



NUMERICAL STUDIES OF LEAF ARCHITECTURE OF DICOTYLEDONS :

SAURAUIA (ACTINIDIACEAE) AS A TEST CASE

by

David T. Blackburn, B.Sc.(Hons.)

Department of Botany, University of Adelaide

Thesis submitted for the Degree of
Doctor of Philosophy, University of Adelaide

October 1978

Awarded August 1979

VOLUME I (containing text and text figures)

ABSTRACT

A leaf architectural study of a genus of extant plants has been carried out to develop numerical approaches for defining leaf characters. Existing architectural classifications of dicotyledonous leaves rely upon terminologies such as that of Hickey (1973). A set of numerical characters has proven efficient in separating the species of the genus Saurauia. They do not suffer from the weaknesses of descriptive terminologies such as class weighting ambiguity and non-ordered and binary characters. The value of using multiple measurements, especially ranges or standard deviation based measures of dispersion is demonstrated. This is recommended in place of single values such as means or those derived from chance exemplars. The inadequacy of existing statistical procedures in handling multiple measurements is also shown. A new similarity/distance metric which generalises previous measures of resemblance is defined. This metric is applied to the set of 10 numerical leaf architectural characters defined for Saurauia to generate a leaf based classification for 92 species. Data derived from type descriptions has also been treated in the same way to prepare a floral based classification for a set of 193 species. The alternate classifications have been compared and the leaf based one shown to be more efficient at defining species. Finally the similarity matrix for type description data is applied to the problem of explaining a major morphological disjunction in the genus which is shown to parallel Wallace's line. This disjunction is explained in terms of the drift of the Asian and Australian continental blocks through the Tertiary and the influence of that movement on the migration and evolution of the Old World Saurauia species. The history of migrations across Wallace's line is shown to be more complex than has been previously indicated.

DECLARATION

This thesis contains no material which has been accepted for the award of any other degree or diploma and, to the best of my knowledge, contains no material previously published or written by any other person except where due reference is made.

David T. Blackburn

ACKNOWLEDGEMENTS

I am indebted to my supervisor, Dr. D.C. Christophel, for his advice and encouragement throughout the duration of this project, and especially for his criticisms of the thesis. I would also like to acknowledge the support of my friends and colleagues in the Botany Department, University of Adelaide, with whom many stimulating discussions have been had. My appreciation is also expressed to Mr. D. Colquhoun, Computing Science Department, University of Adelaide, and Dr. M.H. Andrew and Ms. C. Annells, both of the Botany Department, University of Adelaide, who have assisted me with computing. For his assistance with taxonomic problems, and especially for the many specimens of vegetative material which he donated, I would like to express my thanks to Dr. R.D. Hoogland of the Research School of Biological Sciences, Australian National University, Canberra. I would also like to acknowledge the help that I have received from Dr. M.M.J. van Balgooy, Rijksherbarium, Leiden who specially collected extensive quantities of Malaysian species for me. For access to collections within Australia, I would like to thank Dr. H.J. Eichler, Herbarium Australiense, Canberra, and Mr. S. Everist and Dr. R.W. Johnson, Queensland Herbarium.

Finally, I would like to thank the staff of the following institutions for the long term loan of many precious specimens of Saurauia: Arnold Arboretum, Harvard University; Institutum Botanicum, Academia Sinica, Peking; Botany Laboratory and Herbarium, Agriculture Department, Fiji.

TABLE OF CONTENTS

Abstract

Declaration

Acknowledgements

1. INTRODUCTION	1
Aims of the study	5
Selecting a genus	6
Taxonomic history of <u>Saurauia</u>	8
2. APPROACHES TO THE CLASSIFICATION OF BIOLOGICAL ENTITIES	12
Numerical methods in classification	14
Measures of resemblance	16
Distance coefficients	16
Association coefficients	23
Class intervals and metric characters	29
Correlation coefficients	37
Probabilistic similarity indices	37
Conditional binomial probability	42
3. THE CHOICE OF A LEAF CHARACTER SET	43
Types of leaf characters	43
The set of characters under study	45
Architectural characters	46
Characters of the leaf indument and epidermis	57
Differentiating vein orders on leaves	65
Relationship of the character set to existing terminologies	69
Choice of the ultimate character set	70
4. PRACTICAL PROCEDURES OF DATA COLLECTION	75
Practical approaches to data collection	75
Leaf clearing	82
Measuring non-architectural characters	88
Scoring leaf characters for analysis	89
5. AN ULTIMATE CHARACTER SET	90
A trial set of species	90
Variability and correlation of characters within species	100

Environmental influences on leaf morphology	109
Variance analysis	124
Unique information content of characters	129
Correlation between characters	130
6. TAXONOMIC PARTITIONING OF THE SET OF SPECIMENS	149
Introduction	149
Sample sizes for species	149
The sections of the genus	150
Specific assignment of specimens	153
7. APPLYING MEASURES OF RESEMBLANCE TO THE SET OF OTUS	155
The multistate similarity/distance measure	155
Missing values	155
Characters scored as ranges	159
Distribution of data	164
Recommendations for minimising distortion of distances	165
TAXON, a computer program for evaluating distance/ similarities between pairs of OTUs	165
Size limits on the data set	166
Output	168
Using the program	169
Program limitations	169
Setting the DISCARD value	169
Data format	170
Run name	171
Deck structure	171
Ordination of specimens by similarity	171
Details of the phenogram	176
Defining species limits	177
8. CLASSIFICATION OF <u>SAURAUIA</u> SPECIES	192
Leaf based classification of species	194
Classification of species on type description data	209
Type data phenograms and nearest neighbour trees	217
9. INFORMATION CONTENT OF GRAPHICAL REPRESENTATIONS OF RELATIONSHIPS BETWEEN OTUS	245
Phenograms	245
Higher order neighbour plots	251

10.	BIOGEOGRAPHY AND EVOLUTION OF THE OLD WORLD SPECIES OF <u>SAURAUIA</u>	265
	Biogeographical divisions	269
	Endemism in the genus	270
	Similarity between species	273
	Nearest neighbours	274
	Number of species	275
	Migration routes	291
	Biogeography of the generic sections	292
	Conclusions	310
11.	DISCUSSION	313
12.	BIBLIOGRAPHY	326

VOLUME II

APPENDICES

CHAPTER 1

INTRODUCTION



The bulk of angiosperm taxonomy is based on floral features. Leaves have generally been neglected. The reason for this is in part due to a long standing attitude amongst taxonomists of extant plants, that leaves are unreliable structures, subject to the vagaries of the environment (Deane, 1900; Hamilton, 1916; Odell, 1932; Stebbins, 1950; Good, 1966). To a degree this is true. Flowers, whose specialisations are expressed for only a short period of function, have little bearing on the every day survival of individual plants. Leaves, however, being part of the functional supporting system for the reproductive components, must needs meet the requirements of the local environment.

To most taxonomists, the concept that a leaf based taxonomy for the angiosperms is not only possible but desirable, is foreign. Yet there are commonly situations in which only vegetative structures are available. Frequently, collections of unidentified herbarium specimens languish, because their flowers are immature, or over-mature, or lacking altogether. Educated guesses by workers familiar with the floras will occasionally allow such material to be labelled "sp.aff." This problem is especially manifest in collections of tropical plants, because of their lack of seasonal flowering. It is impossible to plan collecting expeditions to these often inaccessible floras, when a majority of species will be in flower.

For the most part, palaeobotanists are faced with primarily vegetative remains. Fossil flowers, attached to leafy branches are rare finds. The incentive for a leaf based taxonomy has, as a result, come from palaeobotany. A foremost palaeobotanist of the 19th century Baron C. von Ettingshausen, conducted major surveys of Tertiary fossil

floras from both Europe (1852, 1870, 1871, 1878) and the Australian region (1883, 1886, 1891). He largely based his identifications of fossil leaves upon his study of the venation patterns of dicotyledonous leaves (1861). In this work, von Ettingshausen surveyed and illustrated several hundred species of dictyledons. For the most part he applied generic names of living plants to the fossils, and from this he originated his concept of the cosmopolitan flora, with genera such as Fagus, Quercus, Acer, Betula, Banksia and Hakea, spread throughout the Tertiary world.

This floristic approach, whereby extant generic names are applied to fossil plants, has dominated palaeobotanical studies until quite recent times. Inherently the approach is not invalid. The problem with it is that our knowledge of the morphology and variability of extant dicotyledonous leaves is totally inadequate even at the present time. Palaeofloristic studies in the past have been conducted on the basis of the knowledge of living floras that have come within the experience of the palaeobotanist concerned. Coupled with the apparent unwillingness of palaeobotanists to accept that many genera extant in the Tertiary are now extinct, a large proportion of fossil angiosperms are misidentified. It has been estimated that approximately 60% of the Eocene fossils that Berry (1916, 1922, 1924, 1930, 1933, 1941) assigned to extant families and genera are incorrectly identified (Dilcher, 1971). The taxonomic assignment of these fossil leaves has been carried out by "picture matching" (Wolfe, 1972, 1973) to extant leaves with which the worker is familiar.

Recently, the concept of a parataxonomy for extant and fossil leaves has been advocated (Ferguson, 1971; Hickey, 1973; Dilcher, 1974). Ferguson applied this principle to the study of a Miocene flora from Germany. I have applied it to a study of an Eocene flora

from South Australia.¹ In such studies, large sample sizes are important to be able to assess the variability of leaves. Ferguson in Germany investigated approximately 2000 fossil leaves over 76 species. The continuing survey of the Maslin Bay flora is based on a similar number of specimens over about 200 taxa.

The generation of an autonomous parataxonomy for dicotyledonous leaves, whether of fossil or extant plants, requires both a careful analysis of a large sample of specimens and a highly efficient character set for the discrimination of specimens into taxa at the species level.

The characters available for the taxonomic delimitation of leaves can be considered to be of three kinds; architectural, cuticular and anatomical. Architectural characters, sensu Hickey (1973) are those characters that outwardly express leaf structure, including venation pattern, marginal configuration, shape and gland position. Cuticular characters can be considered as those features expressed or retained in the cuticular membrane. Such features include cuticular surface ornamentation, pattern of epidermal cell arrangement, notably the stomatal complex, and the structure and placement of trichomes. Anatomical characters are considered to include features such as the pattern of vascularisation in the petiole, the form and placement of secretory structures within the leaf mesophyll and the placement and structure of sclereids.

Although each of these classes of characters contain taxonomically useful information, they are available to differing extents to students of leaf form. A forester, for example, may not have the necessary equipment or the training to allow an anatomical or cuticu-

1. Blackburn, D.T. 1973. The macromorphological leaf spectrum of a lower middle Eocene, carbonaceous clay lens, at Maslin Bay, South Australia. Unpublished Honours thesis. Department of Botany, University of Adelaide.

lar investigation in the field. A palaeobotanist may find that these features are not preserved. Conversely only dispersed cuticular remains may be preserved and architectural features lacking. This latter case is the rule for fossil plant material recovered in drilling operations.

An investigation of a leaf assemblage should take into account all available features of the leaves (Ferguson, 1971; Dilcher, 1971, 1974). For the most part, however, palaeobotanical investigations can be based on only a subset of features. Leaf impressions which preserve at least some orders of architectural features are the most common mode of preservation of fossil leaves. Cuticular features are less commonly preserved and anatomical characters are rarely present. For an impression fossil to be recognisable as a dicotyledonous leaf, it must have at least some primary and secondary venation details preserved. The spatial relationships of these vein orders are generally sufficient to tie the fossil to class level. The further identification of the leaf within a taxonomic system requires increasing degrees of preservation as greater resolution in assignment is needed. The degree of perfection in preservation of fossil leaves is however limited by a number of depositional and diagenetic factors. Abrasion in transport, degradation by insects and micro-organisms, chemical degradation by high pH and high oxygen potential in the sediment, all influence the condition of the fossil. Sediment grain size is also important in determining the minimum size of structures preserved in an impression. The nature of individual leaves also influences how they will be fossilised. Thin mesophyllous leaves, such as one finds in tropical and subtropical floras, are likely in many cases to lack cuticular remains. This is the case in Tertiary floras of southern Australia, which represent sub-tropical vegetations. Although these

floras often have preserved cuticular remains, there are many cases where the cuticles are not recoverable for study. A distinct taxonomic bias has been noted in species possessing recoverable cuticles in a flora from Nerriga, New South Wales (pers. comm. R.Hill, Dept. Botany, University of Adelaide). Thus in any fossil flora, there will be for all taxa, some degree of missing information. In the case of a marginally workable flora, leaves should possess all orders of architectural features from shape and marginal configuration, to tertiary venation. It is doubtful whether at the present time, fossils preserved only to the level of secondary venation would be of any use, unless they can be assigned as duplicate specimens of taxa from the same flora.

It is important therefore to define a set of characters, of no higher order than tertiary venation, which have been thoroughly studied and tested on a set of taxa of known taxonomic structure. This limitation to only macroscopic features is important if an architectural approach is to find general application in leaf studies. The higher order characters can then be used within this framework where they are applicable in individual studies.

Aims of the study

To test the degree of taxonomic resolution achievable with only the features of shape, size, marginal configuration and primary, secondary and tertiary venation, it was decided to study an extant plant genus in detail, using only those features. The leaf features were then to be applied, via suitable statistical procedures, to the task of generating an independent leaf based classification. Lastly the classifications achieved alternately with leaves and classical taxonomic features could be compared.

Selecting a genus

The chosen extant genus needed to meet several criteria. Firstly, the number of species should be fairly large, e.g. more than 50, to give a reliable data base. Secondly, there should be a fairly reliable taxonomic framework, as a taxonomic revision was not intended. Thirdly, herbarium collections should be accessible, and fourthly, the leaves of the genus should lend themselves to a study of this kind. This latter criterion merely means that venation features should be readily discernible on the leaves. A genus like Eucalyptus where there is almost no vein topography, was considered unsuitable. Finally, it was decided that a tropical genus was most desirable. This last criterion relates to the findings of surveys of Australian Tertiary floras, currently in progress. Studies of three floras from Southern Australia, being conducted in our laboratory, indicate that generally sub-tropical conditions prevailed during the lower Tertiary (Christophel and Blackburn, 1978; Blackburn²).

This is also supported by Tertiary floras from the Northern Hemisphere (Reid & Chandler, 1933; Becker, 1961; Wolfe & Hopkins, 1967; Wolfe, 1970). Other independent methods support these climatic inferences e.g. oxygen isotopes of benthonic animal tests (Emiliani, 1954, 1960; Dorman, 1964) and the palynological record (Martin, 1977). Therefore new leaf architectural methods, tested on a group of tropical organisms, are more likely to produce data which can be extrapolated to the treatment of Tertiary floras, than if established using a temperate genus.

The genus Saurauia Willd. of the Actinidiaceae was selected for study. Saurauia is widespread in the tropics of both the New and Old

2. Blackburn, D.T. Tertiary megafossil floras of Australia. Systematic descriptions of fossil leaves from Maslin Bay. In press.

Worlds, and has a large number of species. Estimates of species numbers range from about 60 (Hooker, 1875; Willis, 1908) to about 250 (Gilg & Werdermann, 1925; Melchior, 1964). If one accepts the more recent estimates of species numbers, there are about 185 species distributed through the region from Himalayan India, to New Guinea and Polynesia. The New World species, numbering about 65, are found from Mexico to Argentina (Hunter, 1966; Soejarto, 1968). The number of species is discussed in more detail in chapter 10.

Saurauia has not been chosen for its size and distribution alone. The genus is and has been in recent times, the subject of some taxonomic revision. Although no complete revision has ever been carried out, local groups of species have been well studied. Hunter (1966) revised the species of Central America. Previous to this Buscalioni (1912-1927) revised the American species, but there are some doubts as to the reliability of his work (Macbride, 1956). In the Indo-Malayan region, no such major revision has been conducted, but revisions of local groups have been commenced by Hoogland (1977). Keys to regional groups are also available for the Old World species (Finet & Gagnepain, 1905; Diels, 1922; Koorders, 1923; Backer & Bakhuizen van der Brink, 1963), and it was felt that these and other literature would provide a workable base with which to compare the leaf classification.

Herbarium collections of Saurauia were available both within Australia (Herbarium Australiense, Brisbane Herbarium) and from institutions in other countries, notably the Arnold Arboretum. However, all of these collections contained only Old World species. It was therefore decided to restrict the study to species of the Indo-Malayan region. This is not considered to be a real limitation, as it has since been demonstrated that the region is the centre of

diversity of the genus. Exclusion of the American species discards very little of the range of leaf morphology within the genus.

The final reason for selecting Sauraula relates to the form of the leaves. Generally they have prominent secondary veins, easily distinguished from other vein orders, and proceeding in simple arcs from the midrib to the margin. Since the study was intended to be a numerical approach to describing leaf architecture, such a vein pattern lends itself to quantification. Sauraula also has a remarkable range (6-60) in the number of secondary vein pairs possessed by different species. This further supports the idea that a leaf study of the genus was possible using only those characters of lower order than quaternary venation.

Taxonomic history of Sauraula

An idea of the complex history of Sauraula species can be gained from the annotated bibliography in Appendix 10. Therefore only a brief account of some of the major events in its history will be given here.

The generic name Sauraula was first used by Willdenow (1801). The correct spelling of the name has been the subject of considerable debate, having been spelt alternatively as Sauramia, Sauravia, Saurauja and Sauraula. The last mentioned spelling is now generally accepted (Hunter, 1966; Hoogland, 1977), and is used in this study.

Several attempts have been made to classify the species into generic sections. Buscalioni (1912) proposed a system based on the indument of leaves, flowers and stems. A similar system incorporating features of the inflorescence as well, was used by Diels (1922) and later accepted by Gilg and Werdermann (1925). There exists no more recent classification for the Old World species than Diels' system. A revision of Mexican and Central American species by Hunter

(1966) includes a proposal for a classification incorporating features of leaf venation, and pubescence of the sepals and ovary. Although this appears to be suitable for New World species, features such as the degree of prominence of secondary and tertiary veins are inapplicable to the large number of morphologically diverse species from the Old World. Hunter also defined the types of trichomes occurring in the set of species with which he worked. This was done in response to the poorly defined terminology of Buscalioni. Although Hunter's trichome types are less ambiguous, they still suffer from the problem of intermediate forms. A particular case is in the difference between sericeous and loriform hairs. The latter type is defined as follows. "The long flexuous tip of the loriform hair, which often becomes entangled with the ends of neighbouring hairs, is the only feature which distinguishes it from the sericeous hair" (Hunter, 1966). This is not an isolated example if one studies Hunter's terms further. The unfortunate aspect of this is that such terms become accepted as describing fundamentally different trichome types. It is also impossible to apply these terms to pre-existing descriptions, simply because such features of the indument have not been adequately described. The problem of applying new and re-defined terminology to previously described species has also been considered by Smith (1941). Burt (1936) criticised the system of Diels on the grounds that features of inflorescence type, bract form and leaf pubescence are not fundamental characters. Smith accepted this, but noted that more fundamental features such as hair type, length and direction, flower size, stamen and style arrangements have been inadequately described. As a result he applied Diels' system.

Diels proposed four main sections, defined principally on the form of the inflorescence and the nature of floral bracts. The last

of these four divisions has been further divided into seven subsections based mainly on leaf and bract indument. The system of Diels is considered in detail in chapter 6. Contrary to the opinion of Burtt (1936), the present study lends support to at least some of the sections representing fundamental divisions of the genus. Diels' sections are later applied in a discussion of biogeography and evolution within the genus.

Other classifications

There has been no classification other than that of Diels, designed for all of the Old World species. A number of regional classifications have been set up in the form of keys. It is worth noting that these closely follow a form resembling the system of either Diels or Hunter (1966). The primary divisions in keys to the Indian and Malayan species are made on features of the inflorescence (Dyer, 1875; Ridley, 1922). Keys to south-east Asian species (Finet & Gagnepain, 1905) and Javan species (Koorders, 1923; Backer & Bakhuizen van der Brink, 1963) base their primary divisions on the indument of the ovary and sepals. Baker (1918) proposed classification of Sumatran species which differs little from the later system of Diels. The most extreme division of the genus was that proposed by Choisy in his "Mémoire sur les Ternstroemiaceae" (1855) which I have not seen, but which is cited in Baker (1918). Choisy divided the species into two genera on the basis of style number. Species with five styles were placed in Saurauia, while those with three were placed in Scapha. Although style number is somewhat variable in many species (Baker, 1918; Hoogland, 1977), evidence offered in chapter 10 for a taxonomic disjunction on the character, supports the existence of two essentially different groups. The discussion in that chapter relates the biogeography of 3- and 5-styled species to

a well-known taxonomic disjunction, marked by Wallace's line. Although the data for that chapter is derived from type descriptions, wherein ranges of style numbers are not generally given, the probability that there is a real disjunction in style number is very high. The present study therefore supports, for a flower based classification, the retention, at least in part, of Diels' system, and a possible division on the basis of style number. The following chapter discusses alternate approaches to classification in greater detail.

CHAPTER 2

APPROACHES TO THE CLASSIFICATION OF BIOLOGICAL ENTITIES

A. page
Classification in the sense in which it is applied here can be defined as the ordination of organisms into groups on the basis of their phenetic relationships. The result of this procedure is the generation of a scheme in which individuals are arranged along the dimensions of t -space. This defines the underlying aim of a classificatory procedure, which is to assess whether in the sample of individuals under study there can be discerned a spatial structure, which we assume reflects their true biological relationships. This can be considered at all levels of refinement to be a pattern recognition procedure. In a simple case of classification such as that of arranging a set of two dimensional geometrical forms into a hierarchy, a visual assessment of relationships may suffice. For example, the set of regular polygons can be ordered alternatively by increasing roundness, increasing number of sides, decreasing area per unit length of side, increasing sum of the internal angles, or increasing number of axes of symmetry. The possibility of a descriptive property having more than one method of expression is clearly illustrated, and the problems of this effect are discussed in the section on correlations between characters. Clearly the visual classification of polygons is not necessarily a simple univariate procedure. There are a number of pathways to the same result. The pathway that is used is dependent on the function of the classification. A designer of wheels would see the series of polygons as an ordination on the basis of those shapes that roll better. An experimental subject in psychology may group the shapes into those that have the highest correlation with food and the lowest correlation with unpleasantness.

It is important to be aware of the non-uniqueness of even simple classifications.

Classification in the biological sense is subject to even greater possibility of non-uniqueness. The classifications required for vegetation studies are fundamentally different from the phylogenetic ordination of organisms in evolutionary studies, from the phenetic classifications of extant organisms, and from those whose purpose is identification. For this study the only requirements of a classification are that it reflect the phenetic relationships of a group of organisms and that it be useful in classifying new individuals into pre-existing phenetically important groups.

Historically the most important approach in systematics has been the classification of organisms on the basis of their phenotypes using visual estimation of similarity or resemblance between pairs of organisms. Such a method was the only one available before the advent of high speed data processing computers. This procedure involves (a) a subjective assessment of the variation of a set of characters within the set of organisms under study, (b) the subjective weighting of some characters over others, based on variability and pre-existing notions of their phylogenetic importance, and (c) a subjective classification to fit the requirements of the taxonomist. The existing taxonomic system for all organisms is based upon such an approach. Not surprisingly the classifications so produced are far from unique, and at all levels the status of organisms and groups of organisms is the subject of much debate. In recent years an approach which seeks to minimise the subjectivity of classificatory methods has become widely used. This approach is variously called taxometrics (Mayr, 1966), mathematical taxonomy (Jardine & Sibson, 1971), multivariate morphometrics (Blackith & Reyment, 1971) and numerical taxonomy

(Sneath & Sokal, 1973). The latter term is defined by Sneath and Sokal as "the grouping by numerical methods of taxonomic units into taxa on the basis of their character states."

Numerical taxonomy is a new and presently developing science which is multidisciplinary both in approach and application. The methods of numerical taxonomy have their basis in the realms of mathematics and statistics, and for all but small studies, their implementation requires the use of computers. This removal of taxonomy from the areas of experience of most classical taxonomists has meant that in those fields where the classical methods have historically dominated, the numerical approach has received little acceptance. This has been compounded by the fact that in many cases numerical methods have not been successful in producing classifications of sets of organisms. These failures may be the result of a combination of effects. The taxa under study may be, in the real biological sense, indistinguishable from one another so that an objective assessment of their variability reveals no clustering into more or less separate groups. Alternatively, failure may result from inadequacies in the procedures used, whether because they were incorrectly applied or because of limitations of the methods themselves. That numerical methods have limitations in respect to the classificatory problems they can accomodate, is amply shown by the fact that there is no one universal algorithm for the classification of organisms. With wider usage of numerical methods it is becoming apparent that there exists a small subset of approaches which are most commonly used. These approaches and their relevance to the present study are discussed below.

Numerical methods in classification

A survey of recent papers on classification of groups of

organisms, particularly plants, has shown that a relatively small number of numerical taxonomic procedures are commonly used. In agglomerative polythetic classifications, one starts with individual organisms, having no prior concept of their classification, and groups (agglomerates) them on the basis of their overall similarity. Similarity is defined by a set of attributes, where no one attribute is fundamental to any step of the classificatory procedure. Polythetic refers to the multiple conditions that need to be fulfilled in a classificatory step. This approach is differentiated from divisive (or splitting down from a combined cluster of all OTUs) in that there is no prior classification and monothetic (where single attributes are fundamental) in that no single feature of the organisms is used to define any grouping. Monothetic procedures are commonly used in identification by keys where groups are defined by a rigid and logical set of divisions. Such a grouping is not "natural" except by chance and should be seen only as a simple, rapid approximation to defining the status of a new individual. Divisive procedures such as the classification of an organism into successively lower levels of the taxonomic hierarchy, i.e. division, class, order, family etc., are usually based on monothetic criteria.

