OF ITS DATA UNITS,  THEIR 0001700
C       SCORES ON SELECTED VARIABLES AND SUMMARY STATISTICS.          0001800
C       TESTS OF SIGNIFICANT DIFFERENCE OVER MEAN OF CLUSTERS CAN     0001900
C       OPTIONALLY BE PERFORMED, THIS CAN EITHER BE LEAST SIGNIFICANT 0002000
C       DIFFERENCE OR DUNCAN'S MULTIPLE RANGES TEST AT 5% OR 5% AND 1% 0002100
C       SIGNIFICANT LEVELS.                                           0002200
C       THE PRINTED OUTPUT IS LIMITED TO 11 VARIABLES EACH RUN.  IF MORE 0002300
C       THAN 11 VARIABLES ARE OF INTEREST,  SIMPLY PARTITION THE VARIABLES0002400
C       INTO SUBSETS AND RUN THE PROGRAM FOR EACH SUBSET.             0002500
C                                                                     0002600
C       ==============================================================0002700
C                                                                     0002800
C                                                                     0002900
C       INPUT SPECIFICATIONS                                          0003000
C                                                                     0003100
C           CARD  1    TITLE CARD                                     0003200
C                THE FIRST 5 COLUMNS SHOULD NOT CONTAIN THE SAME WORDS AS 0003300
C                THE NEXT 5 COLUMNS, ELSE IT WOULD BE READ AS END CARD. 0003400
C                                                                     0003500
C           CARD  2    PARAMETER AND OPTION CARD                      0003600
C                COLS  1 - 5     NE = NUMBER OF ENTITIES (DATA UNITS) 0003700
C                COLS  6 -10     NC = NUMBER OF CHARACTERS  (MAX. 11) 0003800
C                COLS 11 -15    NCL = NUMBER OF CLUSTERS (MAX. 160)   0003900
C                COLS 16 -20   NTIN = INPUT FILE UNIT FOR DATA        0004000
C                COLS 21 -25   ISIG = OPTION FOR SIGNIFICANT TEST OF CLUSTERS0004100
C                                     MEANS.                          0004200
C                                   = 1 FOR LEAST SIGNIFICANT DIFFERENT TEST 0004300
C                                   = 2 FOR DUNCAN'S MULTIPLE RANGES TEST 0004400
C                                   = OTHER INTEGER FOR NO TEST REQUIRED 0004500
C                COLS 26- 30   IPRO = OPTION FOR LEVELS OF SIGNIFICANT 0004600
C                                     DIFFERENCE                      0004700
C                                   = 1 FOR 5% ONLY                   0004800
C                                   = 2 FOR 5% AND 1%                 0004900
C                                                                     0005000
```

210

```
C         CARD 3    LABEL CARD FOR CHARACTERS.
C                   A 5 LETTERS LABEL IS REQUIRED FOR EACH CHARACTER
C                   (11A5 FORMAT).
C
C         CARD(S) 4 LABEL CARDS FOR DATA UNITS.
C              THERE ARE TWO OPTIONS :
C              1.   INCLUDE 1 CARD WITH 'NOLAB' IN COLUMNS 1-5.
C                   NO LABELS WILL BE PRINTED WITH THE DATA UNITS.
C
C              2.   INCLUDE NE CARDS, COLUMNS 1-10 CONTAINING A LABEL
C                   FOR ONE DATA UNIT.
C
C         CARD(S) 5 SEQUENCE LIST FOR DATA UNITS .
C              USES AS MANY CARDS AS NECESSARY TO LIST NE DATA UNITS.
C              (16I5 FORMAT)
C
C         CARD(S) 6 NUMBER OF DATA UNITS IN EACH CLUSTER .
C              USE AS MANY CARDS AS NECESSARY TO LIST THE SIZE OF THE NCL
C              CLUSTERS WHOSE NUMBERS ARE ORDERED IN THE SEQUENCE LIST OF
C              CARD 5.  (16I5 FORMAT)
C
C         CARD 7    FORMAT FOR  PRINTING DATA ON OUTPUT.
C              GIVE FORMAT FOR NC FIELD OF 10 COLUMNS EACH.  USE ANY
C              COMBINATION OF E, F AND G FIELDS.
C              E.G. --- 11F10.5) ---
C
C         CARD 8    FORMAT FOR READING DATA
C
C         CARD(S) 9 ORIGINAL DATA (IF ON CARDS)
C
C         CARD 10   END CARD
C              THE FIRST 5 COLUMNS SHOULD CONTAIN THE SAME WORDS AS THE
C              NEXT 5 COLUMNS TO TERMINATE THE PROGRAM.
C
C
      DIMENSION TITLE(16), NUMBER(160),FMTIN(16), FMTOUT(19)
      DIMENSION LABELC(11), GTOT(11), GSS(11), CTOT(11), CSS(11),
     1          LABELD(2,160), LIST(160), DATA(11,160)
      DIMENSION MEAN(11,160), A(160), KARHDG(3)
      INTEGER FIRST
      REAL MEAN
      DATA FMTOUT(1), FMTOUT(2), FMTOUT(3)/'(1X, ', '2A5,I', '4,1X,'/
      DATA BLANK/'     '/
    5 READ (5, 1000) TITLE
      WRITE (6, 2010) TITLE
      IF (TITLE(1).IS.TITLE(2)) GO TO 200
      READ (5, 1100) NE, NC, NCL, NTIN,ISIG, IPRO
      READ (5, 1000) (LABELC(I), I = 1,NC)
      READ (5, 1000) (LABELD(I,1), I = 1, 2)
      IF (LABELD(1,1).EQ.'NOLAB') GO TO 15
C
C     READ REMAINING LABELS
C
      DO 10 J = 2, NE
   10 READ (5, 1000) (LABELD(I,J), I = 1, 2)
      GO TO 20
   15 CONTINUE
      DO 17 J = 1, NE
      DO 17 I = 1, 2
      LABELD(I,J) = BLANK
   17 CONTINUE
```

```
      20 READ (5, 1200) (LIST(I), I = 1, NE)
         READ (5, 1200) (NUMBER(I), I = 1, NCL)
         READ (5, 1000) (FMTOUT(I), I = 4,19)
         READ (5, 1000) (FMTIN(I), I = 1, 16)
C
C
         WRITE (6, 2100) NE, NC, NCL, NTIN, ISIG, IPRO
         WRITE (6, 2200) (I, NUMBER(I), I = 1,NCL)
         WRITE (6, 2310) (FMTOUT (I), I = 1, 19)
         WRITE (6 , 2300) (FMTIN(I),  I = 1,16)
C
C        READ DATA SET
C
         DO 25 J = 1, NE
      25 READ (NTIN, FMTIN) (DATA(I,J), I = 1, NC)
C
C        INITIALIZE GRAND STATISTICS FOR THE ENTIRE DATA SET
C
         DO 30 I = 1, NC
         GTOT(I) = 0.0
      30 GSS(I)  = 0.0
C
C        COMPUTE STATISTICS FOR EACH CLUSTER AND PRINT RESULTS.