Most numerical taxonomic studies start with a set of unclassified individuals, scored for a large number of numerically definable attributes. Using measures of similarity, clustering (notably UPGMA, Sokal & Michener, 1958) and ordination, e.g. principal component analysis, the relationships between individuals in t-space are determined. These results are then generally condensed to no more than three dimensions and presented by one of the various graphical means, e.g. phenograms, minimum spanning trees, contour plots, stereograms, scattergrams or similarity matrices. It is also possible on some

data sets in which there is some a priori basis for dividing taxa into a number of groups, e.g. plants grouped into geographic regions, fossils in horizons, people into political groups, to investigate relationships using a technique such as discriminant analysis. Conceptually however, the simplest procedure is an analysis of similarity between pairs of organisms using measures of similarity or resemblance. These will therefore be discussed in the following section.

Measures of resemblance

Measuring of taxonomic resemblance of pairs of OTUs (operational taxonomic units) is generally accomplished by one of the following procedures. Similarity coefficients are used to compute the resemblance between the elements of the two rows of the data matrix which represent the attributes of the two OTUs under comparison. A host of similarity coefficients exist, many of them empirically determined, and all of them variously applicable or inapplicable to different kinds of data.

Similarity coefficients can be grouped into four more or less distinct types, (1) distance coefficients, (2) association coefficients, (3) correlation coefficients, and (4) probabilistic similarity indices. These were defined by Sneath & Sokal (1973), but will be considered here concurrent with the derivation of new similarity coefficients. In the following section, all measures of resemblance unless specifically acknowledged by reference to the literature, have been derived in the course of this study.

1. Distance coefficients

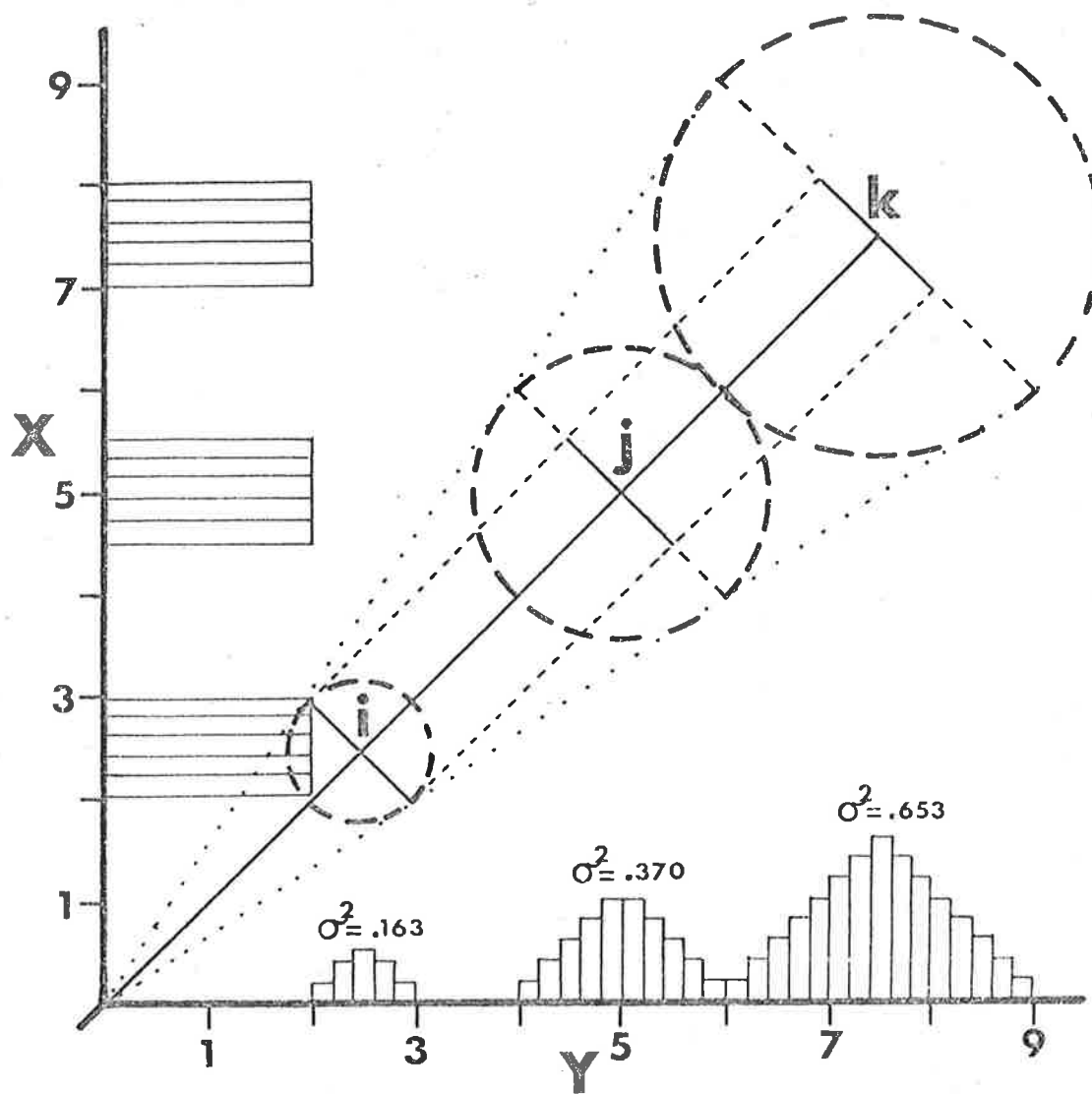
Distance coefficients compute the distance between OTUs in either Euclidean or non-Euclidean space. Distance coefficients can be seen to be the complement of similarity measures, in other words,

measures of dissimilarity. The simplest distance measures compute the numerical difference between the character states of two OTUs where each is represented by a single value on the character. The distances for each character are then either averaged over all characters as in the case of mean character difference (M.C.D., Cain & Harrison, 1958), or else the square root of the sum of squares of character distances is found as in the Euclidean taxonomic distance of Sokal (1961). These two measures are special cases of the Minkowski metrics whose general form is

$$d_r(j,k) = \left[\sum_{i=1}^n |x_{ij} - x_{ik}|^r \right]^{1/r}$$

Two further metrics of this form are the Manhattan metric and the Canberra metric of Lance and Williams (1967). This last is a non-linear metric which is sensitive to proportional rather than absolute differences. For example, the Canberra metric considers that 4 is larger than 2 to the same degree as 2 is larger than 1. Relative to other linear metrics this overestimates the distance between OTUs. For some data sets e.g. those where the expected variation in measurement is proportional to the absolute value, this may be a desirable property (see fig. 1). The effect on numerical procedures of data where variance is proportional to the absolute value seems to be little considered. This effect can be visualised in a scattergram of two metric variables with distinct clustering of the OTUs (fig. 1). Consider three bivariate normal clusters whose centroids are (X_i, Y_i) , (X_j, Y_j) and (X_k, Y_k) . The frequency distributions of the Y variable are projected onto the Y-axis using a triangular distribution as an approximation to normality. If the univariate variances are calculated for each cluster it can be seen that values of variance increase with increasing Y values. Conversely

FIGURE 1. Hypothetical clusters (i, j, k) of OTUs where variances are proportional to cluster centroid values (area outlined by circles) or variances are independent of size (rectangular dashed outline). For further explanation refer to text.



in the case of a variable with a fixed variability, illustrated by three equal length line segments passing through the centroids, the variances computed from frequency distributions projected onto the X-axis are equal. For the first case a non-linear metric could be justifiably used for the computation of distances. However, a non-linear metric would distort the significance of distances computed for the second case. It appears to be true that in most cases, the significance of distances between OTUs is assessed as a linear function independent of distance from the origin. This is illustrated by the fact that most graphical representations of OTUs in 2-space use linear axes, e.g. scattergrams of discriminant analyses, PCAs and regressions on characters for OTUs. Rarely are axes scaled to allow for this effect. Various procedures for scaling are discussed in Sneath & Sokal (1973). However, in practice it appears that the only place non-linear scaling, such as using the log. or nth root of the values, is performed, is where multidimensional characters such as volume and area are being used with single dimensional (length) characters. One instance of a need for log. scaling of length values for an OTU set has been observed in a study of the lengths and widths of achenes of Senecio hybrids (pers. comm., M.E. Lawrence, Botany Dept., University of Adelaide). The clustering of individuals into hybrid groups is such that the variance of the clusters increases with increasing achene size.

In most cases it appears that biological data can be adequately treated by linear approaches. The basic form of existing Euclidean distance measures has been shown above. Inherent in all of this class of measures, is that the significance of individual distances can be assessed only in comparison with all other distance values for the OTU set. This problem is overcome to some extent by

standardising characters, assuming independence of characters and a normal distribution. Standardisation is performed by subtracting the mean value of each character from distance values computed on that character and by then dividing by the variance of the character. This standardises a character assumed to be normally distributed to zero mean and unit variance. Such a metric has an indeterminate range although it has been shown by Rohlf (1962) that the expected value of distance approaches 2 as the number of characters tends to infinity.

Gower (1971) has suggested the use of ranging to equalise size and variability between characters. This procedure gives the smallest state of a character the value 0 and the largest state the value 1. The ranging of characters can be accomplished in one step with the computation of distances, by using a metric of the following form.

$$d_{jk} = \sum_{i=1}^n \left(\frac{|x_{ij} - x_{ik}|}{|x_{\max} - x_{\min}|} \right)$$

This gives a measure of distance which is scaled to the total range of characters over the data set. The disadvantage of this, however, is that where the population is small, addition of new OTUs is likely to have a drastic effect on the range. The instability of the range can be overcome by the use of variance or standard deviation as discussed above. The indeterminate values of distance are however clearly a disadvantage if single values are being considered in isolation. An alternative to the variance is the studentised range. This parameter is dependent on sample size and standard deviation and can be used, for example, to estimate the likely range at a given probability level for k variates estimated from a sample

of v variates. The value of v can be greater than, less than, or equal to k . The likely range of k can be computed from values of the studentised range $Q_{a(k, v)}$, where $Q = \text{Range}/\text{stand. deviation}$. Values of Q can be obtained from Table U of Rohlf and Sokal (1969). The relevance of studentised range to distance metrics lies in the fact that distances can be standardised to the studentised range. For a given critical value such as 1% or 5% a distance metric can be said to lie in the range 0 to 1. The distance measure has the form:-

$$d_{kv} = \sum_{i=1}^n \left(\frac{|x_{ij} - x_{ik}|}{Q_{a(k, v)}} \right)$$

A further approach is the generalised distance of Mahalanobis (1936) and Rao (1948). This approach has found application in discriminant analysis but seems to have been little used elsewhere. Generalised distance is computed by maximising the difference between pairs of means for those combinations of characters that have maximum variance between pairs of groups compared to within group variance. The computation of generalised distance assumes multivariate normality although discriminant analysis is robust to violations of this.

The need to be able to process more than single values for each character has led to the development of Pearson's (1926) coefficient of racial likeness. This coefficient computes distances between samples of OTUs for which mean and variance have been computed, taking sample size into account. For studies of OTUs where the sample sizes of individual taxa are large this approach appears to be the most meaningful metric available. For small samples the use of this method may not be justified because of low confidence in using small sample means and variances as estimators of population parameters.

A linear distance measure capable of handling means and ranges of metric characters, variability in non-metric hierarchical or non-hierarchical characters and binary characters has therefore been developed for use within the context of this study. Before the measure and its derivation are discussed, it is worth considering a further set of resemblance measures, the association or similarity coefficients.

2. Association coefficients

There are a large number of measures of resemblance which can be described as association coefficients. The simplest of these are the well known Jaccard, Yule and Simple Matching Coefficients. These are only readily applicable to binary or two state characters. The information loss in recoding metric (continuous numeric) characters to two states does not justify their usage with this kind of character. The General Similarity Coefficient of Gower (Gower, 1971) is applicable to all classes of characters, assigning a weight for each comparison, such that matches are given a score of 1 and mis-matches are set to 0. The major disadvantage of this is that, unlike distance measures, the significance of a mis-match cannot be judged.

A combined similarity index/distance measure for all classes of characters

A number of measures of resemblance can be seen as both distance measures and similarity coefficients. The coefficient of Gower (1971) is the best example of such a coefficient which is also applicable to mixed character sets.

The simplest statement of a measure of resemblance (I_a) for two OTUs is:-

$$I_a = \frac{C_a}{C_v}$$

where C_a is the number of characters on which the two OTUs are matched and C_v is the number of characters on which they are validly scored. This statement is exactly equivalent to the Coefficient of Jaccard, a measure of association applicable to binary characters, and which can be stated as

$$S_j = \frac{a}{a + b + c}$$

where a = number of matches, and b and c = number of mis-matches.

A 2 x 2 contingency table will serve to illustrate the relationships between a , b , and c .

		OTU j		
		1	0	
OTU k	1	a	b	$a + b$
	0	c	d	$c + d$
		$a + c$	$b + d$	$n = a + b + c + d$

FIGURE 2

The fourth term d , represents the number of double negative matches.

For multistate characters I_a can be rewritten as

$$I_a = \frac{2C_a - T_j - T_k}{T_j + T_k} \quad \dots \dots \text{eqn.1}$$

T_j & T_k = total number of character states across all characters possessed by OTUs j and k respectively. Where only single values are scored for each OTU in each multistate character, $T_j = T_k = C_v$. I_a is now a measure of dissimilarity and can be written as a dissimilarity coefficient of the form

$$I_a = \frac{2a - (a+b) - (a+c)}{(a+b) + (a+c)}$$

$$= \frac{-(b+c)}{2a + b + c} \quad \dots \dots \text{eqn.2}$$

where $T_j = a + b$, $T_k = a + c$ and $C_a = a$.

A worked example will serve to illustrate this.

Example 1

Characters	1	2	3
OTU j	+-	+-	+-
OTU k	+-	+-	+-

where $+$ = 1 and $-$ = 0 or respectively presence and absence of a character state.

Referring back to fig.2 it can be seen that $a = 2$, $b = 1$, $c = 1$.

$$I_a = \frac{-2}{6}$$

The meaning of this value in expressing the degree of mismatch of OTUs j and k is that out of three possible matches, one character is not matched.

It is interesting to note that except for sign this expression (eqn.2) is exactly equivalent to the "non-metric coefficient" of Watson, Williams and Lance (1966, 1967). The logical extension of eqn.1 is to an association measure for multistate characters where OTUs can possess more than a single state for each character. The example below illustrates such a case.

Example 2

Characters	1	2	3
OTU j	++	+-	-++
OTU k	++	+-	++-

This condition cannot be expressed in terms of a 2×2 contingency table. Neither can it be adequately expressed by eqn.1. Working the value of I_a for this example gives

$$I_a = \frac{2C_a - T_j - T_k}{T_j + T_k} = \frac{2(3) - 5 - 5}{5 + 5} = \frac{-4}{10}$$

In Example 1, I_a gave a value which expressed the number of states involved in mismatches as compared with the number of possible matches. To be consistent in Example 2, I_a could be expected to

give a value of $-2/10$. To correct this, eqn.1 can be written as

$$I_a = \frac{2C_a - T_j - T_k + \text{corr.}}{T_j + T_k}$$

where corr. is a correction term. Since only one character is involved in a multiple match in Example 2, the correction can be written as $2(m - 1)$ where m is the magnitude of the match. ^{per character} In Example 2, $m = 2$, and the correction is $+2$. A further worked case will serve to illustrate that this term is the general solution.

Example 3

OTU j	++++
OTU k	++++

$$I_a (\text{uncorr.}) = \frac{2(1) - 4 - 4}{4 + 4} = \frac{-6}{8}$$

$$I_a (\text{corr.}) = \frac{2C_a - T_j - T_k + 2(m-1)}{T_j + T_k} \quad \dots \dots \text{eqn.3}$$

$$= \frac{2(1) - 4 - 4 + 2(4-1)}{4 + 4} = \frac{0}{8}$$

This value clearly means that of 8 total states, there are 0 mismatches. $I_a (\text{corr.})$ can now be seen as a dissociation coefficient ranging from 0 to -1 . For the general case of multiple characters, $I_a (\text{corr.})$ is written as

$$I_{\text{diss}} = \frac{2C_a - T_j - T_k + \sum_1^m 2(m-1)C_m}{T_j + T_k} \quad \dots \dots \text{eqn.4}$$

where C_m = number of characters in a comparison of OTUs j and k with multistate matches of magnitude m .

It is easy to see that eqn.4 can be expressed as a true multi-state distance measure by allowing m to be negative for mismatches e.g.

OTU j	+++-----
OTU k	-----++++

where there are three states between the ends of the ranges of the

character scores for the two OTUs. Eqn.4 can be rewritten as a general Euclidean distance measure in the form

$$I_D = \frac{2C_v - T_j - T_k + \sum_{-m}^{+m} 2(m-1)}{T_j + T_k} \quad \dots \text{eqn.5}$$

It is necessary to substitute C_v for C_a in eqn.5 to avoid a discontinuity in possible values of I_D which arises as a result of the expression being a distance measure and not a measure of association. This can be easily demonstrated by working examples with alternately C_a and C_v in eqn.5.

The relationship between I_D and existing distance measures can be demonstrated by substituting symbols more conventionally associated with distance metrics.

T_j = total range over all characters of character states of OTU_j. This can be rewritten as $\sum_{i=1}^n [X_{ij(\max)} - X_{ij(\min)}]$

Substituting for T_j and T_k in eqn.5

$$I_D = \frac{2C_v - \sum_{i=1}^n (X_{ij(\max)} - X_{ij(\min)} + 1) - \sum_{i=1}^n (X_{ik(\max)} - X_{ik(\min)} + 1)}{\sum_{i=1}^n (X_{ij(\max)} - X_{ij(\min)} + 1)} + \frac{\sum_{i=1}^n 2(m-1)}{\sum_{i=1}^n (X_{ik(\max)} - X_{ik(\min)} + 1)} \quad \dots \text{eqn.6}$$

To distinguish this new form, I_D will be replaced by D_I . Simplifying eqn.6 we obtain

$$D_I = \frac{\sum_{i=1}^n 2 - (X_{ij(\max)} - X_{ij(\min)} + 1) - (X_{ik(\max)} - X_{ik(\min)} + 1) + 2(m-1)}{\sum_{i=1}^n (X_{ij(\max)} - X_{ij(\min)} + 1) + (X_{ik(\max)} - X_{ik(\min)} + 1)}$$

The top and bottom lines are summed independently to avoid inconsistency between values of eqn.5 and D_I if summation of fractions

is performed. D_I is a Euclidean distance measure which expresses the location of character means allowing for ranges of values within characters. The range of D_I is 0 to $-(\text{Sum of character ranges})/\text{no. of characters}$. A value of zero, concomitant with other distance measures, implies that two OTUs do not differ at all, both in location of the mean values for each character and in their character ranges. The maximum distance value is obtained when two OTUs exhibit no ranges within characters and lie at opposite ends of the overall character range for each character. Where metric characters are divided into a number of class intervals (k) the range of D_I values is 0 to $-(k-1)$. D_I is related to a class of metrics mentioned previously - the Minkowski metrics. A special form of the Manhattan metric has been described by Lance and Williams (1967). This is known as the Canberra metric, defined as

$$d_{\text{CANB.}}(j,k) = \sum_{i=1}^n \left(\frac{|x_{ij} - x_{ik}|}{(x_{ij} + x_{ik})} \right)$$

A modification of the Canberra metric has been suggested by Gower in Sneath and Sokal (1973, p.126) which replaces the denominator by $|x_{ij}| + |x_{ik}|$.

The relationship between several metrics can be illustrated by comparing computed distance values for an example of two OTUs.

Character states	1	2	3	4	5	6	7	8
OTU j	+	-	-	-	-	-	-	-
OTU k	-	-	-	-	-	-	-	+

Mean character difference (M.C.D.) = 7

Taxonomic distance d_{jk} = 7

Ranged metric d_{jk} = 1

Studentized distance d_{kv} = 1

(Note. Both the preceding two metrics are scaled to the total

variation of the sample of OTUs and so are hardly meaningful in an isolated two OTU case).

Canberra metric $d_{CANB.}(j,k) = 7$

Multistate distance $D_I = -7$

D_I produces values equivalent to those of existing, well tried metrics but has the advantage of applicability to characters scored for multiple states.

Class intervals and metric characters

The procedure of dividing metric character ranges into a fixed number of intervals is recommended for a number of reasons. Firstly the number of class intervals and their size, although a function of the range of a character, is less sensitive to fluctuations of range than a character standardised to unit range. This latter procedure is widely followed in numerical studies (Sneath & Sokal, 1973). The stability of classed data on fluctuations of the range is important in studies where new OTUs are added after the initial standardisation of data. Where data is unitised there is a greater chance of re-scaling being necessary than when class intervals are chosen. This relates to the fact that the number of intervals can be chosen on the basis of the sample size. Secondly, in reality metric variables can never be continuous, as a consequence of the limits of measuring equipment. Thus, if a measurement is subject to $\pm 1\%$ error, there is little justification for dividing the range of the measured character into more than 50 intervals. This initial doubling of the class width is also the optimum in terms of reducing computational loads as compared to information loss. The unreliability of measurements are further compounded if the values used in an analysis are held to be sample based estimates of population parameters, e.g. means. This problem is simply the sample size and variance dependent

confidence one can place on estimates of measures of location.

A method of estimating optimal class intervals for a given sample size, with particular reference to frequency distributions was described by Sturges (1926). Sturges' method has been used in this study for computing the number of class intervals for metric characters. The expression used is

$$k = 1 + 3.3 \log n$$

where k is the number of class intervals and n is the number of values to be grouped, in this case OTUs. Figure 3 illustrates the relationship between sample size and k for values of n to 100,000.

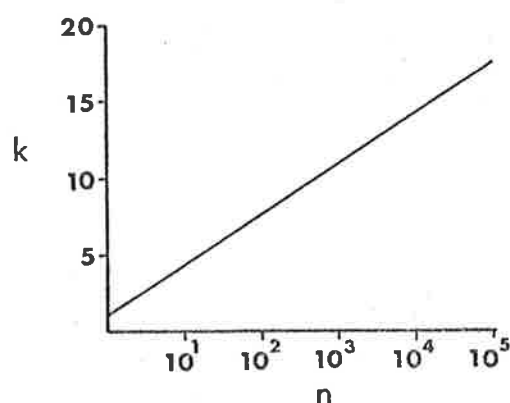


FIGURE 3. Number of class intervals as a function of sample size.

A further reason for converting to interval measurement is that Euclidean distance measures do not allow for the underlying distribution of the data. The result of this is that where data are not randomly distributed, the variance of distance values tends to be low. This has been shown empirically to be true in chapter 7. Outlying values are effectively weighted against values close to the sample mean and the discriminatory power of the distance function is greatly reduced. Using a number of class intervals smaller than

Range
2(% measurement accuracy)

allows the data to be forced into something approaching a random distribution. The randomisation of a normal distribution by adjusting class intervals is illustrated in figure 4. If the data is normally distributed class limits required to achieve an equal frequency in each class can be calculated by the following method.

1. Compute n , $\bar{x}$ and $\bar{X}$ for the set of OTUs.
2. Set the no. of class intervals using $k=1+3.3\log n$.
3. Calculate n/k as the desired frequency in each class.

Since the area to the mean from the extreme class limits is $\pm .5$, the area to the mean from each class limit is $.5, .5-(1/k), .5-(2/k), \dots .5-(k-1)/k, .5-(k/k)$.

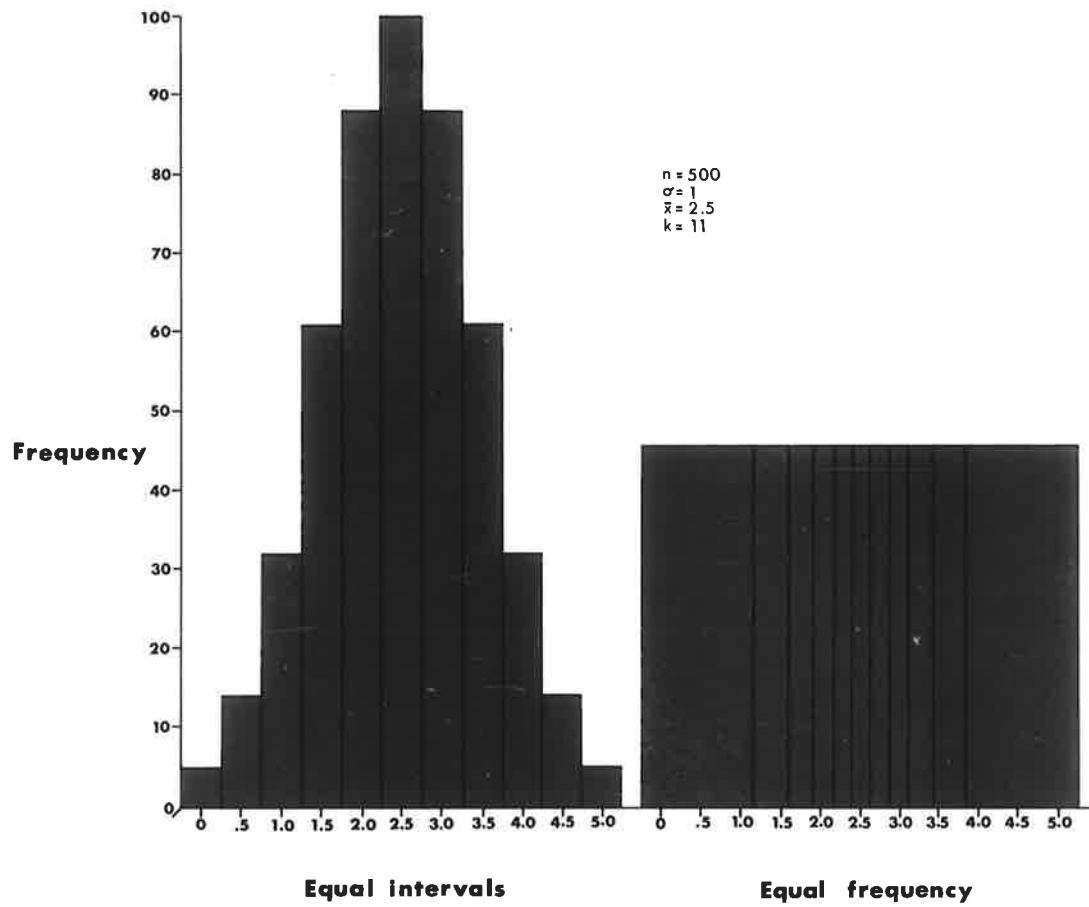
4. Evaluate area to the mean for each class limit.
5. Read values of the ratio Z for each class limit, from a table of areas of the normal curve. Since $Z = \frac{X - \bar{X}}{\sigma_x}$ then

$$X = \bar{X} + \sigma_x Z.$$
6. Compute values of X to give class limits.

Most metric taxonomic data seems to fit a normal distribution more or less closely. Where data fits the binomial distribution the assumption of normality for the purpose of computing class limits will generally suffice, even for small sample sizes. Where the sample is small it may be easier to determine equal area classes by trial and error using a histogram than to bother with the method described above.

Where the distribution is skew rescaling the axis by a non-linear conversion factor such as $\ln X$, $\sqrt{x}$, x^r , will generally allow symmetry to be restored. The rescaled values can then be treated by the assumption of normality and applying the above method.

FIGURE 4. Equal frequency intervals generated from a normal distribution by the method described in the text. The normal distribution is defined by the statistics shown. The number of class intervals (k) for the sample (n) is 11.



Influences on the accuracy of D_I as an estimator of Euclidean distance

The distribution of D_I values is dependent on the underlying distribution of individual characters. The ideal situation for the estimation of Euclidean distance between OTUs is when the data fits a random or equal frequency distribution. This is achieved, as described above, by shifting class limits of metric characters to obtain equal class frequencies. Where class frequencies are not equal, as in the case of a centre weighted distribution, failure to rescale to randomness will mean that Euclidean distances between OTUs will be underestimated. Conversely where a distribution is U-shaped or bimodal, distances will generally be overestimated. For metric characters rescaling is not a problem, but in the case of ordered, qualitative characters rescaling cannot be carried out in the same manner. Where a large number of states are scored, minimal information will be lost if a small number of states with low score frequencies are incorporated into adjacent states to try and approach randomness. Such characters are generally found to be pseudo-normal and so this additive procedure is generally acceptable. It may also be possible to express ordered qualitative characters as metric characters. For example, leaf shape can be expressed either by terminology or in terms of parameters such as length to width ratio, position of axis of maximum width, apex and base angles.

A further problem with ordering qualitative characters is that for a distance metric to be accurate in its estimation of distance, the character states must obey the rule of additivity i.e. $a-b = b-c = \frac{1}{2}(a-c)$. Where this does not apply, every effort should be made to convert to metric characters. Failing this, there are strong grounds for considering each character state as a separate two state character. Ordered qualitative characters are further complicated when

they are not unidirectional but branched. An example of this is where shapes are expressed not only as absolute shapes such as ovate, elliptic, etc., but are qualified by terms such as narrow and broad. Such characters can be termed branched or matrix characters. Figure 5 illustrates a dichotomised series of shapes.

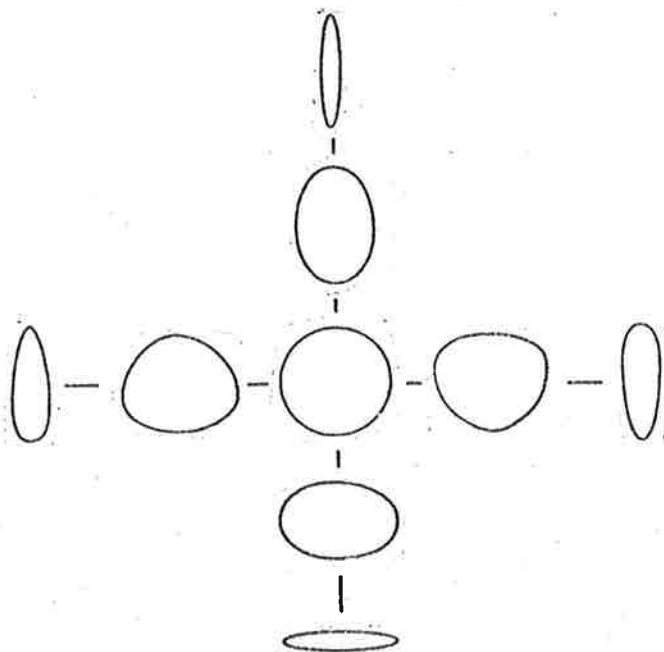


FIGURE 5. Branched or matrix characters illustrated by a series of standardised leaf shapes (after Systematics Association, Descriptive Terminology, Simple plane shapes, 1962).

Clearly a distance metric is inapplicable in this situation. These characters should be divided into a set of separate characters, one for each ordered branch.

When no ordering of qualitative characters is possible, they should be treated as n separate characters where n is the number of states, or at least so that a mismatch is seen to be equivalent to

a mismatch in a binary character.

The use of mixed character sets does not affect the overall values of the multistate distance as long as care is taken in the correct use of the alternate forms I_D and I_{diss} . A worked example will illustrate this.

$$\begin{array}{ccc} \text{OTU } j & + & + \\ \text{OTU } k & + & - \end{array}$$

Using I_{diss} , only positive values of m are allowed viz.

$$I_{diss} = \frac{2C_a - T_j - T_k + \sum_{m=1}^m 2(m-1)C_m}{T_j + T_k} = \frac{2-2-1+0}{2+1} = \frac{-1}{3}$$

Using I_D , character two of the OTU j /OTU k comparison can be considered as a 0 state mismatch.

$$I_D = \frac{4-2-1-2}{2+1} = \frac{-1}{3}$$

This example illustrates a problem caused by binary characters. That is, a mismatch in binary characters is treated as a zero state mismatch in multistate characters e.g.

$$\begin{array}{cc} + & - \\ - & + \end{array} \quad I_D = \frac{2-1-1-2}{1+1} = \frac{-2}{2}$$

$$\begin{array}{cc} + & \\ - & \end{array} \quad I_D = \frac{2-1-2}{1} = \frac{-1}{1}$$

Binary characters are therefore best excluded from character sets. A relatively small number of binary characters will hardly affect distance values, but conversely their exclusion will discard little information. Non-ordered multistate characters must also be handled carefully when included with metric characters. Apart from the problems of achieving equal frequencies, mismatches must not be computed i.e. m must equal 1. A mismatch in a non-ordered qualitative character is not equivalent to a mismatch in a metric character, but is equivalent to a binary character mismatch. The correct treatment

for metric and qualitative characters is illustrated below.

Character 1 is metric, character 2 is non-metric.

	Char.1	Char.2
OTU j	1-----	+-----
OTU k	-----6	-----+
Char.1	$D_I = \frac{2 - 1 - 1 - 10}{1 + 1} = \frac{-10}{2}$	
Char.2	$I_D = \frac{0 - 1 - 1}{1 + 1} = \frac{-2}{2}$	

For char.2, $2(m-1)$ is computed only for $m-1$ 0. These are then added as separate top and bottom lines (e.g. $D_I + I_D = \frac{-12}{4}$) to compute overall similarity.

3. Correlation coefficients

The third class of resemblance measures is the correlation coefficient. These coefficients are generally non-metric functions. The widely used Pearson product-moment correlation coefficient is based on the product of the distance of character scores from the mean and is defined as

$$\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}$$

This metric has been shown to be particularly applicable to studies based on linear dimensions of morphological features. This coefficient however is not essentially different from the multistate distance (D_I) in expressing resemblance. Eades (1965) and Minkoff (1965) discussed weaknesses of correlation coefficients in that they require all characters to have the same dimensions and directions.

4. Probabilistic similarity indices

The final class of coefficients is the probabilistic similarity

measure. The best known of these is the index of Goodall (1966). This method uses the observed frequencies of scores for each character state over a sample of OTUs to determine the cumulative probability that a pair of OTUs will be as similar or more similar than one would expect from the combination of characters. A feature of Goodall's method is that it weights rarer states over more common ones. The question of the justification of this is an important one and is amply discussed in Sneath and Sokal (1973). The equal prior weighting of character states is supported in this study and is the philosophy behind equalisation of class frequencies.

Probabilities are most useful when used to determine the significance of a comparison between OTUs. They cannot be justifiably used as a means of ranking OTU pairs by similarity. The reason for this is that once a set of OTUs is shown to be clustered, the null hypothesis of randomness of character scores is violated. As clusters are demonstrated therefore, probabilities for each cluster should be recalculated. A technical problem associated with computing probabilities is the high computational load, especially for larger studies.

An approach to the determination of probabilities for similarity values derived from multistate distance has therefore been investigated. The approach begins with the assumption that the probability of an OTU possessing a particular character state is randomly determined. This requirement relates back to the need for equal class frequencies discussed previously. On the assumption that character states are independent, the probability of a particular match of states in one character between two OTUs is expressed by the binomial probability function.

$$P(N=n) = \binom{s}{n} p^n q^{s-n}$$

where s equals the number of states of the character and n the number of matched states. If only positive matches are considered significant the probability of any particular pair of events (match) in the two trials (OTUs) is given by

$$\begin{aligned}
 P(N=n) &= \binom{s}{n} .25^n .75^{s-n} = \binom{s}{n} \left(\frac{1}{4}\right)^n \left(\frac{3}{4}\right)^{s-n} \\
 &= \binom{s}{n} \frac{3^{s-n}}{4^s} \\
 &= \frac{3^{s-n} s!}{2^{2s} n!(s-n)!} \quad \dots \dots \text{eqn.7}
 \end{aligned}$$

A special case of this, when the value of s in the two trials differs is given in appendix 2. Clearly in the case of metric characters the assumption of independence of states is unlikely to hold. Non-ordered qualitative characters may well meet this condition. In the case of this happening each character state can be considered as a separate character, so that s now equals the number of characters possessed by each OTU. In the derivation of eqn.7 above, p and q were substituted by the values .25 and .75 respectively. The reason for this is that only one of the four possible outcomes of pairs of events in which each event can be considered to have two states (+, - or presence, absence) is of importance as a match. The possible outcomes are ++, +-, -+, -- where only ++ is a match. The relationship between this probability, here called P_n^s and the dissimilarity index I_D have been computed for a range of values of s and n .

Figure 96 gives the probability of particular matches of two OTUs. Of greater application however is the plot of cumulative probabilities which gives the probability of finding a pair of OTUs matched less than the pair in question (fig. 97). Like the probabilistic similarity index of Goodall such an approach is likely to find its greatest application in ecological studies where species

incidences are being compared on the basis of random sampling of vegetations. If the assumption of randomness can be made then the probability of a pair of OTUs scored on two state character being more or less similar than the pair in question can be read from the figures. An interesting feature to note is the relationship between number of characters and the number of matched states required to meet a particular critical probability value. Thus the critical level of 5% is achieved with 3 matches over 4 characters, 6 matches over 10 characters, 9 out of 20 characters and only 19 out of 50 characters. Clearly this makes nonsense of any general statement regarding the need to achieve a particular percentage similarity before two OTUs are considered to be the same taxon.

Binomial probabilities could be applied to multistate characters if some guarantee of the independence of character states could be made. In the case of metric characters this would admit the possibility of discontinuous ranges being scored for a set of individuals making up an OTU. To the binomial probability model a match of three states is equivalent in either of the following cases.

OTU j	+---+---+		OTU j	---++++-
		or		
OTU k	+---+---+		OTU k	---++++-

A discontinuous range can arise either because the sample size is not large enough to mask random influences or the OTU is not a homogeneous sample. The general binomial model which admits all possible combinations of states is inapplicable when OTUs are scored on a single state or are free to range over several states. The simplest approach to accomodating these possibilities is to use a sample based estimate of probabilities of matches. Where OTUs are represented by a single value for each character and the frequency of OTUs is equal over all states (s) the probability (p_i) of two

OTUs being matched on a character is $1/s$. The summed probabilities over all characters (n) are given by Fisher's approximation to the combination of independent probabilities, expressed as

$$-2 \sum_{i=1}^n p_i$$

which is distributed as χ^2 with $2n$ degrees of freedom (Fisher, 1963). The probability values for this statistic may be read from a table of χ^2 . Where the frequency of OTUs is not equal across character states the probability of a match of OTUs on one character is

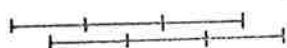
$$p_i = \sum_{i=1}^s \left[\frac{f_i}{\sum_{i=1}^s f_i} \right]$$

When OTUs are scored for more than one state in a character the probability of a match is given by

$$p_i = \frac{f_i}{2 \sum_{i=1}^s f_i - \left(\sum_{j=1}^{s_j} f_j + \sum_{k=1}^{s_k} f_k \right)}$$

where s_j , s_k , f_j and f_k are the number of states and frequencies in each state of OTUs j and k respectively.

An important reason for considering the number of states that OTUs possess in each character is that when character ranges are divided into discrete classes spurious similarities can result. For example, an OTU scoring close to a class limit may be more similar to another OTU in an adjacent class than one in the same class. This can be overcome without dispensing of class limits by scoring borderline values in both classes. This is equivalent to having overlapping classes. The error in assignment to a class is minimised when class limits lie at the mean value of adjacent classes viz.



(for six classes)

Conditional binomial probability

Modelling probabilities on an assumed distribution has been considered above with regard to applying the binomial distribution to measuring resemblance. The simple binomial model discussed is inadequate where discontinuous ranges are inadmissible. If the data is defined by a rigorous set of criteria, expressions for conditional binomial probability can be derived. On the assumption that the frequencies of OTUs are equal over all classes the following three overlapping situations are defined.

1. OTUs can possess a fixed number of states i.e. where a character is divided into s states an OTU can have s_1 states where $1 \leq s_1 \leq s$.
2. OTUs may range over any number of states such that the range of s_1 is 1 to s , 2 to $(s-1)$, 3 to $(s-2)$, ..., $s/2$.

These situations are clearly intersecting and the expressions for each are given in appendix 1.

The computational loads are very great in the application of the conditional binomial distribution to the problem of estimating resemblance. Because of unacceptable inadequacies in all the probabilistic methods, the measurement of similarity between OTUs has been accomplished in this study primarily by the multistate distance measure.

CHAPTER 3

THE CHOICE OF A LEAF CHARACTER SET

The single most important decision in a study of this kind, after the initial definition of the project, is the choice of a final character set to be committed to whatever analyses are appropriate. This choice governs the path of all other aspects of the study and so is here considered in the necessary detail.

The initial phase of the research program on Saurauia was to carry out a detailed survey of architectural features of the leaves of a sample of species. The data derived from this aspect were to then be applied to generating a leaf-based classification of the species.

Types of leaf characters

A number of character sets for describing leaves have been used in the past. The earliest set of significance to this study was that of von Ettingshausen (1861) who prepared a terminology for the description of vein patterns in leaves. Von Ettingshausen's terms have found general acceptance by taxonomists of extant plants and palaeobotanists and are maintained in recent publications such as those of Hickey (1973), Dilcher (1974), Hickey and Wolfe (1975) and Dolph (1976). The most significant of these recent usages has been that of Hickey who collected and defined a set of terms describing the range of architectural features of angiosperm leaves. Hickey's terminology is based on preexisting systems including that of von Ettingshausen and of Turrill in Lee (1948), later re-described by Stearn (1956) who expanded the system somewhat. An independent system of description using metric features was that of Mouton (1966). Mouton's system will be discussed later in further detail, but suffice

it to say here that it is very incomplete when compared to the Hickey system. A recent publication of a vein pattern terminology was that of Melville (1976) who in a misdirected sense of etymological propriety renamed the existing terms, defined some new subdivisions of these and produced an unwieldy system which will not be further considered in this study.

Other systems such as those used by Featherly (1965), Madler and Straus (1971) and Mildenhall (1973) are not important except that Madler and Straus included features of the leaf indument and epidermis.

Features of leaf indument and epidermis have also been collated into descriptive terminologies. The systems of van Cotthem (1970, 1971) and Stace (1965) are important in defining the range of features of epidermal cells, in the former case particularly of the stomatal complex. Dilcher (1974) advocates the use of these systems coupled with Hickey's terminology.

There has been a new awareness in recent times of the value of leaf characters in taxonomy brought about by the growing interest in fossil angiosperm floras. Leaves have been long neglected because of their variability and the inability of taxonomists of extant plants to either handle this variability or to see the constancy of many characters previously poorly defined. Criticisms of their use by Hamilton (1916) and Odell (1932) have been based on a few specific examples of variability of characters. Such critics have noted the easily observable variability in features of leaf size, dentition form, leaf texture and indument density, but have failed to appreciate that many features of venation and epidermis are constant and diagnostic. Numerical approaches to taxonomy have also meant that even variable characters contain some recoverable information. The question

of handling variability of leaves has been considered by Gardiner and Jeffers (1962), Dancik and Barnes (1974) and Dolph (1976(a,b)). The most recent of these has demonstrated the inadequacy of existing character sets in accomodating variability and the high degree of misidentifications that may result from using characters whose inter-relationships are poorly understood. A physiological approach to explaining the diversity of leaves within a species has been taken by Lewis (1972), but this study considered only those features such as shape, size and texture, the environmental dependence of which was noted by Elliot (1891) and used by Webb (1959) to produce a physiogomic classification of rainforests.

The set of characters under study

Although it was intended in this study to restrict the character set to only features of shape, size, margin type, primary, secondary and tertiary venation for reasons discussed in the introduction, it was felt worthwhile initially to investigate a large range of characters. The ultimate choice of a character set was dictated by a number of factors. These can be summarised as information content, relevance to the aim of the study, ease of measurement and universal applicability to the set of species.

Initially a survey of about thirty species represented by pressed herbarium material was undertaken to assess the availability of characters and to prepare a list of characters for more detailed analysis. Eighty characters of venation, indument and epidermis were defined in this initial phase. Of these, thirty-three were ordered or non-ordered qualitative, two were binary and the remainder purely metric characters. The epidermal and indument features did not differ greatly from those of Dilcher (1974) except in details of character states. The architectural features, however, were mostly metric

(27/40) and differed in many respects from those of Hickey or Dilcher.

The whole set of metric architectural characters will now be described. Bracketed letters refer to features illustrated on figure 6. Only those characters that differ from previous usage will be defined in detail. A number of metric characters are discussed within other characters, as they were not considered to be fundamentally distinct.

Architectural characters

1. Leaf length (AB).
2. Leaf width (CD).
3. Position of maximum width.

This character refers to the point (M) at which the broad axis of the leaf (CD) intersects the long axis (AB) and is expressed as a ratio of the distance from the leaf base to this point and the total leaf length i.e. $((BM/AB) \times 100)\%$.

4. Position of basal axis of 20% of maximum width.

This is the position of the intersection (P) of the long axis with the lowermost axis whose length is 20% of the maximum width of the leaf (EF) and is expressed as $((BP/AB) \times 100)\%$. This and the succeeding character are important in defining the base and apex of a leaf.

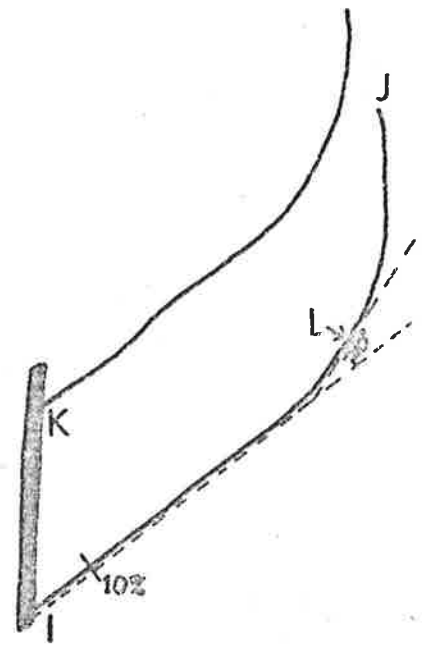
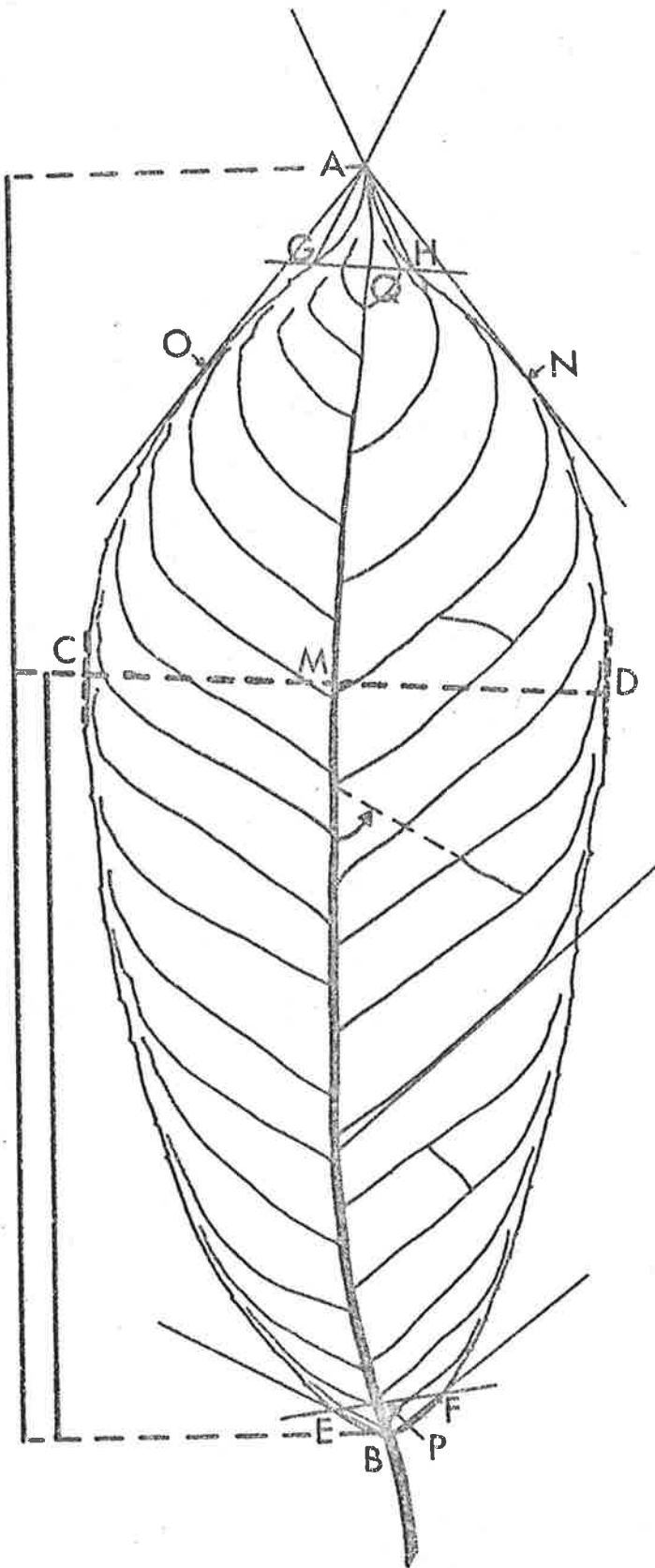
5. Position of apical axis of 20% of maximum width.

As for the preceding character, except that this locates the axis GH as $((AQ/AB) \times 100)\%$.

6. Dentition frequency.

This is the no. of teeth per cm. of leaf margin, computed as an average figure over about half the length of the margin. It may be useful in leaves where the dentition frequency

FIGURE 6. Metric leaf characters for
measuring dicotyledonous leaves.
Detail is of a pair of secondary
veins with relevant characters
shown. For explanation of letters
and characters refer to text.



10%

differs markedly along the margin e.g. leaves becoming entire margined towards the base, to compute dentition frequency separately for the upper and lower halves of the leaf. Likewise this measure can be modified to account for more than one series of dentition.

7. Secondary vein intercostal shape index.

This character is a measure of the shape delimited by the secondary veins and can be an important diagnostic character. It is defined as the ratio of the length of a secondary vein (IJ) and the distance between secondary veins measured along the midrib (IK). This measure is most useful computed as an average over all or a subset of secondary veins. Because this character often changes drastically towards the base and apex an arbitrary cutoff beyond which secondary veins are not measured should be chosen. Secondary veins arising from the midrib above the apical 20% width axis or below the basal 20% axis should not be measured for this or the succeeding two characters. Some studies may find that the variation of this measure along the leaf is a useful character. When measuring parameters of veins, it is essential to be able to define one vein order from the next highest. This is particularly true in the case of secondary veins. Where a secondary vein unambiguously terminates either at the margin (craspedodromous venation) or at the superadjacent secondary, forming a loop with it (brochidodromous venation) or where there is an intramarginal secondary vein formed by fusion of the loops, the location of the end of a secondary is unlikely to cause problems. When secondaries disappear gradually into succeeding higher order veins (eucamptodromous or festooned brochidodro-

mous venations), the exact point at which the secondary becomes a tertiary may be difficult to locate. I have defined, for the purposes of this study, the termination of a secondary vein as that vein junction at which the arising vein branches are of the same thickness and thinner than the previous segment of secondary, regardless of the thickness of other veins in the vicinity and of the course of the veins. Thus as soon as a secondary becomes indistinguishable in thickness from a vein arising from it is considered to be terminated and the branches to be tertiary veins. The criterion of the vein branches being thinner than the previous segment of secondary vein is to account for those cases where secondary veins give rise to branches which themselves are unquestionably secondary veins. In Saurauia such occurrences are infrequent and generally clearly aberrant. In such cases I have excluded the vein from measurement. The distinctness of vein orders has been tested on Saurauia and also on a species of Sterculia. These test cases are discussed in detail at the end of the character descriptions.