C
         LAST = 0
         DFE = 0
         RECI = 0
         DO 90 ICL = 1, NCL
         FIRST = LAST + 1
         NECL = NUMBER(ICL)
         DFE = DFE + (NECL - 1)
         RECI = RECI + (1.0 / NECL)
         LAST = LAST + NECL
         DO 40 I = 1, NC
         CTOT(I) = 0.0
      40 CSS (I) = 0.0
         WRITE (6, 2000) TITLE
         WRITE (6, 2400) ICL, NECL
         WRITE (6, 2500)
         WRITE (6, 2600) (LABELC(I), I = 1, NC)
         DO 70 J = FIRST, LAST
         JE = LIST(J)
         DO 50 I = 1,NC
         CTOT(I) = CTOT(I) + DATA(I, JE)
      50 CSS(I) = CSS(I) + DATA(I, JE) ** 2
      60 WRITE (6, FMTOUT) (LABELD(I,JE), I=1,2), JE, (DATA(I,JE), I=1,NC)
      70 CONTINUE
C
C        UPDATE GRAND STATISTICS AND PRINT CLUSTER STATISTICS
C
         DO 80 I = 1, NC
         GTOT(I) = GTOT(I) + CTOT(I)
         GSS(I)  = GSS(I) + CSS(I)
         CTOT(I)  = CTOT(I) / NECL
         MEAN (I, ICL) = CTOT(I)
      80 CSS(I)  = CSS(I)/NECL - CTOT(I) **2
         WRITE (6, 2700) (CTOT(I), I = 1,NC)
         WRITE (6, 2800) (CSS (I), I = 1,NC)
         WRITE (6, 2600) (LABELC(I), I = 1, NC)
      90 CONTINUE
C
```

```
C       PRINT GRAND STATISTICS
C
        WRITE (6, 2000) TITLE
        WRITE (6, 2900)
        WRITE (6, 2600) (LABELC(I), I = 1, NC)
        DO 100 I = 1,NC
        GTOT(I) = GTOT(I) / NE
   100  GSS (I) = GSS(I) / NE - GTOT(I) ** 2
        WRITE (6, 2700) (GTOT(I), I = 1,NC)
        WRITE (6, 2800) (GSS (I), I = 1,NC)
C
C       LEAST SIGNIFICANT DIFFERENCE TEST FOR THE MEANS OF CLUSTERS
C
        IF (ISIG.LE.0) ISIG = 3
        IF (ISIG.GE.3) ISIG = 3
        GO TO (102, 102, 5), ISIG
   102  CONTINUE
        IF (IPRO.LE.1) PROB = 0.05
        IF (IPRO.GE.2) PROB = 0.01
        HM = NCL / RECI
        DO 110 I = 1, NC
        SEOD = SQRT(GSS(I) * 2.0 / HM)
        DO 105 J = 1, NCL
        A(J) = MEAN(I,J)
   105  CONTINUE
        KARHDG(1) = LABELC(I)
        KARHDG(2) = '      '
        KARHDG(3) = '      '
        WRITE (6, 1300) KARHDG
        CALL SMPRIN (A, NCL, 1, 'CLUST.', 160, 10)
        CALL DIFFS (A,   NCL, DFE, SEOD, PROB, I, KARHDG, 1,
      1              'CLUSTER             ', ISIG)
   110  CONTINUE
        GO TO 5
   200  CONTINUE
        WRITE (6 , 3000)
C
  1000  FORMAT (16A5)
  1100  FORMAT (8I5)
  1200  FORMAT (16I5)
  1300  FORMAT (1H1, 9X, 35(1H*), 10X, 'MEAN VECTOR FOR ', 3A6,
      1              35(1H*)/55X, 22(1H=)//)
  2000  FORMAT (1H1, 4X, 10(1H*), 5X, 16A5, 5X, 10(1H*)//)
  2010  FORMAT (1H1, /20X, 32(1H$), 16H  PROGRAM POSTCA  , 32(1H$),/
      1              20X, 1H$, 80X, 1H$/20X, 1H$, 80X, 1H$/20X, 1H$, 16A5, 1H$/
      2              20X, 1H$, 80X, 1H$/20X, 1H$, 80X, 1H$/20X, 82(1H$)///)
  2100  FORMAT (10X, 7HNE    = , I6/10X, 7HNC    = , I6/10X, 7HNCL   = , I6/
      1              10X, 7HNTIN  = , I6/10X, 7HISIG  = , I6/10X, 7HIPRO  = , I6/)
  2200  FORMAT (/ 9X, 21H SIZE OF EACH CLUSTER ,/, (10X, 2I10))
  2310  FORMAT (/ 9X,   15HOUTPUT FORMAT : , 20A5//)
  2300  FORMAT (/10X, 15HINPUT FORMAT : , 20A5//)
  2400  FORMAT (/10X, 8H CLUSTER, I3, 11H CONTAINING, I4, 12H DATA UNITS.)
  2500  FORMAT (/11H DATA UNITS, 2X, 2HID, 3X, 20H SCORES ON VARIABLES)
  2600  FORMAT (/15X, 11(5X,A5))
  2700  FORMAT (/6H MEANS, 11X, 1P11E10.3)
  2800  FORMAT (/10H VARIANCES, 7X, 1P11E10.3)
  2900  FORMAT (/10X, 31H STATISTICS FOR ENTIRE DATA SET )
  3000  FORMAT (1H1, //42X, 37(1H*)/ 42X, 37(1H*) ///42X,
      1              37HPROGRAM POSTCA BY S.H. TEOW JUNE 1977    ///
      2              42X, 37(1H*)/ 42X, 37(1H*))
        END
```

LIST SYMBOL/CONVER

DATE 01/17/78 TIME IS 14:38

SYSTEM/DUMPALL VERSION 2.9.170

LASTRECORD = 111
MAXRECSIZEIN = 14 BLOCKSIZEIN = 420
*** EBCDIC *** UNITS=WORDS

```
$ SET AUTOBIND
$BIND = FROM CODON/=
FILE 20 = NTMERG, UNIT = DISK
C
C
C      PROGRAM CONVERT
C
C
C      REFERENCE:  E.J. BURR (1970) THE AUSTRALIAN COMPUTER JOURNAL,
C                                   VOL. 2, NO. 3, 98-103
C
C
       REAL NTOT
       DIMENSION N(160), S(160), TITLE(16), FMTIN(16)
C
     2 READ (5,5) TITLE
     5 FORMAT (16A5)
       WRITE (6,7) TITLE
     7 FORMAT (1H1, /20X, 32(1H$), 16H  PROGRAM CONVER  , 32(1H$),/
     1          20X, 1H$, 80X, 1H$/20X, 1H$, 80X, 1H$/20X, 1H$, 16A5, 1H$/
     2          20X, 1H$, 80X, 1H$/20X, 1H$, 80X, 1H$/20X, 82(1H$)///)
       IF (TITLE(1).IS.TITLE(2)) GO TO 960
       READ (5,5) FMTIN
       READ (5, 10) METHOD, NE, NTMERG, TSS
    10 FORMAT (3I5, F20.10)
       WRITE (6,12) METHOD, NE, NTMERG, TSS
    12 FORMAT (//5X,'METHOD   = ',I5/5X, 'NE       = ', I5/5X, 'NTMERG   = '
     1      ,I5/5X, 'TSS      = ', F20.10//)
       WRITE (6,15)
    15 FORMAT (1H1/ 26X, ' AVERAGE DIST. ', ' DIST. BETWEEN ',
     1      '   VARIANCE OF  ', 'INCREASE OF WSS', 15X, '   F-RATIO  ',