8. Secondary vein straightness index.

This character is a measure of the degree of curvature of secondary veins. As for the previous character it is easiest to average this over a set of veins, but the variation along the leaf may be a useful feature. The straightness of secondary vein is defined as follows. The position of the adaxial 10% length of a secondary vein is located. This is then used as the starting point for measuring straightness. The reason for choosing this rather than the point of origin of the secondary vein on the primary vein is to remove the effect of

locally high curvatures at the junction of vein orders. The tangent to this point is located. Taking tangents to the secondary vein by sliding a protractor out along the tangent to the 10% length, locate that point on the vein where the tangent to the vein first makes an angle of 20 with the 10% tangent. This angle is illustrated in figure 6 as the angle between the tangent to the point L and the tangent to the point 10% IJ. The length of vein between these points is then expressed as a ratio with the total length of the secondary vein i.e. (10%IJ to L/IJ).

9. Secondary vein angle.

Since the course of secondary veins is invariably more or less curved, it is necessary to define an arbitrary set of points on the vein, which can be used to express the angle between the vein and the midrib. The region of least curvature of a secondary vein is taken to be most representative of the direction of the vein. This character is therefore defined as the angle between the tangent to the point of least curvature of the vein and the tangent to the midrib at that point where the secondary vein tangent intersects it. Where vein courses are close to circular it is not possible to define a point of least curvature, therefore the angle measured is that between the line joining the ends of the secondary vein (IJ) and the tangent to the midrib at the intersection of this line with the midrib. A vein can be considered to be circular, to the extent that least curvature is undefinable, if it deviates from a fitted circle by more than 2% of the circle radius. A circle may be fitted to the three points represented by the midpoint and ends of the secondary

vein.

10. Number of secondary veins.

This is counted for all lateral veins distinguishable as of secondary rank, and averaging the value for the sets of veins on either side of the midrib.

11. Secondary vein length %.

This is the distance from the leaf base to the point of origin of each secondary vein on the midrib, expressed as a proportion of the total leaf length, beginning with the most basal veins.

12. Tertiary vein angle.

This is the angle between the projection of the line joining the ends of a tertiary vein and the midrib, measuring the adbasal angle. Where tertiary vein angles vary both apically and exmedially on a leaf, it may be useful to measure tertiary vein angle at several points on the leaf, defining the points in terms of % leaf length and % leaf width. The 3^o veins at 50% width and 30, 50 and 70% length are shown on figure 6.

13. Intersecondary vein length %.

The length of an intersecondary vein is measured from the midrib to that point where the intersecondary first becomes indistinguishable from a higher order vein. This length is then expressed as a ratio with the length of the superadjacent secondary vein.

14. Apex Angle 1.

The angle between tangents to the leaf margin passing through the leaf apex (OAN). This, and the following measurements, are severely affected if a portion of the apex is missing.

15. Apex Angle 2.

Angle between lines joining the apex to the intersection of the upper axis of 20% width with the margin (angle GAH).

16. 20% width angle.

This measure of the apex angle may be used where there is the possibility of part of the apex being missing. It is measured as the angle between the tangents to the points of intersection of the apical axis of 20% width with the margin.

17. Percentage acuminate.

This is defined as the ratio of the length of the apical portion of the leaf bound by concave sides to the total leaf length. The length of the apical portion is found by locating those points where the concave margin changes its direction of curvature. These points are joined by a straight line and the intersection of this line with the midrib defines the relevant length of apex. Percentage decurrent may be defined in a similar fashion.

18. Percentage cordate.

The distance between the leaf base and the intersection with the margin of the tangent to both lobes of the base is expressed relative to the leaf length.

19. Base Angle 1.

This is defined slightly differently to Apex Angle 1, in that where the base has concave sides e.g. decurrent, the tangents can be taken to the inside of the margin. The leaf base although less likely to be missing from a leaf than the apex, is in some respects harder to define. Where a leaf has a decurrent base or winged petiole, the position

of the base is that point where the total width of the leaf becomes equal to twice the width of the midrib at that point.

20. Base Angle 2 (EBF).

The angle between lines joining the intersections with the margin of the axis of 20% width (EF) and the base. Where this axis can intersect the margin more than once on each side as in the case of a strongly cordate base, the points relevant to measuring the base angle are those that represent the actual 20% width.

Further characters which have not been closely investigated in this study are given below.

21. Secondary vein angle at its origin.

This is the angle between the secondary and primary vein at the point where the secondary arises. It is useful in cases where the course of the secondary vein changes abruptly close to its origin.

22. Secondary vein curve.

An alternative to measuring vein straightness is to measure the curvature. A simple approach to this is to compare secondary veins to a standard set of curves and choose that curve which most closely resembles the vein course. A standard curve set can be easily generated by using envelope curves. The curves are defined in terms of the angles between the axes and the relative lengths of the axes. The major difficulty with applying this character is that given a large set of curves it is difficult to visually assign observed vein curves to any particular standard. The size of a standard curve set is propounded by the fact that many secondary veins have two regions of high curvature, one near

the base and the other where the vein bends inside the margin. This necessitates generating double envelopes. This can be overcome to some extent by considering the secondary veins on either side of the region of minimum curvature as separate curves. A difficulty with analysing data for this character is that it is not a simple directional metric like linear measurements. Like leaf shape it can be described as a matrix character and constitutes special coding difficulties in a multivariate analysis.

23. Leaf margin curve.

The description of leaf shape by expressing the curvature of the leaf margin in terms of a mathematical function may prove to be a useful character. Attempts to find curves for a number of leaves were made with a polynomial fitting computer program. The major difficulty with this character is that it contains all the characters from the list above that are associated with leaf size and shape and results in a decrease in the number of characters available for study. Secondly, it is, like secondary vein curve, a matrix or multidimensional character. Therefore although the potential information content is high and with the advent of computerised direct scanning systems it is a potentially easily measured character, the inadequacies of existing multivariate analyses mean that the recoverable information is less than the character set of its constituents.

24. Leaf shape index.

A method of expressing leaf shape using a standard grid has been investigated. To overcome the influence of size on a measure of shape it was decided to express leaf shape in

terms of angular coordinates on a radial grid. The requirements of a method of measuring leaf shape can be summarised as below.

A. To overcome the effect of missing base or apex, the measure should be standardised to leaf width.

B. It should be able to distinguish between leaves which are identical in shape except for their orientation e.g. ovate and obovate leaves should be distinctly different in their shape index.

C. Symmetry about the broad axis of the leaf should not perturb the index, i.e. it should not be a measure of difference in shape above and below the axis of maximum width. The effect of being sensitive to this would be to give similar index values to orbicular, elliptic or rhomboid leaves.

D. The index should be insensitive to missing data such as missing base or apex.

E. Leaf curvature should not influence the index, i.e. the shape should be measured such that the midrib is "straightened."

F. It should not be perturbed by multiple intersections of grid radii with the leaf margin. This is likely to happen in deeply lobed leaves.

G. It should be able to compare leaves with single marginal intersections of the radii and leaves with multiple intersections. The index that has been developed measures the distance from marginal intersections of grid radii to the midrib of the leaf. These measurements are made over a standard set of angular intervals with the centre of the

grid placed on the point of intersection of the axis of maximum width of the leaf with the midrib. The grid is oriented so that the 0° line parallels the midrib, at least at the grid centre. The leaf is measured in four quadrants at the angular spacing indicated in figure 31. The index is computed from the sum of distances (d) from marginal intersections to the midrib using the following expression

$$SI = \frac{200 \sum d}{NW}$$

where N is the number of radii intersecting the margin, and W is the leaf width. This produces a value of shape expressed as a percentage. The shape index applied to a species of Saurauia is discussed in chapter 5.

25. Petiole length percent.

This character is expressed as the length of the petiole divided by the length of the leaf, and converted to a percentage.

Characters of the leaf indument and epidermis

Although the features of indument and epidermis have not been incorporated into the analyses of leaf data, their inclusion in the leaf descriptions of Saurauia species, and the occurrence of some unusual cuticular patterns, warrants their description here.

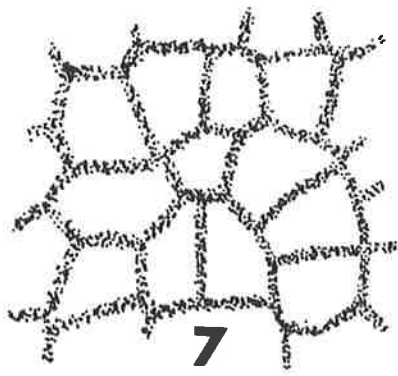
1. Upper cuticular surficial thickening.

This character defines the types of thickening found on the surface of Saurauia cuticles on the upper leaf surface. The character states are

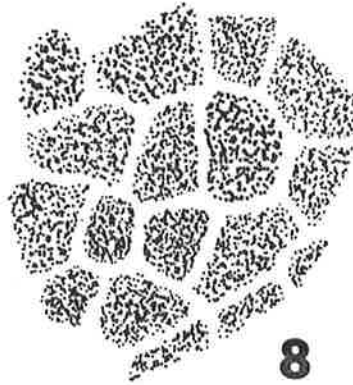
- a. Not thickened, or evenly granular, such that traces of epidermal cell walls are not visible on the cuticle.

- b. Mural. This name is given to the condition where thickening is prominent above the anticlinal walls of the epidermal cells so that they are easily visible as simple lines of thickening on the cuticle (fig.7).
 - c. Surficial. Thickening clearly visible over epidermal cells, but not above the anticlinal walls so that they are marked by regions of thin cuticle (fig.8).
 - d. Lensing. This is an extreme case of the previous state where the cuticle is very thick above epidermal cells and lacking over the walls except in the corners of the cells. The thin regions of cuticle above anticlinal walls are irregularly lens-shaped (fig.9).
 - e. Bordered surficial. Thickened only on the face of cells, not above cell walls but with a thickened distinct border around the region of thickening so that the thin cuticle over the anticlinal walls has a bordered or "tramline" appearance (fig.10).
 - f. Mural and surficial. Thickening above the cells lies in two distinct regions, above the cell walls and on the face of the cells, giving a three lined appearance. Generally the surficial thickening is bordered as well (fig.11).
 - g. Striate. Cuticular thickening in elongate ridges not transcurrent to cell walls. Sub-parallel above single cells but not parallel between adjacent cells and often with striae above and paralleling anticlinal walls (fig.12).
 - h. Ridged. Cuticular thickening in ridges transcurrent to cell walls. Care should be taken to differentiate between ridges and folds produced by drying of the leaf.
2. Upper cuticular granule size. The cuticular thickening generally

FIGURES 7-15. Features of the cuticle in the genus Saurauia. Figures 7-13 illustrate ornamentation on the surface of the cuticle (approx. X300). Figures 14 and 15 illustrate cuticular protrusions between the anticlinal walls of epidermal cells (approx. X1000). Further explanation is given in the text.



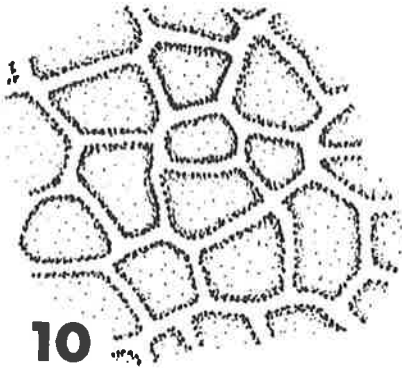
7



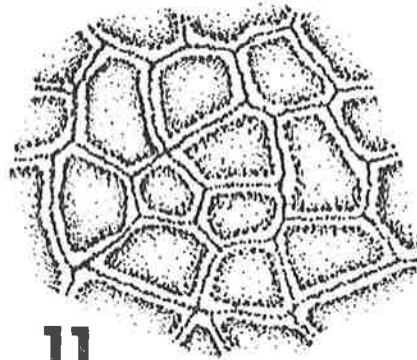
8



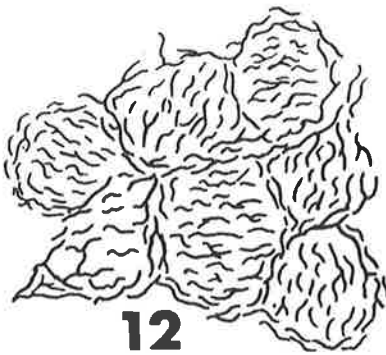
9



10



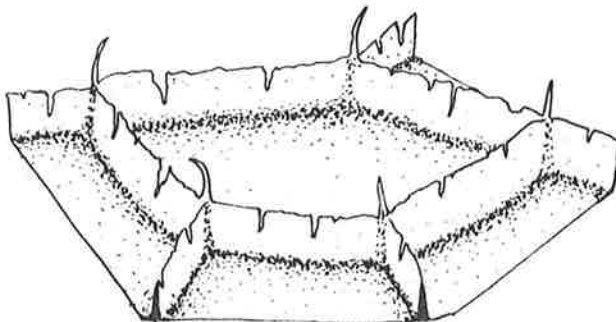
11



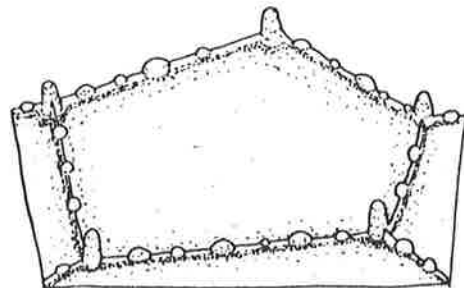
12



13



14



15

consists of distinct granules.

- a. Coarse. Larger than 2um in diameter.
- b. Medium. Diameter between 2 and .2um.
- c. Fine. Less than .2um.

3. Degree of granule contiguity. Granules may be distinct from each other or frequently touching and/or fused.

- a. Almost completely. Forming an almost continuous homogeneous sheet of thickening.
- b. Frequently. Granules fused into large irregular granules or short ridges, in up to 30% of cases (fig.13).
- c. Rarely or never. Granules free and distinct in greater than 95% of cases.

4. Upper cuticular anticlinal thickening. Cuticular thickening is often prominently developed in the corners of cells and intruding between the anticlinal walls of adjacent cells.

- a. Essentially none (normal).
- b. Bladed. Forming acute ridges, generally of even depth along their length but sometimes notched like a breadknife (fig.14).
- c. Beaded. Thickening in spheroidal clumps between anticlinal walls. Sometimes found on bladed thickening (fig.15).
- d. Peglike. Blunt rods of thickening extending between adjacent cells and only in the corners. Sometimes pegs are found with bladed thickening (fig.15).
- e. Spinose. Like pegs, but acute and often curved (fig.14). Occasionally mixed with peglike thickening.

5. Lower cuticular thickening.

- a. None. Outline of epidermal cells not visible. Cuticle very thin frequently with only the thickening of the outer stomatal ledge visible.

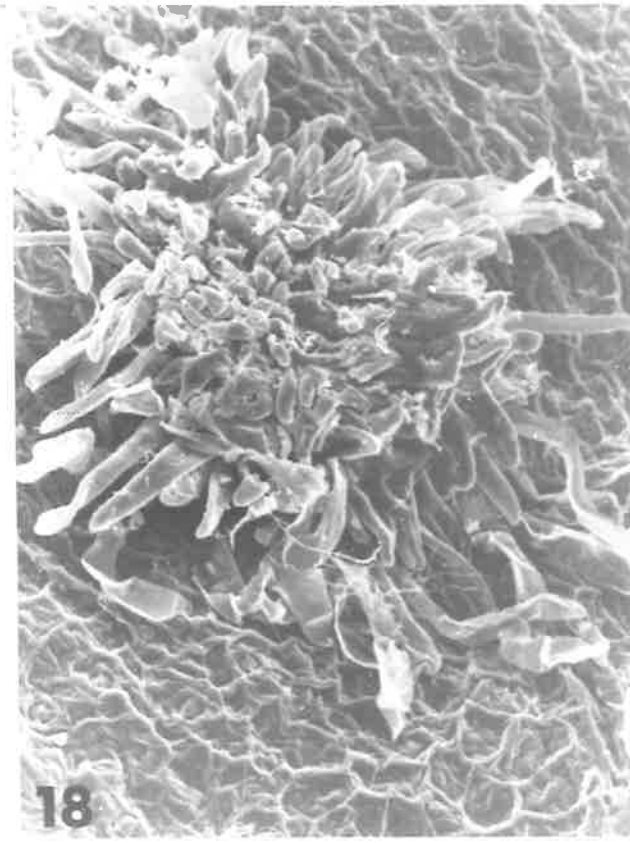
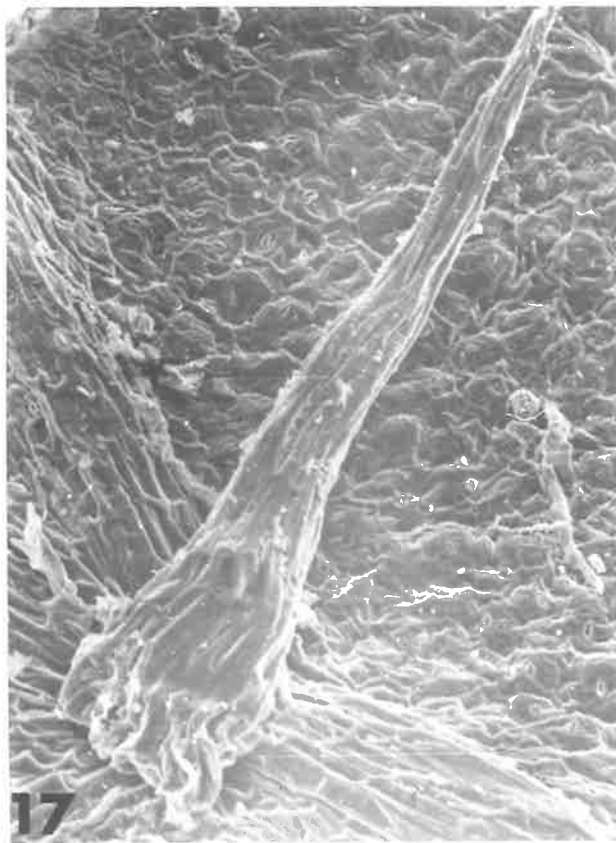
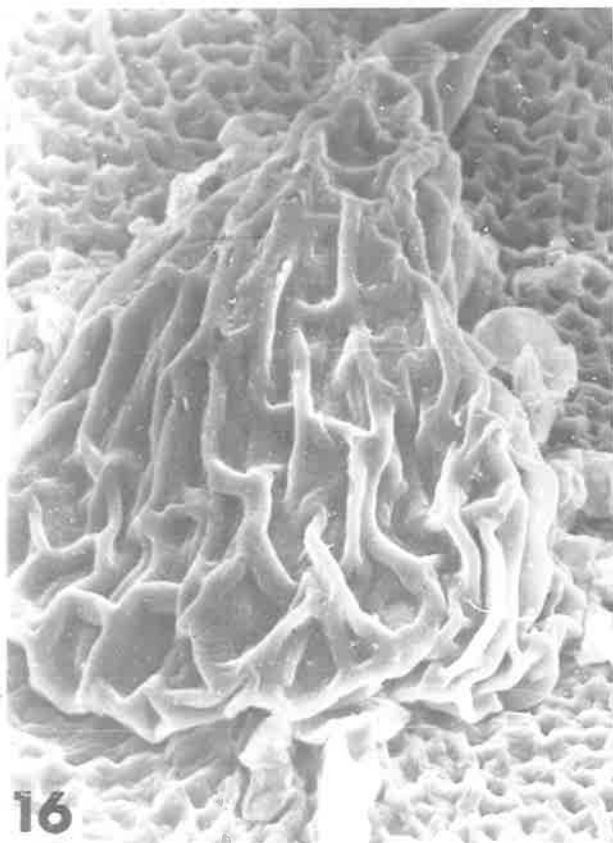
- b. Surficial. Cell arrangement visible due to thickening on the face of cells. Anticlinal walls of epidermal cells not expressed on the cuticle.
 - c. Mural or bordered surficial. Thickening as described for upper surface. Cell walls distinct.
6. Stomatal density. Number of stomates per sq. mm. It has not been possible to apply the commonly used Stomatal Index to Saurauia cuticles as they frequently do not have any expression of the epidermal cells.
7. Upper epidermal hair type.
- a. Uniseriate. Cells of trichome arranged in one series. This includes unicellular hairs. All further hair types are multiseriate.
 - b. Conical. Trichomes generally straight sided, obtuse or acute but never more than three times as long as they are wide. This class of hairs includes the thorns found in some species.
 - c. Strigose. Trichomes more acute than the previous type, up to five times as long as wide, generally bent at the base so they are somewhat appressed to the leaf surface (fig.16).
 - d. Globular. Trichomes as rounded often shaggy clumps, scabrid and not oriented like strigose hairs (fig.18).
 - e. Setose. Hairs acute, at least four times as long as wide. Overlapping in length to width with strigose hairs, but never appressed (fig.17).
 - f. None seen. Hairs absent from cuticular preparations, but possibly very rarely found on primary vein.
8. Lower epidermal hair type. Hairs as for upper surface except usually larger and with tufted hairs as a further type.

S.E.M's of hair types from species of Sauraula

FIGURE 16. Strigose hair on lower surface of
S.purgans leaf (x600).

FIGURE 17. Setose hair from lower surface of
S.bifida leaf (x150).

FIGURE 18. Globular hair on lower surface of
leaf of S.dufaurii (x200).



- a. Uniseriate.
 - b. Tufted. Numerous uniseriate hairs arising from adjacent or the same epidermal cells. Sometimes appearing stellate.
 - c. Conical.
 - d. Strigose.
 - e. Globular.
 - f. Setose.
9. Upper epidermal hair density. Number of hairs per sq. cm.
10. Lower epidermal hair density. As for upper hairs.

Differentiating vein orders on leaves

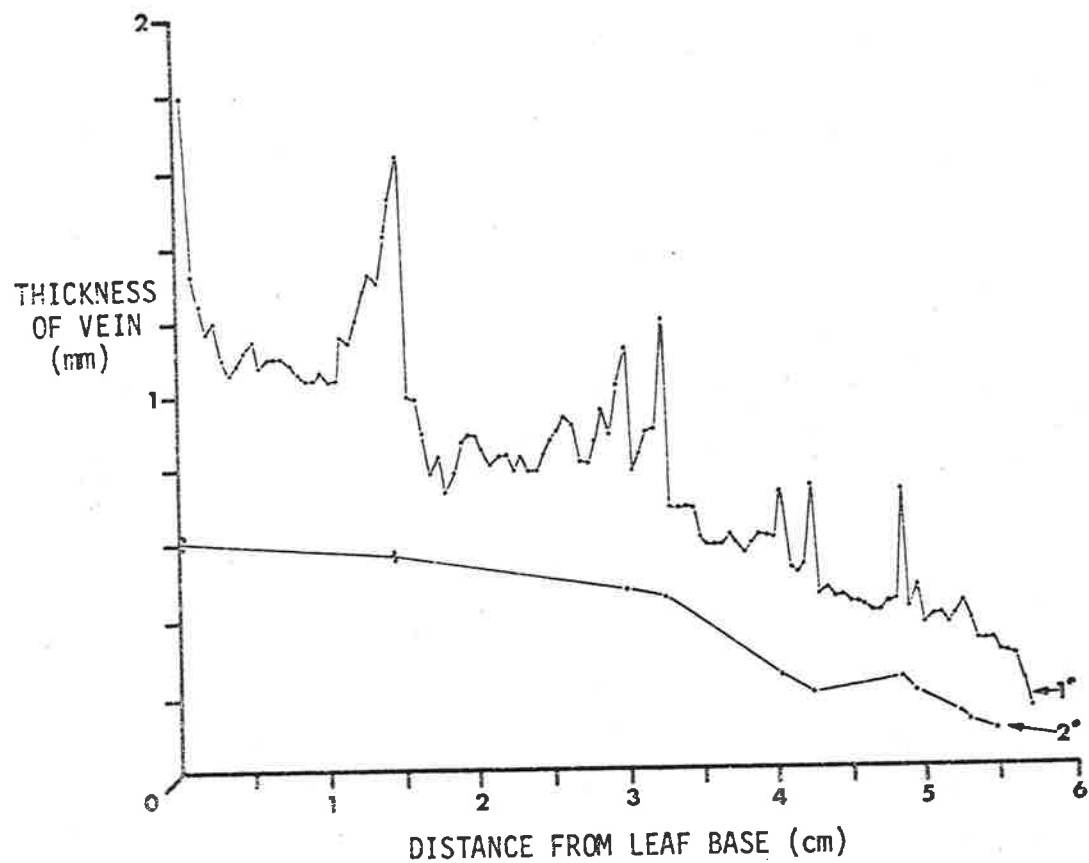
When talking of orders of veins, it is essential to be able to reliably and unambiguously demonstrate the differences between the different orders. At the level of measuring secondary veins for the characters described in this chapter it is important to be able to distinguish the end of secondary veins. The procedure used for making this distinction for the present study has already been outlined. In reality however, is this justifiable, and are there grounds for thinking that vein orders can really be distinguished? To answer these important questions it was decided to make some measurements of vein thicknesses at various points on some leaves to see whether the distribution of vein thicknesses exhibited a discontinuity corresponding to supposed different vein orders. To demonstrate the differences between primary and secondary vein orders a cleared leaf of Sterculia quadrifida (Sterculiaceae) was selected from the research collection of cleared leaves. This species was chosen because it was essential to use a leaf whose primary vein was not distorted by the clearing procedure. All cleared leaves of Saurauia had squashed or split midribs. Primary vein width was measured with an ocular micrometer grid in a binocular microscope, every 0.5 mm along the

vein. Wherever secondary veins branched from the midrib their position was noted, and their width 2 mm out along the vein was measured. Figure 19 illustrates the thickness of the primary vein plotted against length of the vein from the leaf base. Also plotted is the thickness of individual secondary veins against their position on the leaf. At all positions on the midrib, primary and secondary veins are distinctly different in thickness. The high peaks on the primary vein plot are due partly to the fact that the measurement of vein width was taken perpendicular to the primary vein and as a result include the secondary veins at their branching points. The three points on the secondary vein plot, closest to the leaf apex are doubtfully of secondary vein order. Their magnitude is intermediate with tertiary veins. This problem of ambiguous veins near the apex has been handled in the present study by not counting secondary veins which branch from the midrib above the point marked by the axis of 20% width (see character 5 in the previous section). The clear distinction between primary and secondary veins is also found on leaves of Saurauia, though it has not been quantified.