     2      /'   K   II N(II)  JJ N(JJ)', ' BETWEEN II, JJ' ,
     4      '   CENTROIDS  ', ' NEW CLUSTER  ', ' DUE TO MERGE  ',
     5      '   POOLED WSS  ', '  (AMS/WMS)   ', 'PROBABILITY ')
C
       A = 0.0
       SSI = 0.0
       ASS =0.0
       XWSS = 0.0
       DO 20 I = 1, NE
       N(I) = 1
       S(I) = 0.0
    20 CONTINUE
C
       INQUIRE (NTMERG, KIND = KND)
       ICON = 0
       ICOE = 1
       IF (KND.NE.9) ICON = 6
       IF (KND.NE.9) ICOE = 2
       NSTG = NE - 1
```

```
      DO 50 I = 1,NSTG
      IREC = (ICODE * I) + ICON
      READ (NTMERG = IREC, FMTIN) K, II, JJ, CR
      IF(I.NE.K) GO TO 900
      NTOT = N(II) + N(JJ)
      GO TO (100, 200, 300, 400, 500, 600, 700, 800), METHOD
C
  100 CENT = CR
      A = CENT + S(II)/N(II) + S(JJ)/N(JJ)
      SSI = (N(II) * N(JJ) / NTOT) * CENT
      WSS = WSS + SSI
      GO TO 40
  500 A = CR
      CENT = A - S(II)/N(II) - S(JJ)/N(JJ)
      SSI = (N(II) * N(JJ) / NTOT) * CENT
      WSS = WSS + SSI
      GO TO 40
  700 WSS = CR
      SSI = WSS - XWSS
      CENT = (SSI * NTOT) / (N(II) * N(JJ))
      A = CENT +  S(II)/ N(II) + S(JJ)/ N(JJ)
      XWSS = WSS
      GO TO 40
C
   40 CONTINUE
      WMS = WSS / K
      ASS = TSS - WSS
      DFA = NSTG - K
      IF (DFA.LE.0) DFA = 1
      AMS = ASS/ DFA
      F = AMS / WMS
      PF = PRBF(DFA, K, F)
      S(II) = S(II) + S(JJ) + SSI
      VAR =  S(II)/ (NTOT - 1)
      WRITE (6, 35) K, II, N(II), JJ, N(JJ), A, CENT, VAR, SSI, WSS, F,
     1              PF
   35 FORMAT (/5I5, 5E15.7, 2E12.4)
      N(II) = NTOT
   50 CONTINUE
C
      GO TO 950
  200 CONTINUE
  400 CONTINUE
  300 CONTINUE
  600 CONTINUE
  800 CONTINUE
      WRITE (6, 1000)
 1000 FORMAT (10X, 'NO CONVERTION AVAILABLE YET ')
C
      GO TO 950
  900 WRITE (6, 2000)
 2000 FORMAT (10X, ' ERROR IN MERGE SEQUENCE')
C
  950 CONTINUE
      GO TO 2
C
  960 CONTINUE
      WRITE (6, 980)
  980 FORMAT (    ////41X, 39(1H*)/41X, 39(1H*)///41X,
     1           39HPROGRAM CONVER BY S.H. TEOW AUGUST 1977   ///
     2           41X, 39(1H*)/41X, 39(1H*))
```

0011200

END

112

LIST NTIN

DATE 01/17/78 TIME IS 14:38

SYSTEM/DUMPALL VERSION 2.9.170

MAXRECSIZEIN = 14 BLOCKSIZEIN = 14
*** EBCDIC *** UNITS=WORDS

1	NO.	C.DIA	C.DEN	C.ERE	C.HEI	RUST	U.DIS	L.ROL	L.COL	L.WID	F.COL	F.DAT
2	001	6.3333	6.1667	4.5417	4.3750	3.5000	2.7083	1.1667	1.2917	2.7500	2.7500	5.125
3	002	6.1739	6.3913	3.9130	3.7826	3.5217	2.9130	1.3043	1.5217	2.8261	2.6957	5.695
4	003	6.7500	6.8750	4.1667	4.7083	3.2917	2.5000	1.4583	1.5417	2.6667	2.6667	5.125
5	004	6.6667	7.0000	5.1667	4.6667	3.6250	3.0833	1.3333	1.2500	3.0000	2.6667	5.125
6	005	6.5417	6.3333	3.9583	4.1667	3.5833	2.6250	1.3333	1.4167	2.7917	2.5833	5.416
7	006	6.5417	6.1250	4.1667	3.5833	3.3750	3.1667	1.5833	1.5000	2.7500	2.5417	4.958
8	007	6.7917	6.5417	4.1250	4.0417	3.5000	2.4583	1.3333	1.2500	3.1250	3.0000	5.250
9	008	7.4583	6.9583	5.4583	4.6250	3.6667	2.6250	1.3750	1.1667	3.1250	2.5833	4.666
10	009	6.1739	5.8261	4.2609	3.6957	3.6087	2.5217	1.2609	1.3043	2.9130	2.6087	5.173
11	010	5.8750	6.1250	7.8750	5.1250	4.0000	2.9583	1.0417	1.2083	2.9167	2.5833	4.958
12	011	5.8182	5.7273	4.4091	3.5909	3.8182	3.1364	1.1818	1.4545	2.7273	2.7273	5.318
13	012	5.8333	6.7917	6.3333	5.2500	3.7083	3.3750	1.2917	1.2083	2.8333	2.2083	4.416
14	013	7.0417	7.0833	5.2500	5.0417	3.4583	3.0000	1.2500	1.3333	2.9167	2.7500	4.875
15	014	6.5417	5.9583	3.6667	3.6250	3.9583	3.1667	1.5000	1.5417	2.7083	2.7500	4.916
16	015	6.6667	7.0417	4.2917	3.8750	3.2917	2.8333	1.2083	1.5000	2.8750	3.3750	5.708
17	16	6.0000	6.6250	4.0417	4.1667	3.6250	3.2083	1.4167	1.3333	2.9167	2.5417	5.250
18	17	5.7917	6.0000	5.0417	5.0833	3.2083	2.7083	1.2083	1.4583	3.0000	2.3750	5.083
19	18	7.2917	6.6250	4.5417	4.2917	3.4167	2.5000	1.4167	1.2917	2.7083	2.8333	5.458
20	19	6.7917	6.8333	4.9583	4.5417	3.6250	2.5000	1.3333	1.1250	3.0417	2.6667	5.291
21	20	6.6667	6.6667	4.8333	4.5000	3.6667	2.4167	1.6250	1.2083	2.9583	2.8750	5.666
22	21	6.6667	6.2083	4.1667	3.6667	3.7500	3.3750	1.5000	1.4167	2.9583	3.2917	5.333
23	22	7.1739	6.1739	6.0000	4.9130	3.4348	3.0000	1.1739	1.3913	3.0870	2.7826	5.260
24	23	6.2609	6.5217	5.2609	5.2609	3.9130	2.9565	1.4348	1.4783	2.8261	3.0870	5.782
25	24	6.0909	5.8636	4.4091	3.5000	3.7273	2.8636	1.3182	1.5909	2.6818	2.9545	5.454
26	25	7.0870	6.5652	4.5217	3.6522	3.8261	3.2174	1.2174	1.2609	3.0000	2.6957	5.087
27	26	6.2917	6.8333	4.6667	3.6250	3.3750	2.6667	1.5833	1.4583	2.8333	2.8750	5.791
28	27	6.5417	5.7500	4.4583	4.9583	3.2500	2.4583	1.2500	1.3333	2.8333	2.4167	5.458
29	28	6.6818	6.9091	4.2727	3.3636	3.9545	3.1364	1.7273	1.2727	2.8636	2.9545	5.681
30	29	6.4583	6.4583	5.0000	4.4583	3.4167	2.7083	1.2500	1.2500	2.6250	2.8750	5.416
31	30	6.0417	6.0000	5.6250	4.0833	3.7917	2.5833	1.2500	1.5000	3.1250	2.6250	5.125
32	31	6.6957	6.7391	4.4783	3.7391	4.2174	2.7391	1.4783	1.3913	2.9565	2.3913	5.043
33	32	6.6957	6.9565	5.9565	5.2174	3.2174	3.1739	1.3913	1.1739	2.6522	2.3913	5.000
34	33	6.0417	6.1250	4.0000	4.1250	3.7500	2.9167	1.2500	1.5000	3.1250	2.5000	5.211
35	34	6.4583	6.2917	4.5000	4.1250	3.8750	2.8750	1.0833	1.3333	3.0000	2.5000	4.708
36	35	5.8750	6.6667	4.3750	4.0833	4.3750	3.2083	1.4167	1.4167	2.7917	3.1250	5.333
37	36	6.4583	6.5833	3.4167	3.7083	3.7083	2.8333	1.6250	1.4167	2.8750	2.7917	5.416
38	37	6.8150	6.2917	3.6250	4.0417	3.2500	3.0000	1.4583	1.7083	2.9167	3.0417	5.875
39	38	6.0833	6.2917	4.6667	4.2500	3.8333	2.8750	1.5417	1.0833	2.7083	2.5417	5.625
40	39	6.7826	6.1304	4.7826	4.0870	3.7826	2.6957	1.2174	1.1304	3.0435	2.7391	5.087
41	40	5.8696	6.5652	5.6087	4.3043	3.2609	2.4348	1.2609	1.4348	2.6957	2.3478	5.260
42	41	7.1250	6.2917	3.7083	3.7917	3.7083	2.9583	1.2917	1.2917	2.8750	3.2083	4.875
43	42	7.0000	7.0417	4.2083	3.8750	4.0000	2.8750	1.3333	1.2083	3.1250	2.6250	5.000
44	43	6.5833	6.5833	5.0417	4.2917	3.3750	2.5000	1.2917	1.2917	3.0000	2.2917	5.208
45	44	6.3750	6.2917	4.0000	3.3750	3.7500	3.3333	1.1250	1.3750	2.9167	2.7083	5.291
46	45	6.8261	6.7391	4.3913	3.6261	3.8261	3.3043	1.3913	1.2609	3.1304	2.6957	5.087
47	46	5.8750	5.8333	4.2083	4.4583	3.3333	2.5000	1.2917	1.2917	2.8333	2.8750	5.875
48	47	6.6250	6.6250	4.0000	4.0833	3.3750	2.7917	1.3333	1.5000	2.7917	2.6667	5.791
49	48	6.1250	6.5000	4.2500	4.2917	3.3333	2.8750	1.2500	1.2083	2.6250	2.6250	5.583
50	49	6.5417	5.2917	4.8750	4.0833	4.0000	3.0000	1.2917	1.4167	3.5000	3.0417	5.625
51	NO.	C.DIA	C.DEN	C.ERE	C.HEI	RUST	U.DIS	L.ROL	L.COL	L.WID	F.COL	F.DAT

NO.	C.DIA	C.DEN	C.ERE	C.HEI	RUST	U.DIS	L.ROL	L.COL	L.WID	F.COL	F.DAT
50	6.7500	5.9167	3.9167	3.5833	3.6250	3.1250	1.2083	1.2083	3.0000	2.7917	5.208
51	7.1250	6.5833	4.1250	3.2917	3.8750	3.3333	1.6250	1.6250	2.9583	2.8333	6.000
52	6.6957	6.6087	3.7826	3.4348	2.8261	2.4348	1.2609	1.6087	2.8261	2.8261	5.260
53	5.8333	6.4167	5.6667	4.1250	4.1667	3.2500	1.4167	1.3750	2.6667	2.7500	5.291
54	5.3043	5.9130	5.0435	3.8696	4.1739	2.7826	1.4783	1.3913	2.4783	3.0000	5.173
55	6.1250	6.2500	4.7917	4.3750	3.5833	2.6250	1.1667	1.1250	2.7500	2.4167	4.750
56	6.4583	5.5833	4.6667	4.3750	3.2083	2.6250	1.1667	1.9583	2.5417	2.7083	4.583
57	6.7391	6.5652	3.7391	3.5217	3.4783	2.8696	1.6522	1.2609	2.7391	2.3913	5.260
58	6.3750	6.5000	5.5000	4.5833	2.7500	2.6250	1.2500	1.3750	3.1250	2.3333	5.416
59	7.0000	6.4583	4.3333	3.6333	3.8333	2.7917	1.3750	1.3333	3.0417	2.3750	4.916
60	7.0417	6.2083	5.1667	4.7083	2.7917	2.2500	1.2500	1.4167	3.0417	2.6250	5.041
61	6.8261	6.6522	6.1304	4.9130	3.6957	2.8696	1.4783	1.3043	3.0000	2.6087	5.434
62	6.3750	5.9167	3.8750	4.4167	4.0417	3.1667	1.4167	1.3750	2.6667	2.8750	5.125
63	7.0000	7.2917	6.5417	4.3750	3.6667	3.5000	1.1667	1.1250	2.6250	2.9167	5.500
64	5.8333	6.2083	3.3750	3.4583	3.0833	2.7083	1.6250	1.5417	2.7083	2.9583	5.708
65	6.4583	7.3333	4.6250	4.2500	3.5833	2.7500	1.4167	1.5000	2.7917	2.7917	5.750
66	6.7083	6.7917	4.5833	4.2917	3.7917	2.5417	1.3333	1.0833	3.0000	2.5833	5.375
67	6.2500	6.9167	4.8333	3.6667	3.5000	2.8750	1.3333	1.2500	2.5000	2.9167	5.458
68	7.1667	7.1250	7.3750	5.2917	4.2083	3.2500	1.0417	1.0833	3.5417	2.9583	5.166
69	6.9565	6.3043	4.3478	4.1739	3.2174	2.1739	1.3478	1.2174	3.1739	2.6087	5.391
70	6.4167	6.6250	4.3333	5.1250	3.5417	2.3750	1.2917	1.2500	2.9167	2.7083	5.750
71	6.0417	6.0417	4.3750	4.0417	3.6667	2.9583	1.2083	1.2083	2.6667	2.6667	5.458
72	7.1739	6.5217	4.8261	5.0870	4.0000	3.3043	1.2609	1.2174	3.0870	2.3478	5.391
73	6.1250	6.3333	4.6250	3.7917	4.1250	3.5417	1.1667	1.3750	3.0000	2.6667	5.875
74	6.7083	7.0000	4.5833	4.4583	3.5000	3.1667	1.4583	1.5000	2.6250	2.7500	5.458
75	5.9583	6.8333	3.7500	3.9167	3.4583	2.8750	1.4583	1.1667	2.5417	2.7917	5.625
76	6.4167	6.3750	4.2917	3.9583	4.1250	2.7917	1.3333	1.2083	2.7083	2.7500	5.583
77	6.6250	6.2083	4.2500	4.0833	4.1250	3.0000	1.4167	1.2917	2.6250	2.6250	5.791
78	6.5000	6.4167	4.2917	4.1250	3.3333	2.8333	1.2500	1.3750	2.6250	2.5000	5.125
79	4.7083	6.1667	5.0000	3.6250	3.8750	3.0417	1.1667	1.3333	2.5000	2.5000	4.291
80	6.4583	6.4583	4.5833	5.0833	3.5000	2.3333	1.1667	1.5833	3.0833	2.8750	5.375
81	6.2500	6.7083	5.7083	4.9583	3.5000	2.8750	1.2500	1.3750	2.9167	2.8750	5.791
82	6.4583	6.7500	5.0417	4.2083	3.8750	3.1250	1.2500	1.1250	2.8333	2.7500	5.583
83	6.5000	7.0000	3.9583	4.2500	3.7500	2.7500	1.3333	1.3750	2.5833	2.5417	5.791
84	6.4167	6.5417	4.2917	4.3750	3.1667	2.5833	1.2500	1.2083	2.8333	2.7500	5.666
85	6.2500	6.4583	3.1667	3.0000	3.5417	2.9583	1.3750	1.2500	3.1667	3.4583	5.791
86	7.1667	7.1250	3.9167	3.6667	3.7500	2.6667	1.4167	1.4167	2.7917	3.1667	6.416
87	5.7500	6.3333	4.3333	4.0833	3.7500	2.6667	1.4167	1.2083	2.8750	3.2917	6.583