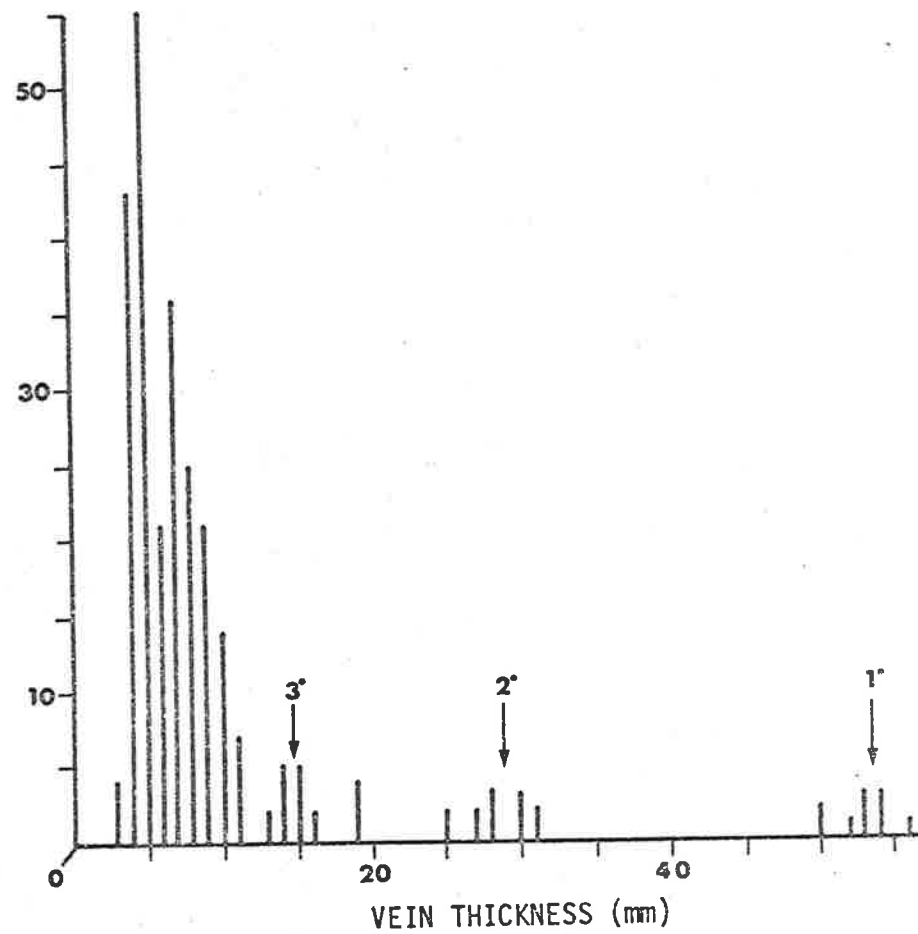
The greatest problems of ambiguity in defining vein orders arise with higher orders. To investigate whether secondary and higher order veins were distinct in Saurauia, a cleared leaf of S. calyptrata (fig. 28) was selected. The thicknesses of all veins contributing to vein junctions were measured from an 8.7 times photographic enlargement of the cleared leaf, using an ocular micrometer on a binocular microscope. A 1 cm wide transect was taken perpendicular to the midrib, at the point of maximum width of the leaf. The photo was a direct print made by inserting the cleared leaf slide in an enlarger. The negative print was made on Grade 4 Ilfospeed photographic paper to achieve high contrast. To prevent

FIGURE 19. Relative thicknesses of primary and secondary veins, measured along the length of the midrib on a leaf of Sterculia quadrifida. The vein orders are distinct at all positions on the leaf.

FIGURE 20. Frequency of vein thicknesses, measured on all veins contributing at vein junctions in a transect across a leaf of Saurauia calyptrata. Only the first three vein orders can be distinguished.



19



20

the possibility of different exposures over the print affecting the apparent thickness of veins, the print was carefully dodged to achieve as even a background as possible.

Frequencies of the different vein orders are plotted on figure 20. Veins of primary and secondary order are clearly distinguishable. The mode of tertiary veins is also distinguishable, though in close proximity to the large mode of higher order veins. Higher vein orders are not distinguishable. This is a result either of inaccuracy in measurement of the fine veins, or of a real lack of distinction in thickness of higher vein orders. The actual reason is of little consequence. The inability to differentiate higher vein orders on thickness, using the maximal resolution of a cleared leaf, means that their distinction in an architectural study is likely to be highly subjective. Vein course is important in defining the lower orders of veins, but its usefulness disappears at higher orders, where the veins are often more or less random. This finding supports the prior decision that the present study should concentrate on vein orders less than or equal to tertiary veins.

Relationship of the character set to existing terminologies

Mouton (1966) described some metric leaf characters for application to the description of fossil leaves. The characters of dentition frequency, number of secondary veins and percentage acuminate are similar to characters of Mouton. Mouton, however, defined the acuminate apex in terms of tangents to the region of reversal of curvature of the margin. Ferguson (1971) used a set of metric characters, defined by Melville (1960) which measure vein angle and secondary vein length in relation to the position of the veins along the midrib. The details of measuring these characters are essentially different from those used in this study and are

considered inadequate. Because the characters have been set up for application to Saurauia a number of features commonly included in leaf character sets have been neglected. These include such features as lobing, stipules and arrangement of leaflets. There seems no reason to doubt that the philosophy governing the choice of characters for this study is applicable to the study of angiosperms in general. A detailed study of higher order vein features has not been undertaken. The reason for this has already been mentioned in chapter 1 and is further illustrated in figure 20. A character set developed and tested in this study should be applicable to studies of fossil leaves in general.

Choice of the ultimate character set

A number of criteria governed the choice of the final character set to be used in detailed analysis of the whole set of species. The number of characters to be used in a numerical taxonomic study has been discussed by Blackith and Reyment (1971) and Sneath and Sokal (1973). Both advocate the use of about 50-60 characters to achieve a stable classification. A study of a single organ such as a leaf is unable to achieve as large a character set as a study of whole organisms. Indeed attempts to generate a large character set may end with largely redundant characters that contain little discriminatory information. Care should be taken to avoid creating a large set of new variables by converting pairs of existing characters into ratios. A small number of ratios may be acceptable, but there is enough uncertainty in understanding the statistical effects of such variables to avoid them wherever possible. The properties of ratios will be considered later in more detail.

The number of characters to be used in a detailed study of a homogeneous group such as a genus bears little relationship to the

number suitable for studying a mixed group of organisms. It may be that a heterogeneous group is easily discriminated in a key or phenogram by a small character set. However, if the real taxonomic relationships are in question a large character set may be necessary to avoid spurious similarities between unrelated taxa. Within a genus, real genetic limitations on the variability of characters are likely to ensure that a classification achieves stability with a reduced character set.

The size of a character set is also dependent on the nature of the characters. Firstly it is obvious that a continuous metric character contains potentially more information than a two state character. The potential information content of various types of characters and their suitability for numerical analyses are summarised in table 1. Suitability of a character for application of numerical tests is the case only when the procedure can be performed without distortion. Thus although it is technically possible to measure Euclidean distance between two OTUs on two state characters by recoding them to 0, 1, the computed value is not a strict distance. This will result in a distortion of Euclidean hyperspace, producing spurious results. Where two state characters are few in a character set this may not cause problems, but as a general principle, I would advocate not using such characters wherever possible. Distortion of distance will also result when the linear separation between states of an ordered multistate character are not equal. Thus the flower colours, red, orange and blue are not equally separated though ordered and so will give an incorrect measure of distance if treated in the same fashion as a metric character. A distance measure will be distorted by character state weighting, such as that brought about by unequal frequencies in states. Therefore discrete, hierarchic characters when scored for one state per OTU

Character type		Number of possible meaningful permuta- tions in a pair- wise comparison	Standard- isation	Similar- ity/Asso- ciation	Distance metric	Corre- lation coeff.	F test	PCA	Discrimin- ant analy- sis	Key gen- eration
Primary	Secondary division									
Binary	Alternate	2(4)	no	yes	no	no	no	no	no	dichoto- mous only
	Two state	4	no	yes	no	no	no	no	no	
Discrete	multistate									
	Non-hierarchic									
	Mutually exclusive states (n)	4	no	yes	no	no	no	no	no	yes
	Multiple values	up to 2^{2n}	no	yes	no	no	no	no	no	yes
	Hierarchic, one dimensional									
	Mutually exclusive states	n^2	no	yes	*	no	no	yes	yes	yes
	Non-exclusive	up to $\frac{1}{2}(n^2-n)^2$	no	only I_D	only D_I	no	no	no	no	yes
	Hierarchic, two dimensional (matrix)									
Mutually exclusive states	n^2	no	yes	no	no	no	no	no	yes	
Multiple values	up to $n^2(2n^2-1)$	no	only I_D	no	no	no	no	no	yes	
Metric	One dimensional									
	Mutually exclusive scores (r)	n^2	yes	yes	yes	yes	no	yes	yes	yes
	Multiple values (r)	up to $\frac{1}{2}(r^2-r)^2$	yes	only I_D	only D_I	++	yes	++	++	yes
	Two dimensional(matrix)									
	Mutually exclusive scores	n^2	yes	yes	no	no	no	no	no	yes
	Multiple values (r)	up to $n^2(2n^2-1)$	yes	only I_D	no	no	no	no	no	yes
	* distorted by unequal frequencies.									
	++ if reduced to mean value									

TABLE 1. Properties of characters.

should be avoided in such analyses unless the equal frequency criterion is approximated.

Matrix characters are an extension of the dichotomised character illustrated in figure 5. Leaf shape extended over the whole set of available states is an example of such a character. These characters are distinguished by the fact that it is possible to find a number of linear ordered series of character states within them. For existing analyses to handle such characters they should be separated into a set of simple unidirectional characters. A standard colour chart such as a Munsell soil chart is a further example of a matrix character where gradations of the character of colour occur in more than one direction. The segregation of such a character into a set of simple unidimensional characters must be decided on the basis of the character itself. No general rule can be applied except that it would be unreasonable to consider each discernible step e.g. in a colour series, as a separate binary alternate character. The effect of this would be to flood the character set with a large number of low scoring characters, so increasing the separation of OTUs. An example of a multiple value discrete multistate matrix character could be indument type, where all types of trichomes were included under one character and an OTU was able to possess more than one type.

It can be seen from table 1 that only one dimensional metric characters were applicable to most of the available tests. The only test on which the two conditions of this class of character i.e. mutually exclusive scores and multiple values, differed was on the F test. Clearly it is not possible to compute within OTU variance when the OTU is represented by single values for each character.

Of the characters described above, all the architectural

characters are of the one dimensional metric type and will be considered in the following chapter in terms of application to a set of species.

CHAPTER 4

PRACTICAL PROCEDURES OF DATA COLLECTION

Initially a small set of species was chosen to assess the character set. Criteria for judging the characters included (a) ease of scoring, (b) reliability in delimiting the species, (c) discriminatory power and (d) degree of redundancy with comparison to other characters. Criteria b, c and d, relate respectively to variance ratio, factor contribution and character correlation. These are discussed in chapter 5.

Because the study was based on herbarium specimens obtained from various institutions the whole set of species was acquired over the period of about a year. The initial survey was based on those species that were first available.

Practical approaches to data collection

Scoring of species was carried out on standard description forms (fig.21). Herbarium label data was appended, generally in the form of a photograph of each herbarium sheet. Wherever possible, an actual leaf was included in an envelope or alternatively a leaf rubbing made. Throughout the study working photographs of all herbarium sheets were taken, often including enlargements of particular leaves. Frequently material was not available for use outside the herbaria. In such cases the photographs served in place of actual specimens and measurements were made directly on them. The procedure of taking leaf rubbings was also extensively used when herbarium material was unavailable for loan.

Leaf rubbings are akin to von Ettingshausen's nature prints in that an image of the leaf is obtained directly from the leaf surface (von Ettingshausen 1861). The method of nature printing involved

FIGURE 21. Standard coding form for
scoring architectural
features from leaves of
Saurauia species.

SAURAUIA LEAF DESCRIPTION

Species:					Specimen No:				
Locality:					Habit etc.:				
Leaf No.	1	2	3	4	Leaf No.	1	2	3	4
Leaf Length					Leaf Width				
Posn. Max. Width					Length to Width				
Apex Angle					Base Angle				
Apex Type					Base Type and symmetry				
Leaf Shape					Margin type				
Teeth Type					2'Venation type				
2'Vein arrangement					No. of 2's				
2' Intercostal Shape Index; Leaf 1									
Leaf 2									
Leaf 3									
Leaf 4									
2'Vein Curve: Leaf 1					Leaf 2				
Leaf 3					Leaf 4				
2'Vein Straightness Index; Leaf 1									
Leaf 2									
Leaf 3									
Leaf 4									
2' Vein Angle; Leaf 1									
Leaf 2									
Leaf 3									
Leaf 4									
2'Vein Length %; Leaf 1									
Leaf 2									
Leaf 3									
Leaf 4									

3'Vein Angle to 2's;1				2				3				4			
3' Primary Angle Variation; 20% 2'Length				50% 2' Length				70% 2' Length							
Posn. on Leaf 30% 50% 70% 80%				30% 50% 70% 80%				30% 50% 70% 80%				30% 50% 70% 80%			
Leaf 1															
Leaf 2															
Leaf 3															
Leaf 4															
3' Vein Course and Pattern:															
3' Areole Shape;															
Higher Order Venation:															
Veinlets:															
Dentition Frequency (teeth/cm.)															
Teeth Size Range:															
Teeth Venation:															
Leaf Ornament, Upper Surface:															
Leaf Ornament, Lower Surface:															
Epidermis, Epidermal Cell Patterns and shape:															
Epidermal Cell dimensions															
Stomatal Frequency and Pattern, Upper Surface:															
Lower Surface															
Stomatal Size:															
Stomatal Type:															
Further Leaf Features:															

placing a leaf between a block of hard metal such as copper, and one of a soft metal such as lead. The blocks were then forced together in a press, so that the leaf veins were impressed into the lead. This was then used as a printing block to reproduce the venation patterns.

Leaf rubbings are a much simpler technique. The most satisfactory method of taking rubbings was found to be as follows. First, a sheet of thin easily transferred carbon paper is placed face up on a flattened leaf. Poorly flattened or broken leaves, especially when thin, should not be treated in this manner, as they may become damaged. Secondly, place a sheet of thin (airmail) paper over the carbon paper. Holding the paper down firmly with the thumb and forefinger of one hand, rub the paper with the back of the nail of the index finger of the other hand. If the paper is held down above the apex and base of the leaf, rubbing should be carried out parallel to the secondary veins, covering each intercostal area carefully before proceeding to the next. Thin paper will allow you to see the transferred image appearing so that it is possible to judge when sufficient detail has been recorded. Of prime importance is the need to prevent relative movement between paper and leaf during this procedure as it is almost impossible to realign in register.

The detail obtainable from leaves is dictated by the degree of vein expression on the leaf surface, the surface finish of the paper and the degree of slippage during rubbing. In general quaternary vein details have been consistently obtained from Saurauia leaves. This often exceeds the resolution obtainable with photographs of the leaf surface, especially in the case of very hairy leaves. Some examples of leaf rubbings from Saurauia and other leaves are illustrated in figures 22 to 25. An alternative to this method is

Leaf rubbings taken from several dicot. leaves.

The technique is described in the text.

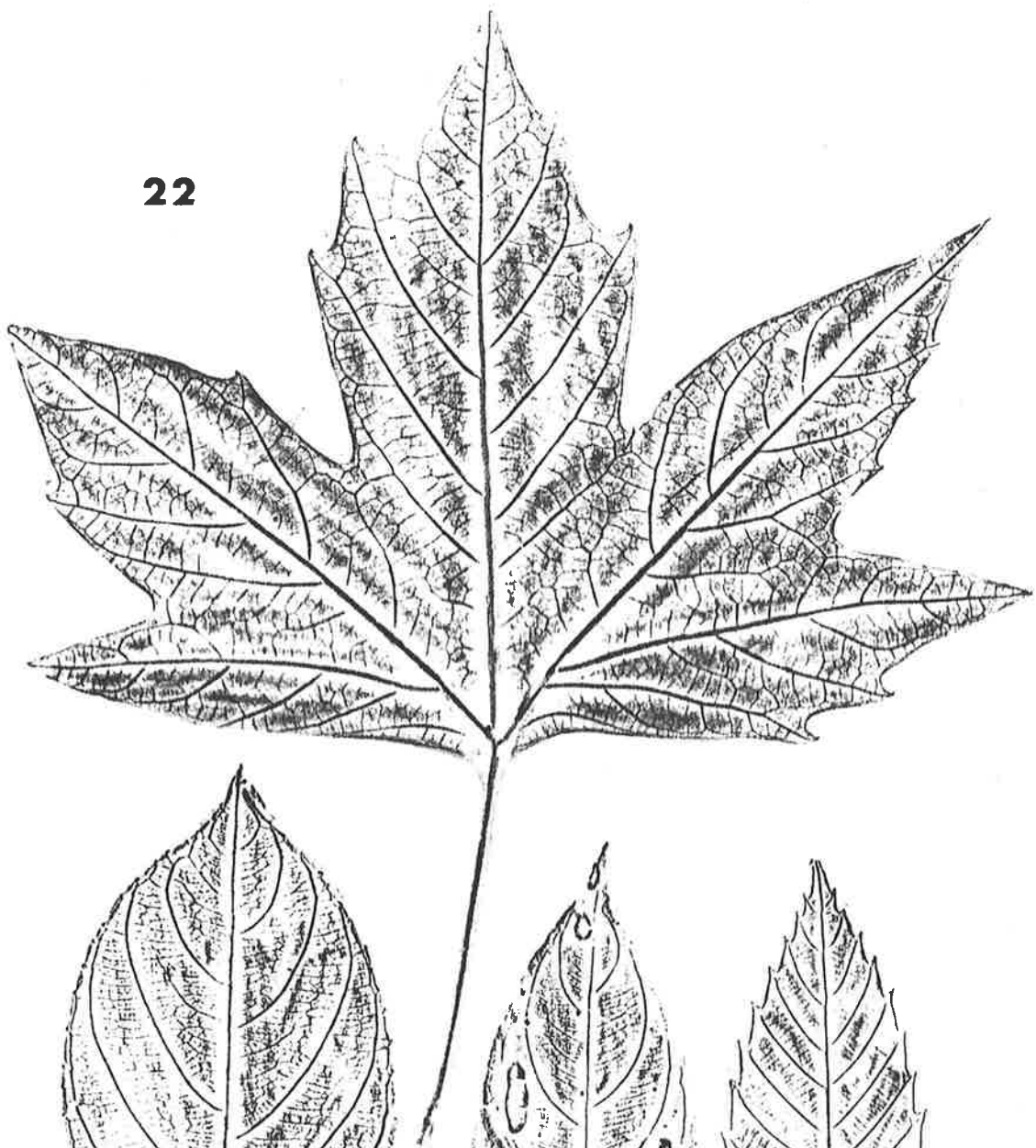
FIGURE 22. Leaf of Platanus occidentalis.

FIGURE 23. Saurauia pentapetala leaf.

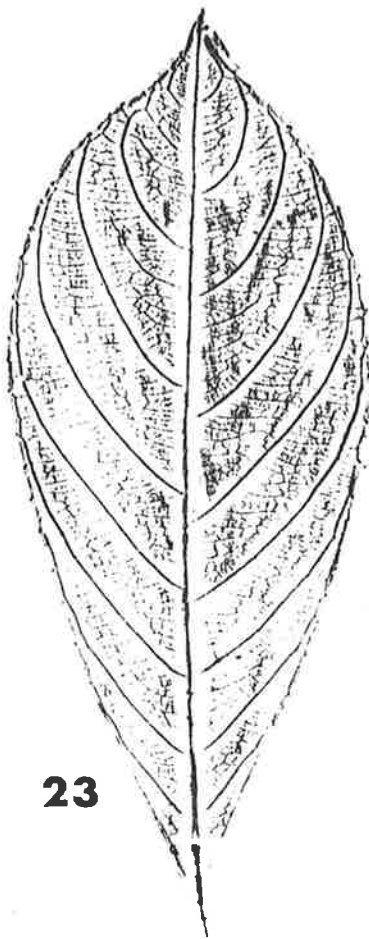
FIGURE 24. S. javanica leaf rubbing.

FIGURE 25. Rubbing from leaf of Quercus incarna.

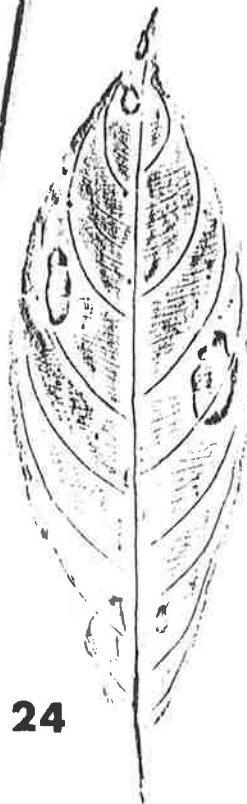
22



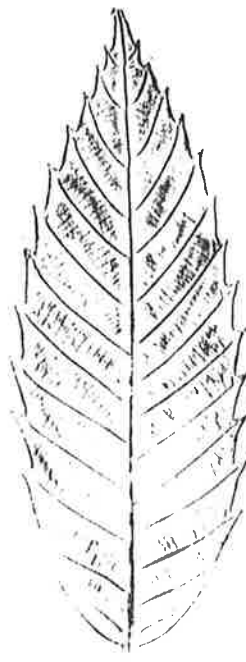
23



24



25



is to dispense with the carbon paper and simply rub the sheet of paper with a soft pencil to obtain a tracing of the leaf beneath. I have found that this method is much more likely to result in damage to the leaf and generally produces poorer contrast between the veins and mesophyll.

A further transfer technique which has been developed in this study is the use of cellulose acetate peels. This method involves the transfer of leaf features onto a thin sheet of cellulose acetate. Transfer techniques have been widely used in recovering fossil leaf remains from the matrix (Voight, 1936; Krausel, 1950; Koch, Friedrich, Christensen and Friis, 1973; Dilcher, 1974). They do not seem to have been used on whole extant leaves to any great extent, though their use in undergraduate botany courses for studying epidermal features is well known. Generally transfers are effected by painting on a transfer medium dissolved in a volatile solvent. Collodion in ether, or parlodion in amyl acetate are commonly used. Cellulose acetate has the advantage of being obtainable in sheets of various thickness. It is also soluble in acetone. Transfers of structures preserved in coal balls have been produced by covering a smoothed and etched surface with acetone, laying a sheet of thin acetate film onto the surface and allowing to dry (Kummel and Raup, 1965). A small quantity of acetate in solution penetrates the surface of the coal ball and dries. When the film is removed cellular remains on the surface are transferred. A similar technique can be applied to whole leaves as follows. A leaf is painted thinly with a mixture of 5 % paraffin wax, 10% petroleum jelly and 85% petroleum ether and allowed to dry. Excess jelly is blotted from the surface. The greasy film left behind serves as a release agent so that the transfer may later be removed from the leaf. The leaf, resting on

a glass plate, is covered in acetone and a piece of thin (130um) cellulose acetate is placed on top and allowed to soften. The acetate is then very carefully smoothed onto the leaf to exclude bubbles. The acetone is left to evaporate for about four hours at 25°C, when the hardened acetate transfer can be removed. The acetate film can then be placed in a photographic enlarger and direct negative prints made. If the technique has been successful, the acetate film should show all venation features expressed on the leaf surface plus the external epidermal features. Transfers produced in this manner are durable and easily stored in conventional card files. A direct print of an acetate transfer leaf of Saurauia sp. is illustrated in figure 26. Bubbles on the preparation are mainly the result of outgassing from large trichomes, during drying and shrinkage of the acetate.

Leaf clearing

The method of highest resolution for recovering venation detail is to use a leaf clearing technique. The rationale behind leaf clearing is that various chemical reagents will remove the pigments from leaves and render mesophyll and veins transparent. Venation can then be selectively stained to increase the contrast for observational purposes. A number of clearing techniques are in existence and these can be divided into two main classes. Firstly, hydroxide clearing methods rely on strong bases such as sodium or potassium hydroxide to render chlorophyll, tannins and other leaf pigments soluble. Secondly, chloro-lacto-phenol methods use a mixture of compounds including chloral hydrate, lactic acid phenol and clove oil to remove pigmentation. The first of these methods is most suited to the clearing of ligneous material and is the most widely used type for leaf clearing (Foster, 1952; Hickey, 1973). Hickey has modified this general technique by pre-treating with

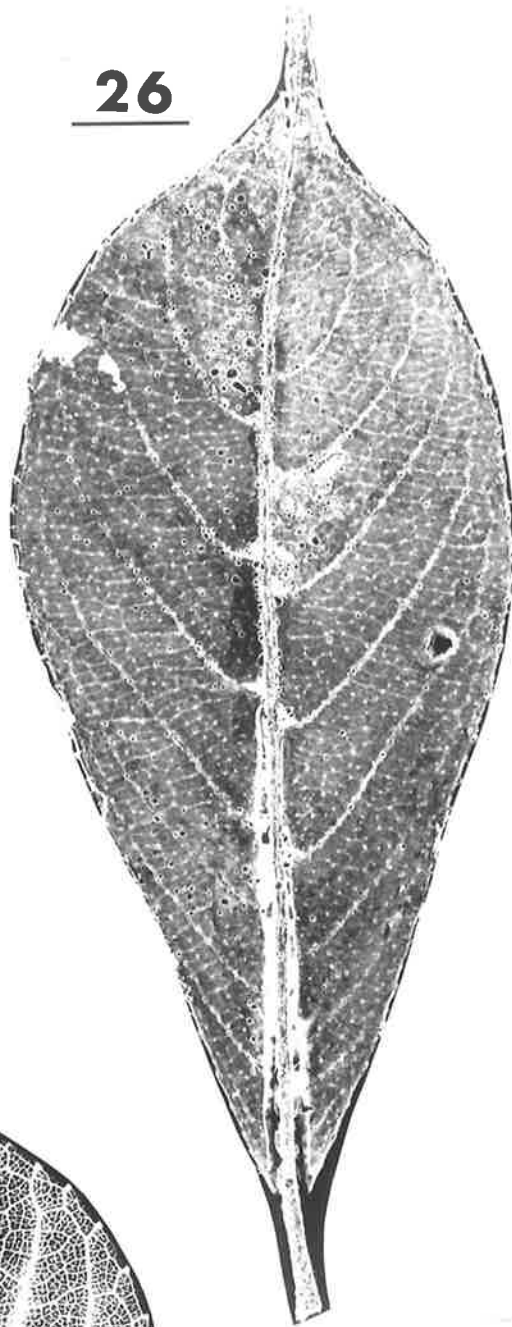
FIGURE 26. Acetate transfer from a leaf of Saurauia sp. (Womersley N.G.F.11376).
The technique is described in the text.