88	6.1304	6.9565	4.8696	4.0435	3.7391	2.6087	1.4783	1.4348	2.7826	2.9565	6.130
89	5.5217	6.0000	3.9130	3.3913	4.2174	3.1304	1.3043	1.3043	2.8696	3.0435	5.782
90	6.2174	7.1739	4.1739	3.0870	3.7391	2.8261	1.3478	1.4783	2.9130	2.8696	5.739
91	6.2093	6.2917	6.2917	5.1667	3.5000	2.7083	1.2083	1.1250	2.7917	2.2083	4.958
92	6.8750	6.9583	4.1667	3.4583	3.5833	2.6250	1.3750	1.3333	2.5000	2.9583	5.875
93	5.7083	5.7500	4.5000	3.7083	3.5833	3.1667	1.3750	1.2083	2.7917	2.7917	4.750
94	6.7083	6.8333	4.6667	4.4167	3.7083	2.5000	1.6250	1.2500	2.6667	2.7500	5.291
95	6.2083	6.2500	4.9167	4.4167	3.4167	3.0000	1.5000	1.4583	2.4167	2.5833	5.250
96	6.1250	6.1667	5.0417	4.0833	3.6250	2.8750	1.3750	1.3750	2.3750	2.6250	5.125
97	6.2917	6.8333	3.6667	3.5833	3.5417	2.7917	1.5417	1.5833	2.6667	3.1250	5.458
98	6.1739	6.9565	2.6957	3.0435	3.9130	3.2609	1.6957	1.4783	2.5652	2.8261	5.434
99	6.0000	6.7917	4.2500	3.5833	4.4583	3.1667	1.5000	1.1667	2.9167	3.0417	5.958
100	6.4583	6.3750	6.2500	4.7083	3.7500	2.4167	1.1667	1.1667	3.1250	2.9167	5.041
101	6.8750	6.2500	3.0000	3.9167	3.6250	3.2917	1.5000	1.2083	2.7500	2.8750	5.833
102	6.1250	6.9167	3.7917	3.7500	3.0000	2.1667	1.4167	1.0833	2.6667	2.8750	6.500
103	5.9583	6.7500	5.6250	4.7500	3.2917	2.6250	1.2917	1.2917	3.0417	2.5417	5.750
104	6.7500	6.7917	4.7083	4.6250	2.9583	2.6250	1.2083	1.2917	2.8333	2.8750	5.833
105	6.2083	6.9167	5.8333	4.3750	3.6250	2.9167	1.2083	1.2500	3.2083	3.1667	5.625
106	6.8750	7.3750	4.5000	3.7917	3.8750	3.0417	1.5000	1.1667	2.7917	2.7500	5.791
107	6.7500	6.2500	3.2917	3.7500	3.8750	3.4583	1.5000	1.3750	2.9583	2.9167	5.416
108	5.9583	6.7500	3.8333	3.7917	3.4167	2.2083	1.3750	1.3750	2.5833	2.8750	5.958
109	6.0833	6.2917	3.9583	4.2500	3.6250	2.8750	1.2917	1.2500	2.7917	2.6667	5.833

110	6.0000	6.3333	5.0833	3.9583	3.8333	2.6250	1.2917	1.1667	3.0000	2.7083	5.583
111	7.0870	6.4348	4.5652	4.0870	3.0435	2.7391	1.3478	1.3043	2.9130	2.4783	5.217
112	6.9130	6.3043	4.3043	4.1304	4.3478	3.4348	1.3478	1.0343	3.0000	2.8696	5.608
113	6.5833	6.8750	5.4583	4.6250	3.4583	3.0417	1.3333	1.4167	2.6250	2.6250	5.500
114	6.2917	6.3750	4.3750	3.5417	3.9583	2.9583	1.2083	1.2500	3.1250	2.9583	5.833
115	6.1667	6.2500	3.2917	4.0000	3.3750	2.8333	1.6250	1.2917	2.7917	3.1250	5.916
116	6.3333	6.4583	5.0000	3.2500	3.7500	3.0000	1.5833	1.0833	2.8750	2.8750	5.625
117	6.7917	6.5833	3.2083	4.2500	3.6667	2.8750	1.2917	1.2917	2.5833	2.9167	6.250
118	6.5217	6.2609	3.9565	3.2609	3.7391	2.8696	1.3043	1.3913	2.9130	2.7826	5.521
119	6.1667	6.3750	3.5000	3.6750	3.6250	2.9583	1.5833	1.2083	2.7917	3.2083	5.875
120	7.0417	6.2917	4.5000	4.1667	3.7083	2.8333	1.5417	1.4167	2.8750	2.6667	5.208
121	6.2609	6.5652	3.1739	3.4783	3.7391	2.8261	1.5652	1.6522	2.4783	2.9130	5.652
122	6.7917	6.2083	5.1250	4.6667	3.8333	2.6667	1.3333	1.4167	2.8333	2.6667	5.208
123	6.5000	6.4167	4.4583	4.5417	3.3333	2.8750	1.4167	1.3333	2.7917	2.8750	5.708
124	6.1250	6.3750	4.3750	4.0833	3.3333	2.4167	1.2083	1.6250	2.7917	3.4167	5.666
125	6.5833	6.3333	4.6250	4.0833	3.6667	2.8750	1.3333	1.5000	3.1250	3.0000	5.416
126	6.1250	6.5417	3.7083	3.6750	3.6250	2.6250	1.3750	1.2500	2.8333	2.9167	5.458
127	6.7083	6.1667	3.4167	3.1250	4.0833	3.3750	1.6250	1.4583	2.6667	3.2083	5.791
128	7.0417	6.6250	4.9583	4.3333	3.5417	2.9167	1.3750	1.0833	3.1250	2.8750	5.833
129	6.5000	7.0833	4.6250	3.7083	3.5000	3.0000	1.3750	1.3750	2.5417	2.4583	5.750
130	6.9167	6.5000	3.5833	4.5417	3.9583	3.3333	1.4583	1.3333	2.8333	3.0000	6.041
131	6.0435	6.7391	3.3478	3.3478	3.7826	3.0435	1.7826	1.2174	2.5217	3.4348	5.782
132	6.3478	6.6522	4.1739	3.9565	3.2609	2.9565	1.4348	1.2174	2.5652	3.1304	6.130
133	7.3333	6.8333	3.8333	3.7917	3.8333	2.9167	1.6667	1.4167	3.0833	2.6250	5.375
134	7.1304	5.9565	5.5652	4.0435	4.0870	2.6522	1.1304	1.1304	3.2609	2.5217	3.956
135	5.9167	6.5833	4.5417	4.0833	3.7917	2.8333	1.2917	1.2917	2.9167	2.6250	5.625
136	6.4167	6.1250	4.3333	4.4167	3.5000	2.8333	1.4167	1.4583	2.8333	2.5833	5.500
137	6.6250	6.5833	5.5417	4.5833	3.7500	2.9583	1.2917	1.1250	2.5000	2.5833	4.541
138	5.6957	5.6957	5.4783	3.4783	3.8261	2.8696	1.1304	1.2609	3.1739	2.5217	3.695
139	6.5417	6.1250	4.3750	4.4167	3.2500	2.5000	1.2917	1.2500	2.8333	3.1250	5.625
140	6.4583	6.5417	4.3333	4.0833	3.3750	2.7083	1.4167	1.1667	2.9167	3.0833	5.625
141	6.2917	6.3333	5.0833	4.5000	3.4583	2.7083	1.3750	1.3333	2.5833	2.6250	5.416
142	6.0417	5.9583	5.2500	3.9583	3.7500	2.5417	1.2083	1.2083	3.2917	2.5833	3.916
143	6.7500	6.7083	5.0417	3.6333	3.7917	3.0000	1.3750	1.1667	3.2083	2.8333	5.375
144	6.0417	7.2917	7.4583	6.2500	3.5833	2.6667	1.0833	1.4167	2.8333	2.2500	4.791
145	7.2500	6.4583	4.9167	4.5417	3.4583	2.6667	1.4167	1.1667	2.9583	2.5833	5.458
146	6.4167	6.2083	4.2500	3.9583	3.2500	2.4167	1.3750	1.2500	2.8750	2.9167	5.666
147	6.0909	6.1364	4.5455	4.0818	3.6818	2.6091	1.3636	1.4091	2.8182	3.0455	5.954
148	6.3750	6.8750	5.2917	4.1250	3.3333	2.6667	1.2917	1.3750	2.5833	2.7083	5.125
149	6.9583	6.5417	4.6667	4.5000	3.5417	2.8333	1.3333	1.4167	2.8750	2.7917	4.875
150	6.2083	6.4583	4.5833	4.2500	3.7500	2.9583	1.2917	1.2500	2.6667	2.7917	4.125
151	6.3750	6.6667	6.0417	4.0833	3.8750	2.8750	1.2083	1.0833	2.7917	1.3333	4.458
152	5.1667	6.2500	4.0833	3.6333	3.3750	2.6250	1.2917	1.4583	2.9167	1.7917	5.833
153	6.4783	6.5652	5.3043	3.4783	3.3913	2.6522	1.3043	1.1739	3.0000	2.3478	5.478
154	6.2917	6.2083	6.2083	5.2500	3.5000	2.5000	1.2917	1.4167	2.8750	2.5417	4.833
155	6.5833	7.0000	5.9167	5.3750	3.2083	2.6667	1.3333	1.3333	2.9167	2.8333	5.208
156	6.7917	7.2500	5.7500	5.3333	3.4167	2.3333	1.5833	1.2500	2.8333	2.5833	5.041
157	6.8750	6.6250	4.8750	4.0000	3.2500	2.5000	1.1667	1.2917	2.8750	2.4167	5.083
158	6.5417	6.4583	6.8750	4.1667	2.9167	2.2083	1.1250	1.1250	2.9583	2.4583	4.083
159	5.4583	6.2917	7.0417	5.1667	2.8750	2.6667	1.2917	1.2917	3.0417	2.4167	5.166
160	7.1250	6.5000	4.2917	3.9167	3.7500	2.3333	1.4583	1.4583	2.5000	2.7917	5.541
NO.	C.DIA	C.DEN	C.ERE	C.HEI	RUST	U.DIS	L.ROL	L.COL	L.WID	F.COL	F.DAT

CHARACTER	SCORE											MEAN	S.D.
	0	1	2	3	4	5	6	7	8	9	10		
C.DIA		21	28	63	278	462	839	1337	594	132	49	6.45	1.45
C.DEN		24	20	46	139	438	1242	1093	603	183	15	6.49	1.34
C.ERE		198	371	647	620	660	556	474	168	74	35	4.65	2.00
C.HEI		89	433	736	1236	622	414	137	87	30	19	4.16	1.58
RUST	45	674	1362	1287	359	76						2.39	0.99
O.DIS	4	100	715	1749	899	336						3.17	0.93
L.ROL	2508	1249	46									0.35	0.50
L.COL	2700	973	125	5								0.33	0.54
L.WID		25	1082	2156	518	22						2.85	0.72
F.COL		85	1319	1886	455	58						2.76	0.75
F.DAT		14	27	67	516	1360	1460	300	52	7		5.38	1.02

APPENDIX C-2 The Frequency Table And Summary Of All Characters For 3803 Plants.

There Are 37 Missing Plants Not Included.

The S.D. (Standard Deviation) Are Based On The Total Mean Sum Of Squares Over The Whole Population.

BIBLIOGRAPHY

ALLARD, R.W.(1960)
Principles of plant breeding. John Willy and Sons Inc., New York. 485 p.p.

ALLARD, R.W.(1970)
Problems of maintenance. Chapter 41, pg. 491-494 of Genetic resources in plants, edited by Frankel, O.H. and Bennett, E., Blackwell scienctific publication, Oxford. 554 p.p.

ALLARD, R.W. and JAIN, S.K.(1962)
Population studies in predominantly self-pollinated species. II. Analysis of quantitative changes in a bulk hybrid population of barley. Evolution 16:90-101.

ANDERBERG, M.R.(1973)
Cluster analysis for applications. Academic Press, New York. 559 p.p.

ANDERSON, T.W.(1958)
An introduction to multivariate statistical analysis. John Wiley and Sons Inc., New York. 374 p.p.

ANDERSON, T.W.(1963)
A test for equality of means when covariance matrice are unequal. Annals of Mathematical Statistics 34:671-672.

ANDREWS, D.F., GNARADESIKAN, R., & WARNER, J.L.(1971)
Transformations of multivariate data. Biometric 27:825-840.

ANONY(1968)
IBM application program GH20-0205-4. System/360 scientific sub-routine package. Version III. 454 p.p.

ANONY(1975)
Burroughs advanced statistical inquiry system (BASIS) user's mannual.

BAARS, J.A.(1976)
Seasonal distribution of pasture production in New Zealand. VIII Dargaville. New Zealand Journal of Experimental Agriculture 4:151-156.

BAARS, J.A.(1976)
Seasonal distribution of pasture production in New Zealand. IX Hamilton. New Zealand Journal of Experimental Agriculture 4:157-161.

BAARS, J.A., RADCLIFFE, J.E. and BRUNSWICK, L.(1975)
Seasonal distribution of pasture production in ew Zealand. VI Wairakei, pasture and lucerne production. New Zealand Journal of Experimental Agriculture 3:253-258.

BARTLETT, M.S.(1938)
Further aspect of multiple regression. Proceedings of the Cambridge Philosophic Society A 160:268-282.

BARTLETT, M.S.(1947)
The use of transformations. Biometrics 3:39-53.

BASYNAT, M.B.(1957)
Thesis, Massey University.

BENNET, E.(1970)
Adaptation in wild and cultivated plant populations. Chapter 10, pg. 115-130 of Genetic Resources in Plants, edited by Frankel, O.H. and Bennett E., Blackwell Scientific Publication, Oxford 554 p.p.

BHATT, G.M.(1970)
Multivariate analysis approach to selection of parents for hybridization aiming at yield imporvement in self-pollinated crops. Aust. J. Agric. Res. 21:1-7.

BHATT, G.M.(1976)
An application of multivariate analysis to selection for quality characters in wheat. Aust. J. Agric. Res., 27:11-18.

BOX, G.E.P. and COX, D.R.(1964)
An analysis of transformations. Journal Royal Statistical Society, B 26:211-243 (discussion 244-252).

BOYCE, A.J.(1969)
Mapping diversity: A comparative study of some numerical methods. Numerical Taxonomy, edited by A.J. Cole, 324 p.p. Academic Press, London. pg. 1-31.