FIGURE 27. Cleared leaf of S.sterrolepida, prepared by the technique of Christophel and Blackburn (1975).

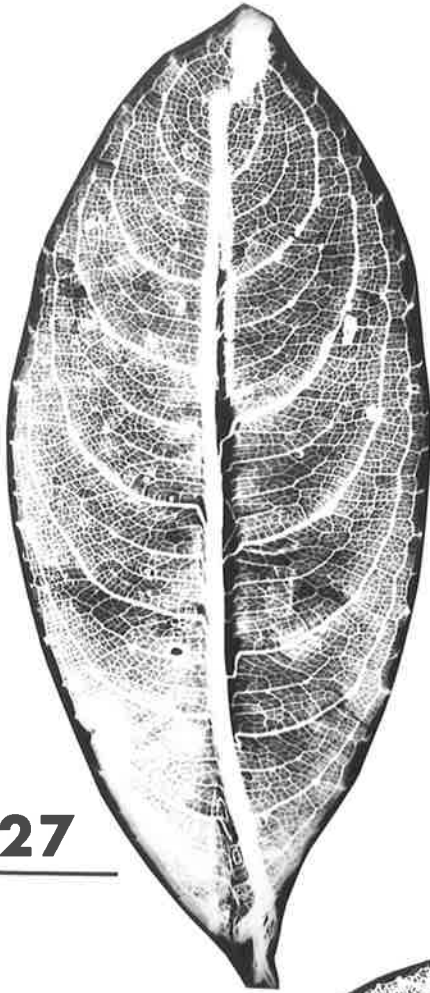
FIGURE 28. Cleared leaf of S.calyptrata, prepared as for the previous figure.

All photographs of cleared or transferred leaves were made as direct prints by inserting the leaf in an enlarger.
Scale bars represent 1 cm.

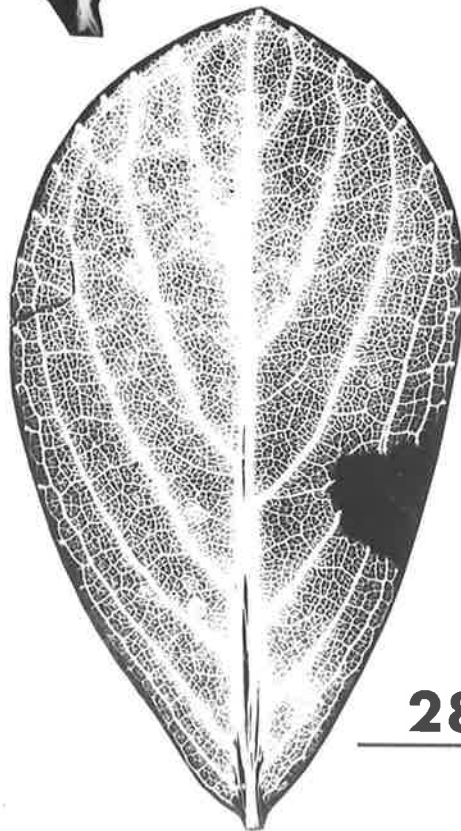
26



27



28



sodium hypochlorite. I have found that this is a less satisfactory method in many cases due to the liberation of chlorine which may destroy the leaf by its gas pressure. A new clearing and mounting technique adapted from the general hydroxide method has been described by Christophel and Blackburn (1975). This technique particularly differs from previous methods in the use of a new mountant. The high refractive index of polyester resin so nearly approaches that of leaf tissue that leaves so treated achieve a high degree of transparency. Details of the technique may be found in a copy of that paper which is bound at the end of the thesis.

The second category (chloro-lacto-phenol) of clearing techniques has found application largely in other fields of biology. The best of these methods is the procedure described by Herr (1971). Herr's reagent is generally known as "four-and-a-half clearing fluid." It has been used to clear such diverse tissues as angiosperm ovules (Herr, 1972, 1973), whole insects and small vertebrates (Harrison and Anderson, 1974) and brain sections (Perry, Harbison and Lumb, 1975). Its application to a wide range of tissue types with modifications has been demonstrated by Blackburn and Christophel (1976).

The general applicability of Herr's technique has been limited by the nature of the clearing fluid. The combination of compounds produces a mixture which is unstable in that its constituents are readily separated. The addition of almost any compound, including most of the constituents of the mixture, results in a rapid and irreversible separation into two fractions. Where this occurs within tissue, it results in an opacity which is impossible to remove. This has previously meant that tissues cleared in Herr's fluid could not be permanently mounted, a serious limitation for the preparation of

long term collections. Two procedures for mounting tissue cleared in this fluid were demonstrated by Blackburn and Christophel (1976) who once again applied polyester resin as a mountant. A copy of this latter paper is also bound at the end of this thesis.

Where high resolution is an important criterion the application of relatively laborious and time consuming clearing techniques is justified. With clearing it is possible to guarantee that all vein orders to the level of veinlets will be demonstrable. In practice it has been found that the most satisfactory approach is to apply the clearing and mounting technique of Christophel and Blackburn (1975) but to prepare two mounts. Whole leaves are mounted between relatively thick glass (1-2 mm), while a leaf slice taken from the midrib to the margin and bounded by two secondary veins is mounted beneath a cover slip. By this method both gross and fine venation details are accessible.

Over the period of time that this method has been applied, it has been found that the stain, Safranin O, tends to fade. The reason for this is that the catalyst for the mountant is a strong bleaching agent. Accordingly, a new approach to staining leaves has been developed. This method uses the property of lignin to reduce aqueous potassium permanganate to manganese dioxide, to stain the venation of leaves. The resultant stain, apart from a little preliminary leaching as the mountant sets, is both colour fast and fine grained. The contrast between veins and mesophyll is also higher than that achievable by any other stain I have encountered. This method has been recently written up and submitted for publication. A manuscript copy is included at the end of the thesis.

A modification of the hydroxide clearing technique of Christophel and Blackburn (1976) has been developed for application to

strongly tanniferous leaves. Leaves such as those of Eucalyptus are highly resistant to normal clearing. They usually require long periods, from a few days to weeks in the hydroxide solution. This often results in the dissolving of ligneous tissues with the result that the leaves are unsuitable for mounting.

The clearing time of heavily tanniferous leaves may be shortened considerably by the following procedure.

Leaves are soaked in a 15% aqueous solution of potassium hydroxide at 40°C for a period generally no greater than four hours. The actual period is dependent on the clearing rate of the leaf involved. Some less tough leaves may in fact be dissolved by this treatment. Such leaves are probably more suitably treated by gentler clearing techniques. Once the leaf has turned completely black or dark brown it is left until it begins to show signs of fading. At this time it is removed and washed under a gently running stream of tap water for about two hours. Washed leaves are then placed into 1:2 100 vol. (40% w.w) hydrogen peroxide/water, and weighted with a piece of light plastic screen. The clearing procedure is hastened by placing the dishes of leaves in sunlight or in a warm cabinet under ultra-violet lamps. Generally even the toughest leaves will clear in four or five hours with this treatment. Cleared leaves are then washed once again in running water for two hours and transferred to saturated aqueous chloral hydrate. This final clearing should take no longer than two hours, after which the leaves can be washed again for about one hour. Staining with the potassium permanganate stain is recommended at this point. Mount the leaves in polyester resin after the usual dehydration steps (Christophel and Blackburn, 1975).

This technique is so rapid that it has been possible to have a

Eucalyptus leaf mounted and ready for photography twenty-four hours after starting the procedure. A limitation of this method is that delicate leaves may be damaged by the liberation of oxygen from the hydrogen peroxide. It is recommended, therefore, that its use be restricted to those leaves that have proven stubborn in other clearing methods.

Because this study has been by intent based mostly on features no finer than tertiary venation, in general sufficient resolution was achievable with photographs of herbarium specimens and with leaf rubbings. The labour involved in preparing cleared leaf mounts was not generally warranted. Several small leaved taxa, however, were treated by this method and cleared leaves of two species are illustrated in figures 27 and 28.

Measuring non-architectural characters

Although not applied in the detailed architectural analysis, features of the leaf indument and epidermis were recorded for most specimens.

Trichometypes and densities were recorded both from untreated leaves and from cuticular preparations. This was done to compare the results of different approaches to measuring these features. Many species of Saurauia have deciduous hairs and appear glabrous if only naked eye observation of untreated leaves is carried out. Secondly, it was suspected that differing descriptions of hair types would arise from the two approaches.

Hair densities were measured from herbarium specimens using a specially prepared grid and viewing the leaf under a binocular microscope. The measuring grid consisted of a 1 mm thick perspex circle with a 1 cm square, divided into four square sectors, inscribed on its surface. Four smaller 1 mm squares were also

inscribed at the centre of the grid. Hair densities were counted on the appropriate area of the grid, using the finer graduations for more dense induments. Hair densities were then converted to hairs/sq. cm. Hair types were classified according to existing terminology (Stearn, 1973). Observations of hairs were also made from cuticular preparations, recording density using an eyepiece grid. Hairs observed in this manner were classified into the set of types given in chapter 3.

Scanning electron micrographs of some major hair types were obtained from a study of Saurauia indument carried out by P.S. Green (unpubl. honours project). These are illustrated in figures 16-18.

Scoring leaf characters for analysis

The most efficient method of measuring leaves for the set of metric architectural characters was to work with photographs or leaf rubbings. The reason for this is twofold. Firstly, they are flat, and secondly, it is possible to mark the relevant points for particular characters directly onto the paper. Herbarium specimens were thus handled as little as possible. The single most convenient piece of measuring equipment was found to be a 45° set square with protractor and graduated base line. All architectural characters could be measured with this. Data were then written directly on the standard description form.

The approach taken in this study with regard to collecting data has been one of maximising information retrieval with minimal work load. This is important if numerical methods are to find general applicability in taxonomy. An objection to existing methods such as the use of terminologies for leaf description, is that they reduce the labour of data collection, but at the expense of information. In the next chapter, the critical choice of an efficient character set from a larger number of characters is therefore discussed.

CHAPTER 5
AN ULTIMATE CHARACTER SET

A trial set of species

Specimens labelled as Sauraula napaulensis DC., S.tristyla DC. and S.macrotricha Kurz. from Yunnan, China, were chosen for an initial study of the character set. For reasons discussed by Hoogland (1977) the specimens labelled as S.tristyla are probably referable to S.roxburghii Wall. They will therefore be hereafter referred to as aff.roxburghii. The specimens labelled S.macrotricha do not possess the conspicuous setae of this species so will be referred to as aff.macrotricha. The specimens of S.napaulensis almost certainly belong to that species.

Leaves of these species were scored for the architectural characters described in chapter 3. In addition indument type was recorded. A single character of the epidermis was also tested here. Stomatal orientation relative to secondary and tertiary venation was scored and is illustrated for a number of specimens in figure 30.

Each character was assessed for the parameters defined in the first paragraph of the previous chapter. The characters in the order in which they appear on the standard coding form (fig.21) will now be discussed and reason for their inclusion or exclusion with regard to further analyses on a large OTU set will be given.

Although there are a priori reasons for suspecting that leaf length and leaf width are highly correlated characters, the ease of scoring and the fact that one of them would probably prove useful, meant that these characters were retained for later analyses. Over the set of specimens under initial study it was evident that S.napaulensis was easily distinguishable generally from the other two

species on these characters. This also proved to be the case with most of the characters. For similar reasons it was decided to retain the characters of position of maximum width, length to width ratio, apex angle, base angle, and number of secondary veins. In the set of specimens the last-mentioned character appeared to be the best discriminator between the species. The character "percentage acuminate" was found to vary considerably on single specimens. However, until the whole range of this character was known for the genus I felt safer retaining it. Leaf shape expressed by shape terminology was a hopelessly variable character, compounded by the fact that the irregularity of many leaves made it difficult to assign them to an inexplicit leaf shape class.

Leaf base type was found to be essentially invariant in the set of specimens. It was decided that this character was less ambiguously expressed by the character of base angle and so was excluded from further analysis. Margins were consistently serrate, so that the character of margin type was of no value in this study. The characters of dentition frequency and tooth size were considered as suitable for further study. The teeth types of Hickey (1973) were found entirely inapplicable to Saurauia even in the whole set of OTUs because of the variability of this character. The general descriptive terms of teeth type such as setose, callose, spinose, were included in the leaf descriptions (appendix G) but excluded from detailed analyses. Petiole length was considered to be a useful and easily stored character and so was included for further analysis.

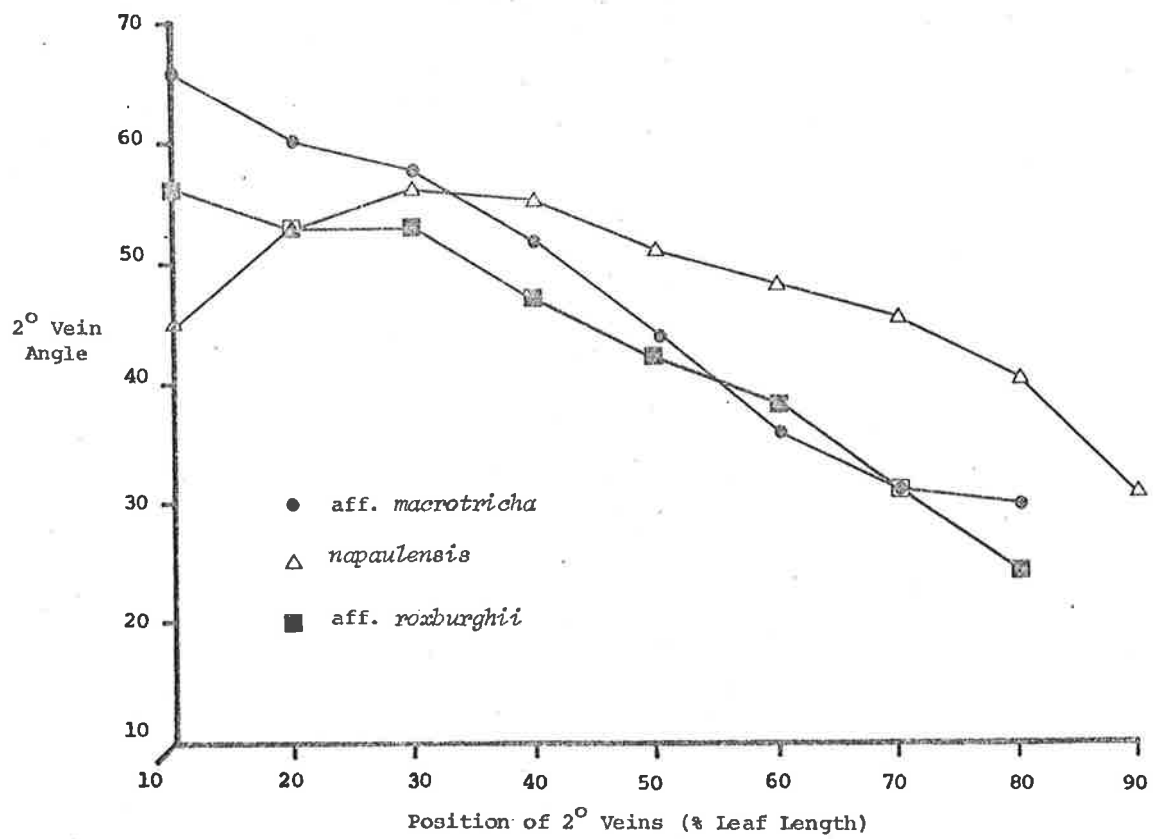
Secondary venation characters appeared to be the best discriminators of the set of three species. Secondary vein arrangement defined by terms such as "alternate" and "opposite" was rejected. No leaf was found which could be classified clearly as belonging to

one of those two classes. Expressing vein arrangement in terms of the percentage of veins that were either opposite or alternate was also tested and abandoned because of the arbitrariness of the boundary between the two character states and the extreme variability exhibited within single specimens.

The number of secondary veins possessed by the leaves was particularly useful for separating S.napaulensis from the other two species. Secondary intercostal shape was also an efficient discriminator. Although this is a laborious character to measure, it was retained for further study. Secondary vein curve, however, was rejected. This character is defined by the closest fit of a secondary vein to a set of standard curves. Problems of scaling for comparison of curves, subjective choice of the best fit and the variability within single leaves meant that this character was unsuitable for further consideration. The characters of secondary vein straightness and secondary vein angle could express most of the information, without the problems of secondary vein curve.

It was observed that the spacing of secondary veins along the length of the leaf was not constant. This variation affects the character of intercostal shape, secondary vein angle, and secondary vein length percent. A similar set of characters of vein spacing and variation of secondary vein length with position on the leaf, were used by Ferguson (1971) in describing fossil leaves. The variation of secondary vein angle, intercostal shape and straightness were closely studied in the three species. The relationship between vein angle and position is illustrated in figure 29. Clearly in all three species there is a trend towards reduced vein angle adapically. This is also the case for intercostal shape and straightness. There were three approaches considered for handling each of these characters.

FIGURE 29. Variation in secondary vein angle along leaves of three Sauraulia species. The position of secondary veins is measured from the leaf base.



Firstly the mean value of each character, computed over all, or a subset of secondary veins, is a simple method of expressing the characters. Secondly the individual values for each vein could be scored. This approach is particularly difficult to treat satisfactorily with existing statistical procedures. The problem is that each leaf is likely to have a different number of veins and individual vein scores cannot be considered to be independent. The third approach was to score the mean value and the slope of the secondary vein character against vein position plots, e.g. the slope of the plots in figure 30. To compare values of slope of the plots obtained from different species requires firstly plotting the values of secondary vein characters for all leaves of each specimen. The mean plot is then determined by interpolating the individual leaf plots to find the values at a number of standard positions on the leaf, e.g. every 10% of the leaf length. The mean value at each standard point is then computed and a regression curve fitted. Where the curve is close to a straight line the slope of a fitted straight line can be used as an approximation. Where at least some plots are markedly curved, problems in comparing slopes arise. The work load in applying this procedure is very great, especially when one considers that the starting values such as intercostal shape are themselves laborious to score. It was decided therefore that mean values of secondary vein characters for each leaf were most suitable. If leaf scoring ever becomes possible by the use of computer handled, pattern recognition procedures, the variability of venation characters along the length of leaves may prove to be useful. A somewhat less accurate method of considering this variability is to use descriptive terms such as those proposed by Hickey (1973) for describing the variation of tertiary vein angles both exmedially and adapically on leaves.

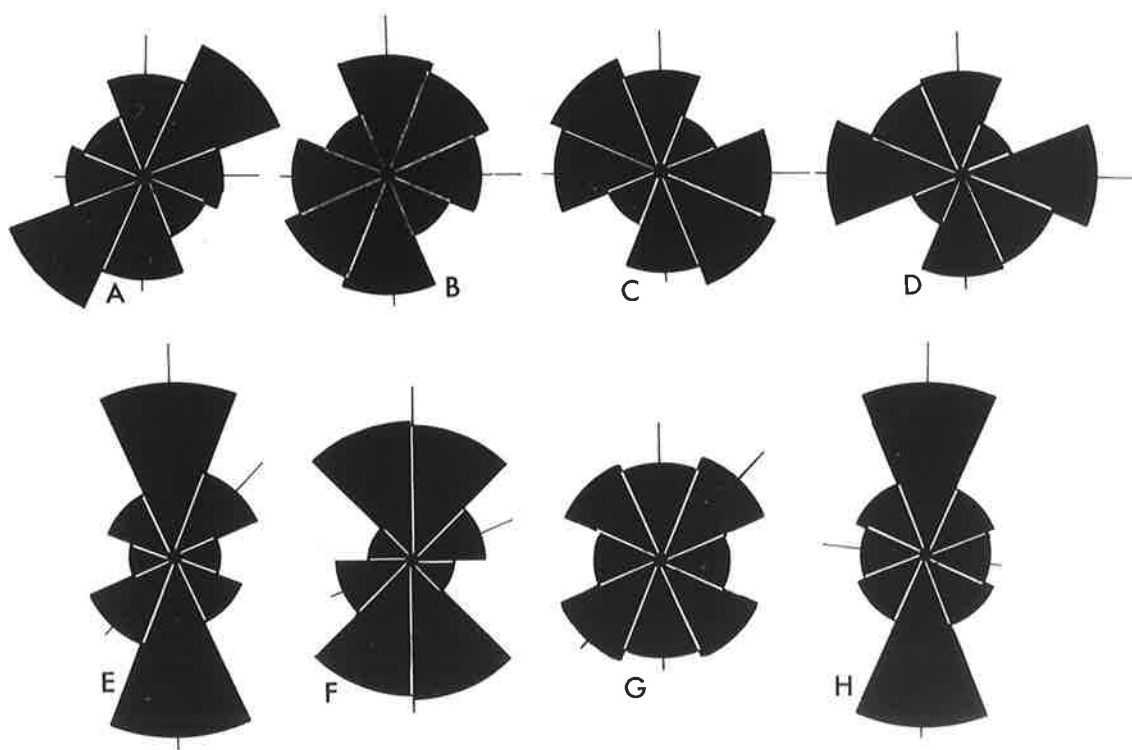
FIGURE 30. Radial frequency plots (rose diagrams) of stomatal orientation in Saurauia spp. Vertical axis is secondary vein direction, other axis represents tertiary vein direction.

(a-d) S.aff.roxburghii, specimens 515633, 373615, 420714, 411248.

(e-g) S.napaulensis, specimens 53212, 420672, 41953.

(h) S.aff.macrotricha, specimen 53207.

All specimens from Yunnan, China.



However, since it was found that Saurauia leaves consistently exhibit a trend of adapical reduction of secondary intercostal shape, secondary vein angle and secondary vein straightness, such terms as "decreasing adapically" were of no value in this study. The variation of tertiary vein angle at various positions on leaves, however, was considered. Measurements of tertiary vein angles were made initially at twenty-four positions on the leaf blade. These were taken as four measurements at each of three positions along the length of the leaves i.e. at the 30%, 50%, 70% and 80% half blade widths for each of 20%, 50% and 70% of the leaf length. This was repeated for each side of the leaf (i.e. alternate sides of the midrib) to obtain the full set of measurements. Once again the labour involved was considerable. To test the relationship between mean values computed from all measurements on the leaf and single values, twenty-four linear regressions were performed, one for each position of measurement. The highest correlation between a single value and the mean value over all positions on the leaves was found with the leaf centroid value. The centroid value is that at 50% of leaf blade width and 50% of leaf length. The correlation coefficient between this value and the mean over fifteen leaves from the three species was found to be .963, which is very significantly different from zero even at the .01 level. It was therefore decided that the single centroid value would suffice as a compromise between information loss and increased labour in scoring the character for twenty-four positions per leaf.

The variation of tertiary vein angle across Saurauia leaves was also found to be consistent. Tertiary vein angle always increased both exmedially and adapically. Tertiary vein course was consistently strongly percurrent. Quaternary veins were consistently polygonal in

arrangement, generally quadrangular or pentagonal. Both of these characters were retained for the purposes of leaf description, but were excluded from analyses because it was found difficult to score them in a completely unambiguous form without considerably increasing scoring loads.

Characters of the leaf indument and epidermis with one exception were considered as essentially descriptive. The single epidermal character which was closely studied from a numerical point of view was stomatal orientation. This was done to investigate whether there was any species dependent consistency in stomatal relationships.

Leaf fragments were taken from near the centre of leaves, trying to include a piece of secondary vein. Cuticular preparations were then made and the orientation of stomates in a number of fields scored. Fields for scoring were chosen where possible to maximise the distance from secondary tertiary veins. This was done to try and minimise local effects of the veins on stomatal orientation. The number of stomates scored was between 110 and 260 with an average of about 170. The actual number was dependent on the size of prepared cuticular fragments. The orientations of individual stomates were grouped into one of eight equal angled sectors on a radial grid. Rose diagrams i.e. radial frequency plots, were then prepared for each leaf. These were oriented with respect to the direction of secondary and tertiary veins on the leaf fragment. A selection of the rose diagrams are illustrated in figure 30. It is clear that at least in some cases there is a strong influence of vein direction on stomatal orientation. However, there is also considerable variation about this e.g. within the specimens of *S. aff. roxburghii*. Without a very carefully controlled and detailed study it was decided that stomatal orientation was of little value as a discriminatory character.

Variability and correlation of characters within species.

Having performed a preliminary assessment of the character set on three species, it was decided to study the variation of characters within a single species in detail. Thirty specimens labelled as Saurauia conferta Warb. were therefore chosen. This species was selected for no other reason than that it was the best represented species obtained to that time. The correct identifications of these specimens were not considered important for the reason that they formed a morphologically homogeneous group. It was decided that testing for morphologically defined groups of OTUs is more important than testing taxonomic species limits.