BRYNANT, E.H. and ATCHLEY, W.R.(1975)
Multivariate statistical methods: Whithin-groups covariation. Dowden, Hutchinson and Ross Inc. Pennsylvania, p.p. 436.

BURR, E.J.(1968)
Cluster sorting with mixed character types. I. Standardization of character values. The Australian Computer Journal 1:97-99.

BURR, E.J.(1970)
Cluster sorting with mixed character types. II. Fusion strategies. The Australian Computer Journal 2(3):98-103.

BURT, R.L., EDYE, L.A., WILLIAMS, W.T., GROF, B. and NICHOLSON, C.H.L.(1971)
Numerical analysis of variation patterns in the genus stylosanthes as an aid to plant introduction and assessment. Aust. J. Agric. Res. 22:737-757.

CHEESEMAN, T.F.(1923)
Manual of the New Zealand Flora. Government Printer, Wellington. 2nd edition. 1163 p.p.

CLEGG, M.T., ALLARD, R.W. and KAHLER, A.L.(1972)
Is the gene the unit of selection? Evidence from two experimental plant populations. Proc. Nat. Acad. Sci. USA, 69:2474-2478.

CLIFFORD, H.T. and STEPHENSON, W.(1975)
An introduction to numerical classification. Academic Press, New York. 229 p.p.

COCHRAN, W.G.(1947)
Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics 3:22-38.

COOLEY, W.W. and LOHNES, P.R.(1971)
Multivariate data analysis. John Wiley & Sons, Inc. N.Y. 364 p.p. ISBN0-471-17060-7.

CORMACK, R.M.(1971)
A review of classification. Jour. of The Royal Statistical Society A:134:321-367.

CROVELLO, THEODORE J.(1970)
Analysis of character variation in ecology and systematics. Annual Rev. of Ecology & Systematics. 4:55-98.

DAVIES, E.B.(1944)
The lime requirement of soils. New Zealand J. Agric. 69:529-534.

DRAPER, N & SMITH, H.(1966)
Applied regression analysis. John Wiley & Sons, Ltd. New York., 407 p.p.

EDYE, L.A., BURT, R.L., WILLIAMS, W.T., WILLIAMS, R.J. & GROF, B.(1973)
A preliminary agronomic evaluation of stylosanthes species. Aust. J. Agric. Research 24:511-525.

EISENHART, C.(1947)
The assimptions underlying the analysis of variance. Biometrics 3:1-21.

ENGELMAN, L. and HARTIGAN, J.A.(1969)
Percentage points of a test for clusters. Journal of American Statist Ass. 64:1647-1648.

FISHER, R.A.(1936)
The use of multiple measurements in taxonomic problems. Ann. Eugenics 7:179-188.

FISHER, R.A.(1938)
The statistical use of multiple measurements. Ann. Eugenics 8:376-386.

FISHER, R.A.(1948)
Statistical methods for research workers. Oliver and Boyd, Edinburgh. 11th edition.

FRANKEL, O.H.(1947)
The theory of plant breeding for yield. Heredity 1:109-125.

FRANKEL, O.H.(1975)
Genetic resources centres - a cooperative global network. pg.473-481 Crop Genetic Resources for Today and Tomorrow, edited by Frankel, O.H. and Hawkes, J.G., Cambridge University Press, Cambridge, 492 p.p.

FRANKEL, O.H. and BENNETT, E.(1970)
Genetic resources - introduction. Chapter 1 of Genetic resources in Plants. pg:7-17.

FREUND, J.E.(1972)
Mathematical statistics. Prentice/Hall International, Inc. London 463 p.p.

GLENDAY, A.C. and FEJER, S.O.(1956)
The use of discriminant functions in the selection of pasture plants, with particular reference to the lolium species. Proc. 7th Int. Glassland Congress : Special Method in Plant Breeding: 461-470.

GOULD, S.J. and JOHNSTON, R.F.(1972)
Georgraphic variation. Annual Review of Ecology & Systemmatics. 3:457-498.

GOWER, J.C.(1966)
Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325-338.

HUBBARD, C.E.(1945)
Grasses. Pelican Books.

HUSSAINI, S.H., GOODMAN, M.M. and TIMOTHY, D.H.(19771)
Multivariate analysis and the geographical distribution of the world collection of Finger Millet. Crop Science 17. No. 2. pg. 257-263.

ITO, K.(1969)
On the effect of heteroscedasticity and non-normality upon some multivariate test procedures. Multivariate Analysis. II, edited by P.R.Krishnaiah, pg. 87-120. Academic Press, New York.

JACQUES, W.(1962)
Yorkshire fog as a pasture grass. Proceedings of The New Zealand Grassland Association. 24:139-150.

JACQUES, W.(1974)
Yorkshire fog (Holcus lanatus) : Its potential as a pasture species. Proceedings of The New Zealand Grassland Association. 35(2):249-257.

JAIN, S.K.(1961)
Studies on the breeding of self-pollinating cereals. The composite cross bulk population method. Euphytica 10:315-324.

KASHIRSAGAR, A.M.(1972)
Multivariate analysis. Marcel Dekker Inc. New York. 534 p.p.

LANCE, G.N. & WILLIAMS, W.T.(1966)
A generalized sorting strategy for computer classifications. Nature, 212:218.

LANCE, G.N. & WILLIAMS(1967a)
A general theory of classificatory sorting strategies. I. Hierarchical systems. Computer Journal 9:373-380.

LANCE, G.N. & WILLIAMS, W.T.(1967b)
Mixed-data classificatory programs. I. Agglomerative systems. Aust. Comput. J. 1:15-20.

LANCE, G.N. & WILLIAMS, W.T.(1968a)
Mixed-data classificatory programs. II. Divisive systems. Aust. Comput. J. 1:82-85.

LANCE, G.N. & WILLIAMS, W.T.(1968b)
Note on a new information statistic classificatory program. Computer J. 11:195.

LANCE, G.N. & WILLIAMS, W.T.(1971)
A note on a new divisive classificatory program for mixed data. Computer J. 14:154-155.

LEVY, E.V.(1970)
Grasslands of New Zealand.

MARRIOTL, F.H.C.(1971)
Practical problems in a method of cluster analysis. Biometrics 27:501-514

MARSHALL D.R. & BROWN, A.H.D.(1975)
Optimum smapling strategies in genetic conservation. Chapter 4, pg. 53-80 of Crop Genetic Resources for Today and Tomorrow. Edited by Frankel, O.H. and Hawkes, J.G., Cambridge University Press, Cambridge. 492 p.p.

MATHER, K.(1949)
Biometrical genetics - The study of continuous variation. Methuen & Co. Ltd., London. 1st edition. 162 p.p.

MATHER, K. & JINKS, J.L.(1971)
Biometrical genetics - The study of continuous variation. Chapman & Hall Ltd., London. 2nd edition. 382 p.p.

MAYER, LAWRENCE, S.(1971)
A method of cluster analysis when there exist multiple indicators of a theoretic concept. Biometrics 27:143-155.

MENGESHA, M.H.(1975)
Crop germplasm diversity and resources in Ethiopia. Chapter 33. pg. 449-456. Crop Genetic Resources for Today and Tomorrow. Edited by Frankel, O.H. and Hawkes, J.G. Cambridge University Press, Cambridge. 492 p.p.

MITCHELL, K.J.(1956)
The growth of pasture species under controlled environment. I. Growth at various levels of constant temperature. New Zealand J. Sci. Tech. A38:203-216.

MITCHELL, K.J. & LUCANUS, R.(1960)
Growth of pasture species in controlled environment. II. Growth at low temperatures. New Zealand J. Agric. Res. 3:647-655.