Specimens of S.conferta were each scored for fifteen metric variables. Various statistics are given in table 2 for each character.

Character	Range	Mean	Variance	Var/Mean
Leaf length	9.8-32.7	20.92	30.9	1.477
Leaf width	3.2-10.7	7.33	3.77	0.514
Length to Width	2.02-4.1	2.91	0.32	0.11
Apex angle	37-85	58.6	237.1	4.046
Base angle	25-79	44.1	246.5	5.59
Posn. Max. Width	47-71	60.16	20.27	0.337
Percent acuminate	3.3-13.1	6.3	4.01	0.637
No. of 2° veins	10-20	14.25	10.71	0.752
2° Intercoastal shape	4.0-7.2	5.19	0.65	0.125
2° Straightness	.34-.73	0.53	0.0044	0.0093
2° Vein angle	25-49	34.6	41.1	1.188
3° Vein angle	58-89	72.8	81.3	1.117
Dentition freq.	1.6-4.7	2.36	0.24	0.101
Petiole (% leaf lth.)	4.2-12.9	8.51	4.11	0.483
Shape index	57.3-70.5	64.1	9.80	0.1529

In no case was the distribution of a character significantly different from the normal distribution.

TABLE 2. Statistics for various characters scored over thirty specimens of S.conferta.

Clearly several characters are highly variable within this species e.g. leaf length, apex angle, base angle, 2° vein angle, 3° vein angle and number of secondary veins. Note that these characters are commonly used in descriptions of leaves and are amongst those considered as of limited value by critics of the usefulness of leaf features (Hamilton, 1916; Odell 1932). The variability of characters in determining their value in the final character set is discussed later in this chapter.

A further approach to assessing the value of a character for discriminatory purposes is correlation analysis. Using Pearson's product moment correlation coefficient the correlation between pairs of characters has been assessed. Correlation gives a measure of redundancy of information in a character set. For example, a correlation coefficient approaching one for a pair of characters suggests that they possess essentially the same information across the set of OTUs. Therefore one of such a pair of characters should be deleted from the variable set. Significance of a correlation coefficient can be determined from a t-test for significant difference from zero. Those pairs of characters with correlation coefficients significantly different from zero at the .01 level are given in table 3. This significance level simply means that there is at least a 99% chance of obtaining a correlation coefficient nearer zero than the one in question. In interpreting correlation coefficient (r) values the sign is important in understanding reasons for and direction of, the correlation.

Character pair	r
Leaf length x leaf width	0.832
Leaf length x no. of 2° veins	0.749
Leaf width x no. of 2° veins	0.776
Leaf width x 2° intercostal shape	0.729

Character pair	r
Length to width x apex angle	- 0.776
Length to width x base angle	- 0.665
Length to width x 2° Intercost. shape	- 0.537
Length to width x 2° Vein angle	- 0.534
2° Intercostal shape x no. of 2° veins	0.688
2° Intercostal chape x 2° Vein angle	0.532

TABLE 3. Correlation coefficients for variable pairs,
computed for thirty specimens of S.conferta.

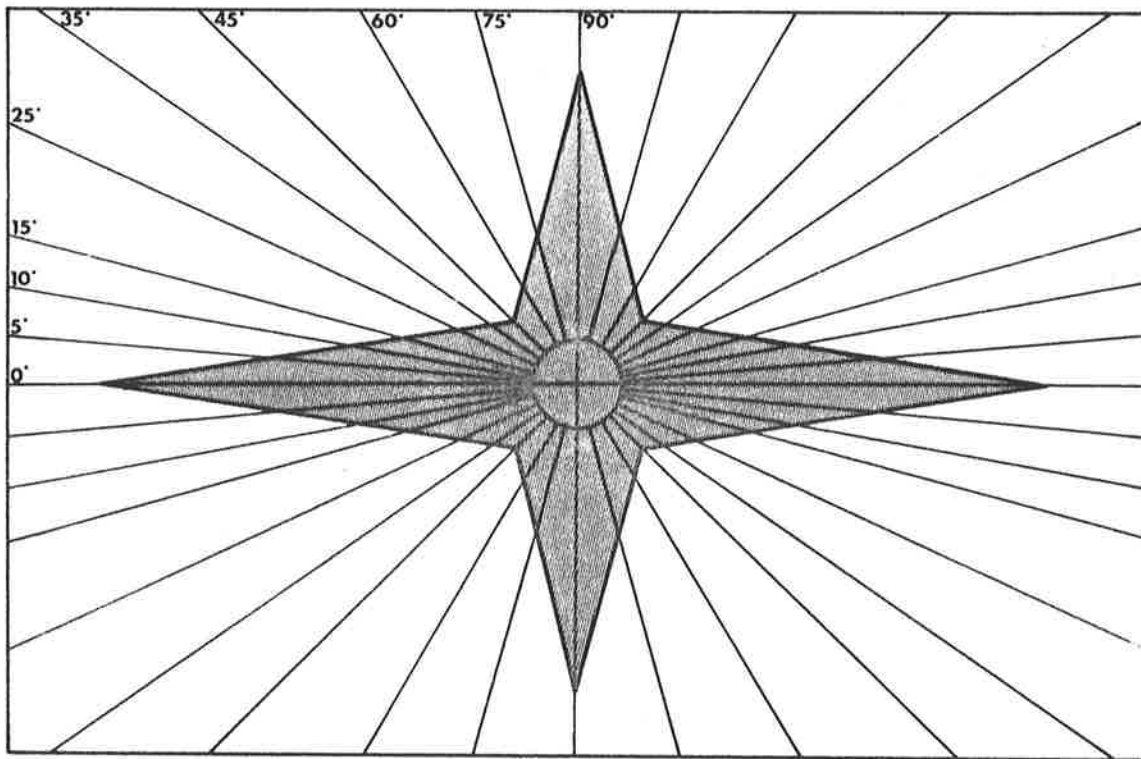
Given that the set of specimens of S.conferta are genetically nearly homogeneous at the species level, then it is possible to suggest plausible and simple reasons for the observed correlations of the characters. For the set of specimens it can be seen that there is little variation in leaf shape. All leaves of the specimens are simple, elliptic to obovate, varying mainly in size. Therefore it is obvious that leaf length and leaf width are correlated for the reason of homogeneity of the set of OTUs. It is possible to generate all the leaf shapes and sizes found in the set of specimens from the "mean" specimen by a simple transform grid (Sneath and Sokal, 1973). This procedure is illustrated for a pair of generalised leaves in figure 32.

Simple geometric properties of the characters can be used to explain correlations between leaf shape expressed as the length to width ratio of the leaf and apex and base angles. If one models a leaf as an ellipse then increasing the length to width ratio must make the ends progressively narrower angled. This explains the reason for the negative correlation between L/W and terminal angles.

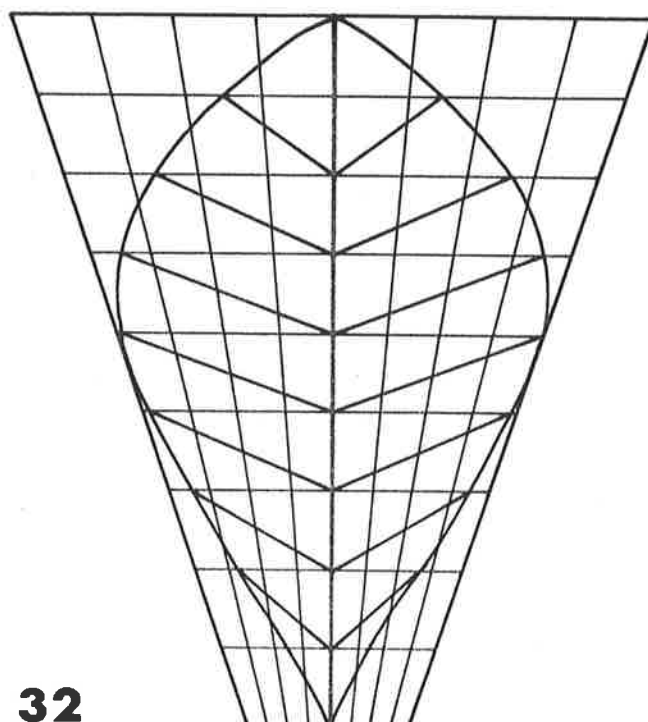
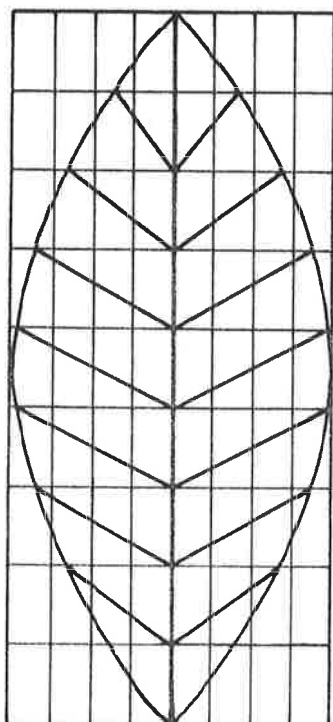
Functional reasons can be suggested for the empirical correlation between leaf length and number of secondary veins. Given that in a particular environment a single secondary vein is able to supply

FIGURE 31. Leaf shape index measuring grid overlain on a hypothetical leaf represented by the shaded area. The value of leaf shape, measured by the method described in the text, is 25.8%.

FIGURE 32. Simple transforms of generalised leaves to illustrate relationships between leaf shapes.



31



32

a fixed area of leaf then an increase in leaf area measured as an increase in leaf length could be expected to be accompanied by increase in the number of secondary veins. This correlation is, of course, influenced to an indeterminate extent by both genetic and environmental mechanisms. The question of environmental influences on leaf characters will be considered later.

Lastly simple geometric restraints can be invoked to explain the correlation between length to width ratio and the characters of secondary veins. By altering no other factor than length to width ratio it can be seen that both vein angle and vein length, hence intercostal shape must increase. Likewise, secondary vein intercostal shape is dependent on the spacing of veins and therefore number of veins. Given constant vein angle, then increasing width will tend to increase vein straightness. The inter-relationships between various geometrically dependent characters can be seen once again by studying figure 32. This figure illustrates the effect of a simple transform from an original hypothetical leaf shape (a) to a new shape (b). Characters of leaf shape, apex and base angle and secondary vein course are affected by the transform.

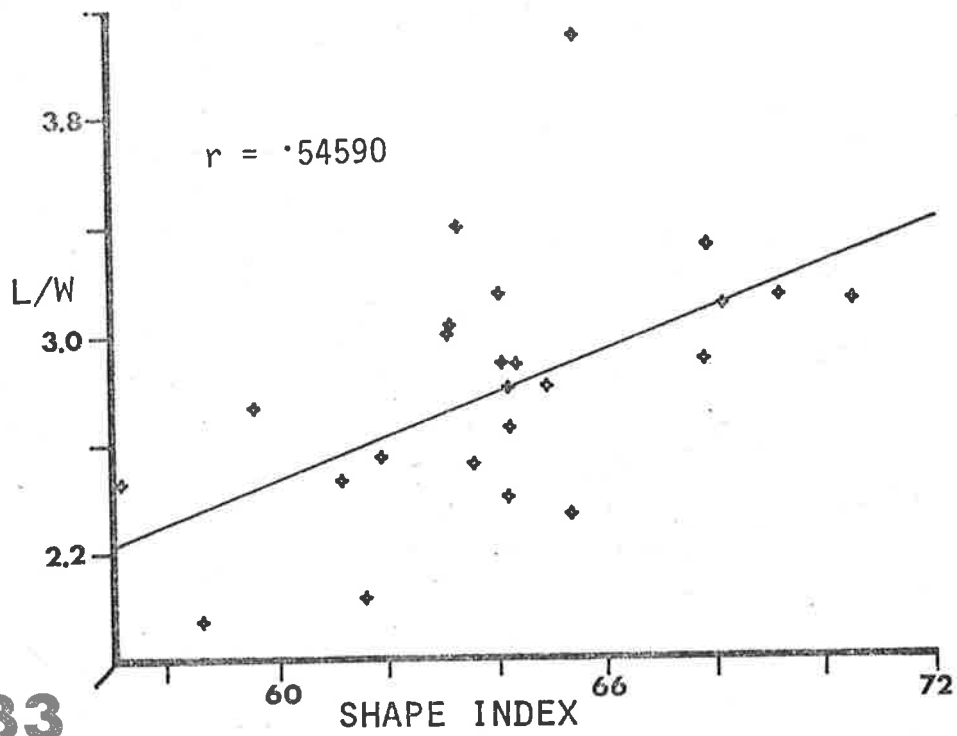
There are a priori reasons for supposing that at least some of the metric characters used in this study are interdependent. The correlation between various characters has been demonstrated empirically by using the leaves of S.conferta as an example. Based on these observations there would be strong grounds for rejecting a number of the more highly correlated characters if the whole study was of S.conferta alone. However, as has already been observed in the three other species, at least some of the characters are useful discriminators. Before considering further the question of correlation of characters, one final character will be considered.

An index of leaf shape has already been considered. Leaf shape is generally expressed by a set of terms, most recently defined by the Systematics Association for descriptive terminology (1962). Leaf shape can also be expressed in terms of other characters influenced by it, such as length to width ratio and apex and base angles. A further approach is to use a leaf shape index. The method described in chapter 3 computes the mean distance from the midrib to the margin by means of a radial grid. Where the individual distances are large as in the case of an oblong leaf the value of the index will be large. Conversely where a leaf has a long attenuated apex and base the index value will be small (fig.31). Clearly then there is evidence for supposing that there are geometrical reasons for correlation between an index of leaf shape and parameters such as apex and base angle and length to width ratio. To determine whether such a correlation exists, bivariate scattergrams of leaf shape against these other measures of shape have been prepared for S.conferta (figs. 33 and 34). Pearson correlation coefficients have been computed for each pair of variables. A plot including base angle is not illustrated but has an r value of -0.38426 . Empirically the most significant correlation is between length to width ratio and shape index. This is expected from geometrical determinants of the shape index. Since fixed angular references are defined for measuring leaves, an increase in length to width ratio of a leaf with convex margins will mean that the intersections of radii with the margin will be generally greater. This is especially true around those regions where the radii are close spaced i.e. apex and base. The high negative correlations between apex and base angle and length to width ratio have already been considered. It is reasonable to expect therefore a negative correlation between these

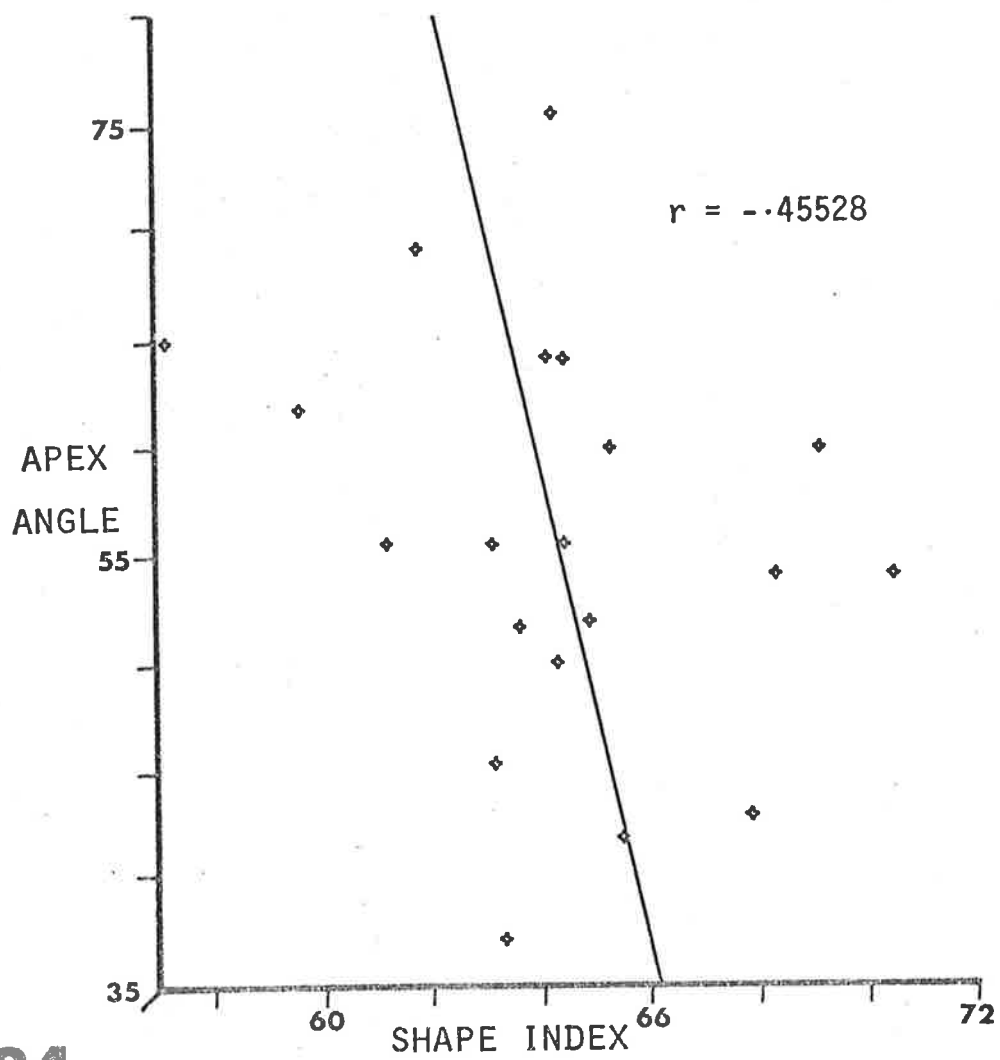
FIGURE 33. Scattergram of S.conferta specimens plotted for length to width ratio against leaf shape index. Correlation coefficient (r) given for fitted linear regression.

FIGURE 34. Scattergram of S.conferta specimens plotted for apex angle (degrees) against shape index (%).

33



34



angles and the shape index. Figures 33 and 34 illustrate that this is indeed the case. In the case of leaves with essentially convex margins, one would expect this to occur because of the geometric limitations on termination angles and length to width. Such a leaf with a low (25.8%) shape index is shown with the shape index grid overlain, in figure 31.

The shape index has been shown to be a useful discriminatory character and to be quite stable within species (pers. comm., R. Hill, Botany Dept., University of Adelaide). However, it is difficult to measure and is correlated to other measures of leaf shape, such as apex and base angle and length to width ratio. These characters are also simple to measure and so more applicable to a survey of a large number of species. Where these characters are consistently available on a set of specimens, it is hard to justify reducing them to a single shape character. The value of the shape index lies in those studies where leaf bases and apices are frequently missing e.g. in a suite of fossil leaves. In cases of such missing characters, the shape index is able to recover useful information which would otherwise be lost.

Environmental influences on leaf morphology

Before venturing into a detailed discussion of the variability and relationships between characters in a large set of OTUs it is worth considering some influences on leaf characters. The variability of leaves can be divided into three components, genetic, environmental and experimental. Ideally in a study which hopes to find a leaf morphology based reflection of the natural relationships between OTUs, only the genetic component is desired. The experimental component of variability is a result of two factors, inaccuracy in scoring and measuring leaf characters, and misidentification of

specimens. A discussion of minimising the influence of inaccurate measurements has already been given in chapter 2. The errors of identification are not considered to be of great importance in the present study. The limits of species as they correspond to existing taxonomic ideas are not particularly important. Of greater relevance is the clustering of species into morphologically homogeneous groups. In cases where there is some doubt in the specific identification, it has been found possible to at least relate the specimens to a general group of species with similar morphology.

Components of variability can be assessed for their relative contribution by means of an analysis of variance. Unfortunately the sample size for most of the species in this study is quite small. This precludes an analysis of variance across the whole species set. Further, to distinguish environmental from genetic variance it is necessary to have a set of pure environmental variables. The only such data extractable from the set of OTUs comes from herbarium label data. Altitude, geographic location and forest type are the only environmental factors noted commonly on the labels. Of these, geographic location is of almost no value as the ranges of species involved are not generally very great. For an analysis of the influence of altitude and forest type on leaf characters, the best represented species, S.conferta was chosen. Unfortunately other well represented species did not have as complete label data as this species. Correlation analysis was conducted on S.conferta to assess whether there was any relationship between leaf characters and environmental variables.

The only characters which showed environmental dependence were features of leaf size and number of secondary veins. No correlations were found which were statistically significant. There was, however,

an interesting trend noted in several characters, in terms of their ordination by forest type.

The variable, forest type, needs to be explained further. On the set of S.conferta specimens it was noted that label data included generally a description of the growing site. These sites could be classified into three types, (1) primary forest, (2) secondary forest and (3) disturbed or open site. The names of each of these are derived from those commonly used on the labels. Some specimens were not specifically described as belonging to one or other of these types. Therefore an attempt was made to classify them on the basis of information supplied.

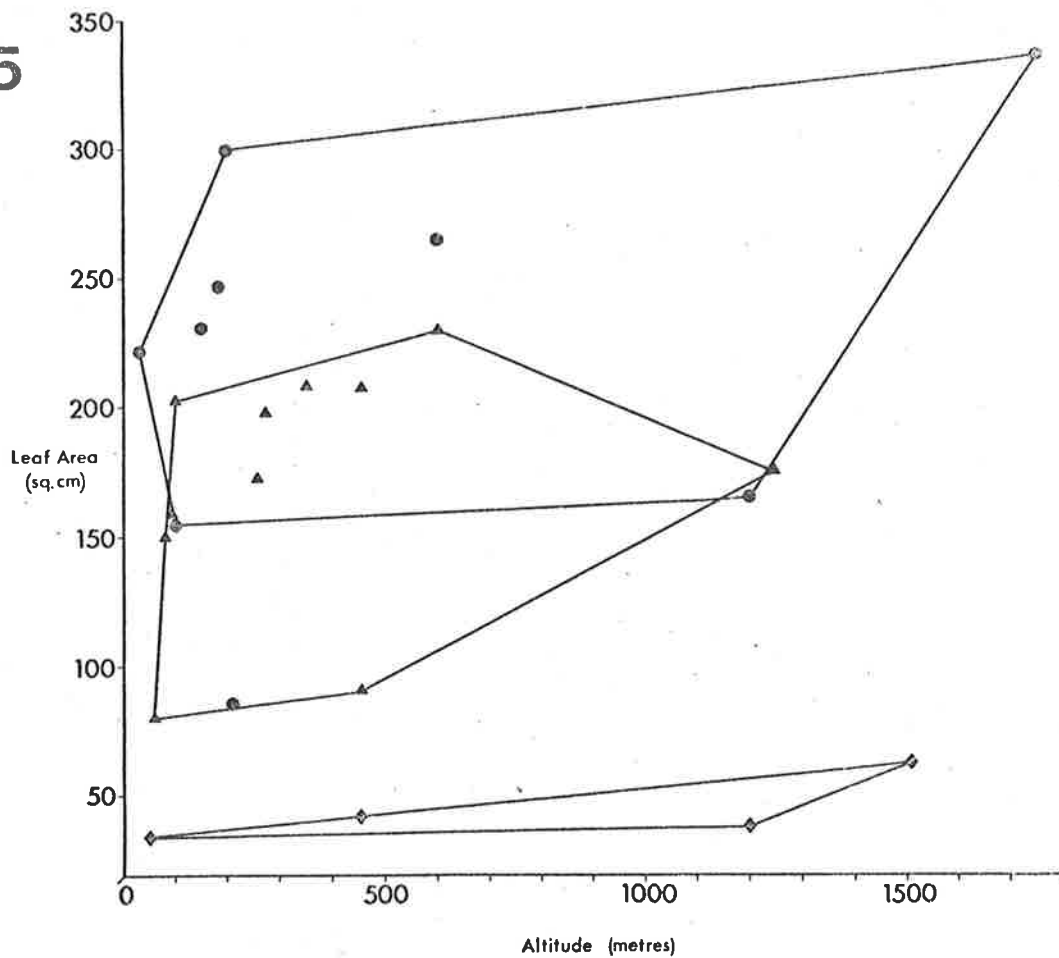
Primary site when not specifically mentioned on the labels, was defined for terms such as climax forest, old forest, dense forest in gullies. Arbitrary secondary forest assignments included descriptions such as regrowth and primary forest margins. Disturbed sites included roadsides, landslides, and cliff faces. Disturbed is used here in the broad sense to mean open, more or less stressed sites.