MUNGOMERY, V.E., R.SHORTER & D.E. BYTH(1974)
Genotype X environment interactions and environmental adaptation. I. Pattern analysis - Application to soya bean populations. Aust. J. Agric. Res., 25:59-72.

MUNRO, M.M.(1961)
Thesis, Massey university.

PERSON, E.S. & HARTLEY,H.O.(1972)
Biometrika tables for statisticians. Volume II. Cambridge University Press, Great Britain, 385. p.p.

PILLAI, K.C.S. & GUPTA, A.K.(1969)
On the exact distribution of Wilk's criterion. Biometrika 56:109

PRESS, S.J.(1967)
Structured multivariate behrens - Fisher problems. Sankhya Series A, 29(Part 1):41-48.

PRESS S.J.(1972)
Applied Multivariats Analysis. Holt, Rinehart & Winston, Inc. New York. 521 p.p.

PRICHARD, N.M. & ANDERSON, A.J.B.(1971)
Observations on the use of cluster analysis in botany with an ecological example. The Journal of Ecology, 59:727-747.

RADCLIFFE, J.E.(1974)
Seasonal distribution of pasture production in New Zealand. I Methods of measurement. New Zealand Journal of Experimental Agriculture. 2:337-340.

RADCLIFFE, J.E.(1974)
Seasonal distribution of pasture production in New Zealand. II. Southland plains. New Zealand Journal of Experimental Agriculture 2:341-348

RADCLIFFE J.E.(1975)
Seasonal distribution of pasture production in New Zealand. IV. Westport and Motueka. New Zealand Journal of Experimental Agriculture 3:239-246.

RADCLIFFE, J.E.(1975)
Seasonal distribution of pasture production in New Zealand. VII. Masterton (Wairarapa) and Maraekakaho (Hawke's Bay) New Zealand Journal of Experimental Agriculture 3:259-265.

RADCLIFFE, J.E.(1976)
Seasonal distribution of pasture production in New Zealand X. Rangitikei district. New Zealand Journal of Experimental Agricultre 4:163-170.

RADCLIFFE, J.E. & COSSENS, G.G.(1974)
Seasonal distribution of pasture production in New Zealand. III. Central Otago. New Zealand Journal of Experimental Agriculture 2:349-358.

RADCLIFFE, J.E. & SINCLAIR, D.P.(1975)
Seasonal distribution of pasture production in New Zealand. V. Gisborne Plains. New Zealand Journal of Experimental Agriculture 3:247-251.

ROO, C.R.(1952)
Advanced statistical methods in biometric research. John Wiley & Sons, New York.

RATKOWSKY, D.A.(1976)
Personal comminication.

RICKARD, D.S. & RADCLIFFE(1976)
Seasonal distribution of pasture production in New Zealand. XII. Winchmore, Canterbury Plains dryland and irrigated pastures. New Zealand Journal of Experimental Agriculture 4:329-335.

RIVEROS, F.(1963)
Thesis, Massey University.

ROHLF, F.J.(1971)
Perspectives on the application of multivariate statistics to taxonomy. Taxon 20(1):85-90

ROUND-TURNER, N.L., SCOTT, R.S., & RADCLIFFE, J.E.(1976)
Seasonal distribution of pasture production in New Zealand. XI Otago downland and Taieri plain (Invermay). New Zealand J. of Experimental Agriculture 4:321-328.

SAILA, S.B. & FLOWERS, J.M.(1969)
Georgraphic morphometric variation in the American lobster. Syst. Zool. 18:330-338

SCHALZOFF, M.(1966)
Exact distribution of Wilk's likelihood ratio criterion Biometrika 53:347

SCOTT, A.J. & SYMONS, M.J.(1971)
Clustering methods based on likelihood ratio criteria. Biometrics 27:387-397.

SEAL, H.L.(1968)
Multivariate statistical analysis for biologist. Methuen & Co., Ltd., London. 1st published 1964. 201 p.p.

SEARLE(1966)
Matrix alyebra for biological sciences.

SHORTER, R., BYTH, D.E. & MUNGOMERY, V.E.(1977)
Genotype x environment interactions and environmental adaptation. II Assessment of environmental contributions. Aust. Journal of Agricultural Research 28(2):223-236.

SIMLOTE, K.M.(1947)
An application of discriminant function for selection in durum wheats. The Indian Journal of Agricultural Science. XVII(Part V):269-280

SIMMONDS, N.W.(1962)
Variability in crop plants, its use and conservation. Biol. Rev. 37:422-465.

SMITH, H.F.(1936)
A discriminant function for plant selection. Ann. Eugenics 7:240-250.

SOKAL, R.R.(1966)
Numerical taxonomy. Scientific American 215(6):106-116. Dec., 1966.

STEEL, R.D.G. & TORRIE, J.H.(1960)
Principles and procedures of statistics: with special reference to the biological sciences. McGraw-Hill. New York.

STEPHENSON, W., WILLIAMS, W.T., & COOK, S.D.(1972)
Computer analyses of Petersen's original data on bottom communities. Ecological monographs 42:387-415.

SUCKLING, F.E.T.(1960)
Productivity of pasture species. New Zealand J. Agric. Res. 3:579-591.

VINAL, H.N. & HEIN, M.A.(1937)
U. S. department of agriculture Year book:1032 U.S.D.A.

WARD, Jr. J.H.(1963)
Hierarchical grouping to optimize an objective function. Journal of American Statist. Ass. 58:236-244.

WATKIN, B.R. & ROBINSON, G.S.(1974)
Dry matter production of "Massey Basyn" Yorkshire Fog (Holcus lanatus). Proceedings of The New Zealand Grassland Association 35(2):278-283.

WHITE, J.G.H.(1973)
Improvement of hill country pastures. Pastures and Pasture Plants, edited by R.H.M. Longer, 428 p.p.

WILKS, S.S.(1932)
Certain genetalizations in the analysis of variance. Biometrika 24:471-474.

WILLIAMS, W.T.(1971)
Principles of Clustering. Annual Review of Ecology and Systematics 2:303-326.

WILLIAMS, W.T.(1972)
The problem of pattern. The Australian Mathematics Teacher 103-109.

WILLIAMS, W.T. & CLIFFORD, H.T.(1971)
On the comparison of two classifications of the same set of elements. Taxon 20(4):519-522.

WILLIAMS, W.T. & DALE, M.B.(1965)
Fundamental problem in numerical taxonomy. Advancesin Botanical Research 2:35-68.

WILLIAMS, W.T. & LANCE, G.N.(1968)
Choice of strategy in the analysis of complex data.
The statistician 18(1)31-43.

WILLIAMS, W.T. & LAMBERT, J.M.(1959)
Multivariate methods in plant ecology. I. Association-analysis in plant communities. J. Ecol. 47:83-101.

WILLIAMS, W.T. & LAMBERT, J.M.(1960)
Multivariate methods in plant ecology. II. The use of an electronic digital computer for association analysis. J. Ecol. 48:689-710.

WILLIAMS, W.T. & LAMBERT, J.M.(1961)
Multivariate methods in plant ecology. III . Inverse association analysis. J. Ecol. 49:717-729.

WILLIAMS, W.T., LAMBERT, J.M. & LANCE, G.N.(1966)
Multivariate methods in plant ecology. V. Similarity analyses and information analysis.

WISHART, D.(1969a)
An algorithm for hierarchical classification. Biometrics 25:165-170.

WISHART, D.(1969b)
Numerical classification method for deriving natural classes.
Nature 221(January 4, 1969):97-98.