The architectural features which showed a consistent ordination by site, are leaf length and width, leaf area (as length x width), number of secondary veins and secondary veins per unit area. Figures 35 and 36 illustrate this consistent ordination. Figure 36 is plotted as $-(\log. \text{ of vein density})$ to make the scale linear. The overlap between primary and secondary site clusters is partly explained by incorrect assignment of specimens to these groups. The subjective assessment of these two types by collectors is highly likely to result in incorrect description. There may also be error introduced by the assignment of unspecified specimens to the site types. One thing is clear however, plants from disturbed sites have smaller leaves and lower vein densities. It is worth noting that a single

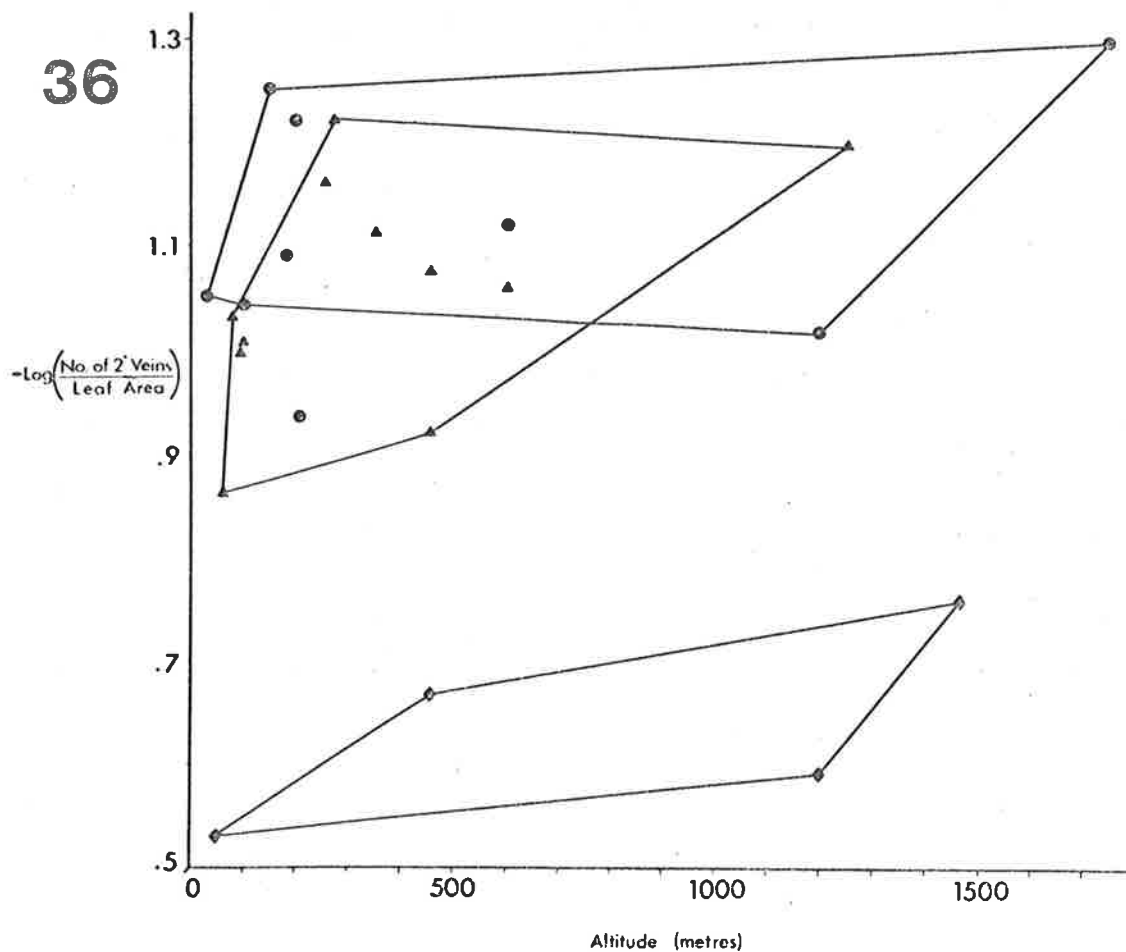
FIGURES 35 & 36. Scattergrams of S.conferta specimens plotted by environmental factors, and clustered into site groups.

- primary forest
- ▲ secondary forest
- ◆ disturbed site

35



36



primary forest specimen lying near the bottom of the secondary forest cluster in both plots, was specially noted by the collector as coming from well drained, sandy soil.

It is possible to suggest some reasons for this ordination. Plants growing in open, or well drained sites are likely to be exposed to more water stress than those within closed canopy forests. The broad mesophylls with open vein networks are likely to be at a disadvantage in such sites. The reduction of leaf size within a species appears to be related simply to environmental influences. The secondary vein density must, however, be seen as a combination of two effects. Firstly one would expect that stressed leaves might require a higher vein density to supply water to the mesophyll. In the case of S.conferta, there appears to be a correlation between leaf size and vein density. If one considers the actual number of secondary veins, one finds that the correlation is the reverse of this. That is, there is a significant positive correlation between leaf size and the number of secondary veins. Small leaves therefore have fewer secondaries. However, since leaf area is a product of linear dimensions, the reduction in leaf area could be expected to be proportional to the square root of the number of secondary veins. Leaf area would therefore "run ahead" of vein number so that the number of secondary veins per unit area would increase with decreasing leaf size. This expected logarithmic correlation has been found to be the case with S.conferta and is demonstrated in figure 37.

Figures 38 and 39 also illustrate the lack of correlation between leaf parameters and altitude. This is perhaps not surprising since the altitudinal range of S.conferta is not very great (1800 m). Altitude is plotted here to demonstrate the clusters more effectively.

The relationship between leaf characters and environmental

FIGURE 37. Logarithmic relationship between leaf area (length x width) and vein density (no.2° veins/leaf area) illustrated by means of a scatter-gram of 24 specimens of S.conferta. Specimens are coded by symbols into growing site type as for figures 35 and 36.

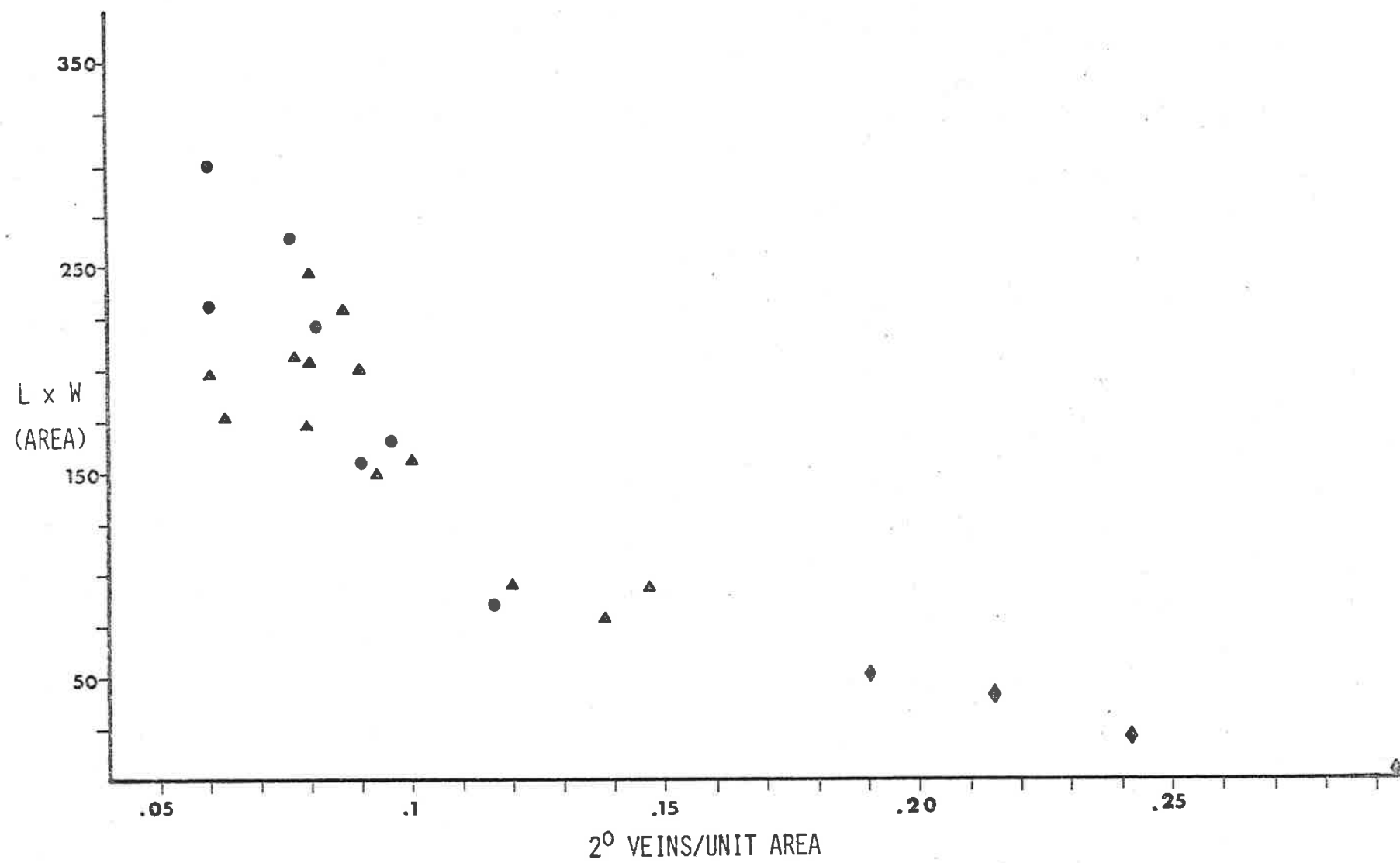
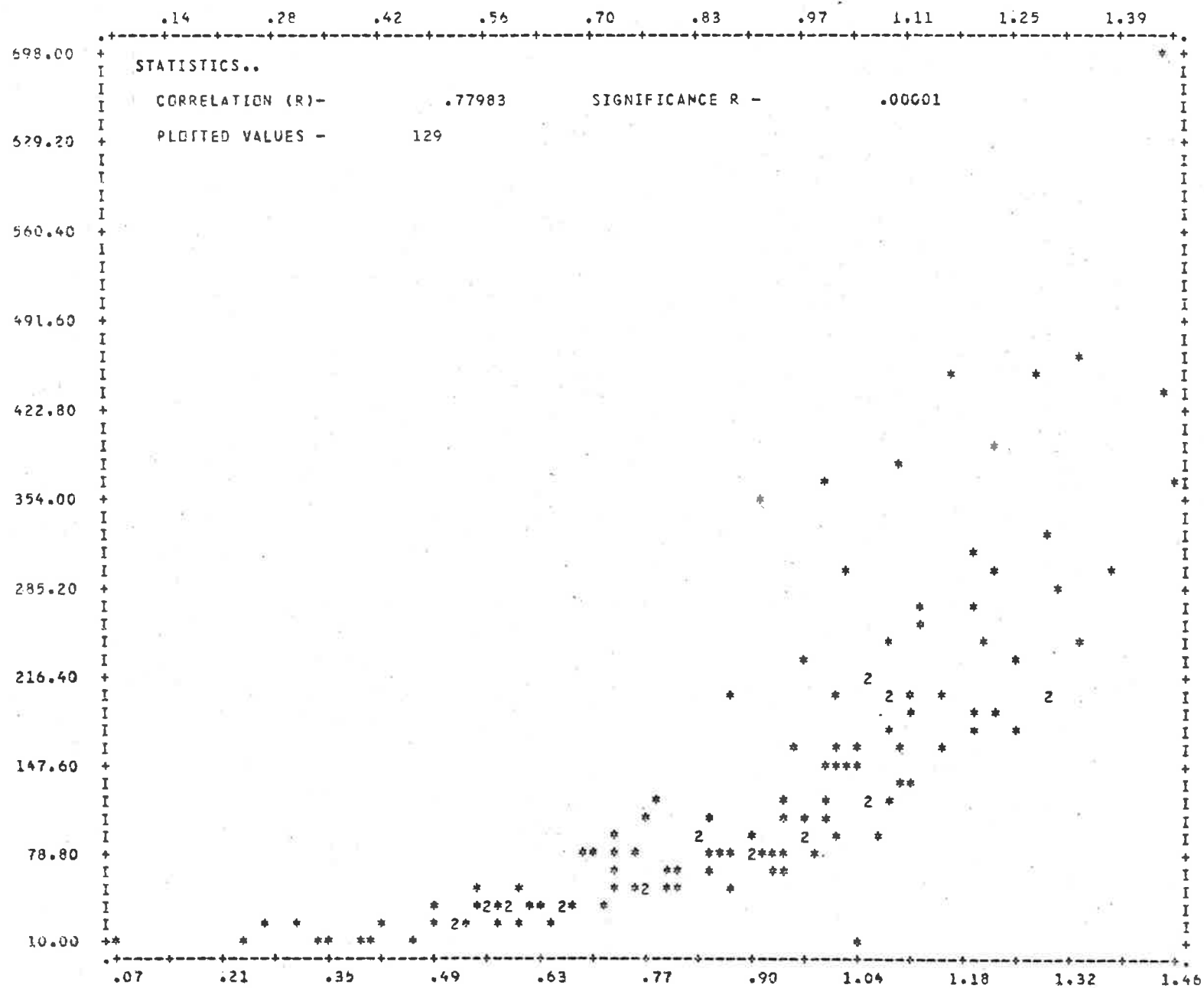


FIGURE 38. Scattergram of 129 specimens of Saurauia species, plotted on leaf length x width (vertical) and number of secondary veins (horizontal).

FIGURE 39. Scattergram of 129 specimens of Saurauia species plotted on leaf length x width (vertical axis) and (-log. no. of secondary veins/area) (horizontal axis).

SCATTERGRAM OF (DOWN) A2

(ACROSS) A7



variables was tested further on a larger set of species. Label data for 129 specimens over 42 species contained equivalent information to that tested on S.conferta. The set of specimens was treated in the same way as for that species. The most significant correlations over the data were between leaf area and number of secondary veins (fig.38), leaf area and vein density (fig.39) and altitude and vein density (fig.40). The first two of these were expected from the observations on S.conferta. The last mentioned, however, had not been demonstrated in that species. Although the correlation coefficient of $-.37$ for this correlation is statistically significant, the actual scatter of points is quite wide. It does illustrate the trend to increased vein density at higher altitudes, in this case up to 3800 m. Subjectively this can be related to the fact that the three specimens from above 3000 m have small, thick, spiny leaves with a dense indument on the lower surface and thick cuticles. At lower altitudes, however, this grouping becomes obscured by the influences of forest type and the fact that the specimens tested were from the whole range of the genus in the Old World.

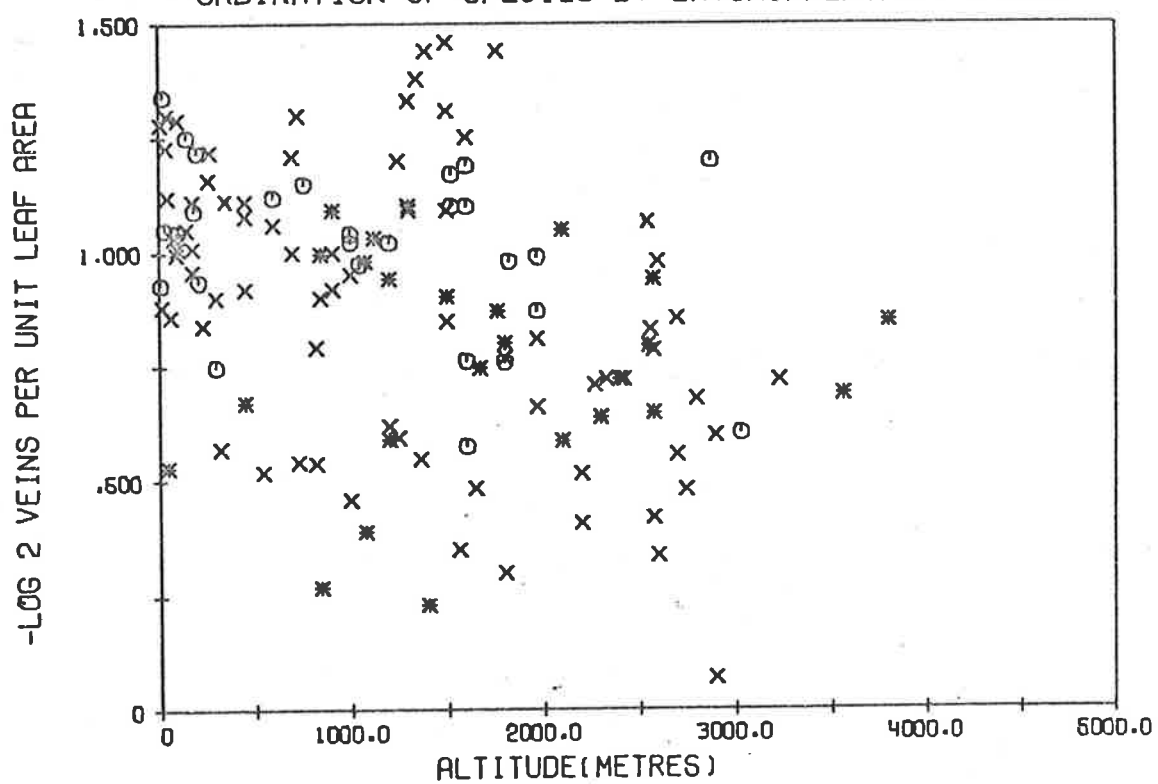
The lack of consistent ordination of specimens by environmental influences has meant that except for one species, the relationship between environmental and genetic influences on leaf form is not known. Although unfortunate, there is no reasonable way of overcoming this in the present study. The only approach that could be taken to resolve this further is to conduct a detailed field survey of the genus. This was impossible within the present study. Therefore no further distinction between environmental and genetic variance will be considered.

Ignoring environmental variance, there are two components of genetic variance. These are the variance within species and that

FIGURE 40. Secondary vein density for 129 specimens over 42 species of Saurauia plotted against altitude. Species grouped into site type.

- ⊙ primary forest
- X secondary forest
- * disturbed site

ORDINATION OF SPECIES BY ENVIRONMENTAL FACTORS



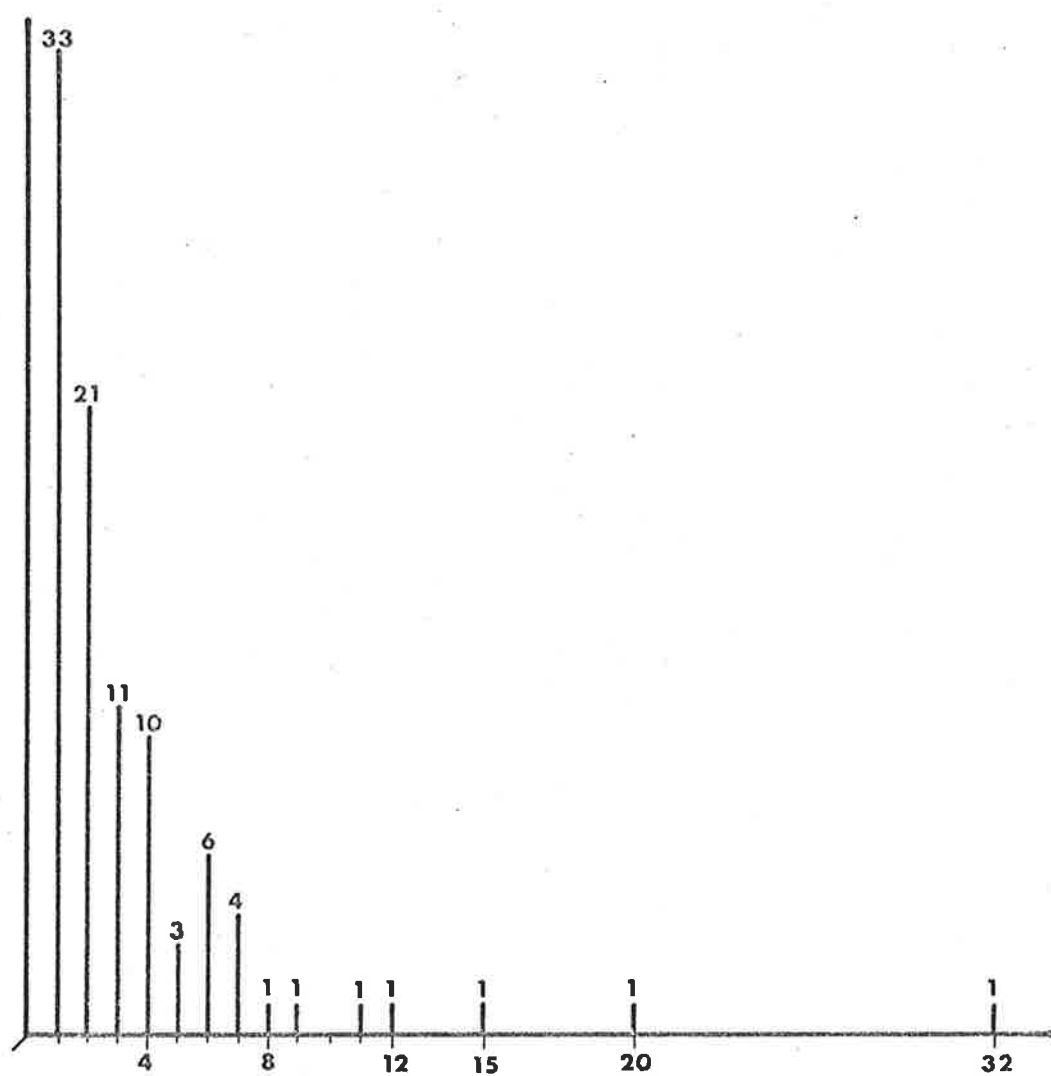
between species. A consideration of these components is important in selecting an efficient character set. Where the ratio of within species to between species variances approaches one for a character, that character has little power to discriminate specimens into species, or species into species clusters. Where the variance ratio is very high, the converse is true. The variance ratio is known as the F-ratio and a test of significance for differences between variances is known as an F-test.

Variance analysis

To apply an analysis of variance, using the F statistic, to Saurauia it was necessary to make some prior decisions about the limits of species. In general, identifications of herbarium specimens to species were considered reliable. However, of more importance than whether individual specimens were correctly identified was that the groups so formed should be morphologically homogeneous. Correct identification is implicit in this statement, but this may not always be the case. For example a species that has been split unnecessarily may be morphologically homogeneous, yet this group transcends taxonomic boundaries. Likewise a heterogeneous group may not have been recognised as separate species by taxonomists. Therefore based on a subjective appraisal of morphological homogeneity, the set of specimens has been grouped into 94 species.

Variance ratios have been computed over the whole set of species rather than on individual pairs of species for the reasons given above. A further reason for this is that sample sizes were not constant over all species. Figure 41 illustrates the frequencies of species with particular sample sizes. Since the reliability of a sample based estimate of population variance is sample size dependent, then comparisons of variance for species represented by only two or

FIGURE 41. Frequency distribution of number of specimens (horizontal) against number of species (vertical) for 94 species of Saurauia included in the study.



three specimens is generally unjustified.

Sample size for the present study has been a major problem in assessing the reliability of individual comparisons between species. Since it was impossible to make field collections myself, the whole sample of specimens consists of previously collected herbarium specimens or a limited number of collections made for me by Dr. M. M.J. van Balgooy. It is doubtful whether anything less than planning a major collecting expedition to all the countries of the Old World where Saurauia grows would help this problem greatly. Many species are known from single collections and have not been recognised since. Because of the rapid fall off in the number of species against sample size (fig.41) initially it was considered that variances might have to be computed from individual leaf data where a species was considered in terms of individual leaf values rather than specimen means. There is some justification for this approach in a study where actual variances are of limited importance (e.g. where they are not used to define species limits). To explain the philosophy behind this approach it is worth giving a "species" definition. A species can be considered as a local cluster in character space of centres of morphological homogeneity, each of which is a specimen. Thus individual leaf values contributing to specimen variance can be seen as contributors to the species variance.

However for at least part of this study it was considered important to be able to define species limits. That is, it was eventually desired to use species rather than specimens as OTUs. Accordingly, only a subset of species was chosen to estimate within species variance. From figure 41 it can be seen that 29 species are represented by four or more specimens. It decided that 29/95 species would give an acceptable estimate of within species variance. This sample

size gives an estimate of variance with confidence of about $\pm 7\%$. Species with three or fewer specimens were rejected on the ground that variances computed from so few specimens were likely to be very unreliable. The required variance ratio for significance at the 5% level for 2 and 100 degrees of freedom is 2.3 times that required for three degrees of freedom.

At this stage one species, S.archboldiana, was excluded from the species set since it was represented by a single incomplete leaf. It is worth noting here that the taxonomic status of this species must be questionable since the incomplete specimen in question is the holotype.

Variances were computed within species on 29 species for the following variable set. Between species variances were computed from 93 species. The variables, F ratios and significance are given in table 4.

No.	Character	F	Significance
1	Leaf length (L)	2.2234	.01
2	Leaf width (W)	2.8595	.01
3	L/W ratio	2.9778	.01
4	Apex angle 1	2.0245	.05
5	Base angle 1	2.8611	.01
6	Position of max. width	3.2082	.01
7	Percent. acuminate	2.9306	.01
8	No. of secondary veins	7.5240	.0001
9	Secondary intercostal shape	3.2827	.01
10	Secondary vein straightness	1.8236	.05
11	Secondary vein angle	3.0029	.01
12	Tertiary vein angle	2.5703	.01
13	Dentition frequency	5.7725	.001
14	Petiole length (% leaf L)	1.5455	.5

TABLE 4. F ratios for leaf characters

computed for 93 species of Saurauia

Of all the listed characters only petiole length does not have an F value within accepted statistical limits. It is, therefore, excluded from further analysis. The two characters, number of secondary veins and dentition frequency, have extremely significant F values.

Unique information content of characters

The F-ratio for characters is a measure of the discriminatory power of a character. In a set of characters such as the remaining thirteen of the list above however, it must be possible to determine the uniqueness of the information content of each character. This has already been considered in character correlations within S.conferta. For the most part correlations are explainable within a species by simple geometrical constraints. A species however represents a particularly homogeneous set of OTUs. The genetic influence on characters could be expected to be much greater between species. Therefore there is reason to expect lower correlations between characters in a set of species than within a single species. Correlation between characters reduces the available information and increases redundancy. For example, in a given sample leaf width is found to increase with leaf length i.e. larger leaves are also broader. Let us say the correlation is perfect i.e. a plot of species using length and width as the axes shows the points to lie on a straight line or smooth curve with zero scatter about the curve. It is clear that the information content of both characters is the same. Leaf width can be seen simply as a transform or rescaling of length. In this case one of the characters should be deleted from the character set. The reason for this can be seen in an extreme example. A set of characters made up of leaf length, leaf length squared, log leaf length, leaf length/mean length clearly contains no more information than length alone. The result of this is that length receives

an inordinately high weighting in the analysis of OTUs on that character set. The same argument applies to a set of characters where correlations are high but characters have different names e.g. length and width. Therefore, although the prime choice of a character set has been made on the basis of the F-ratio, further weeding of characters should be carried out on the basis of a correlation analysis.

Correlation between characters

A bivariate correlation analysis of 94 species was therefore conducted. This is accomplished by comparing the distribution of OTUs over pairs of characters and using regression analysis to determine the dispersion of points about a fitted line. In the present study regression analysis was accomplished using the REGRESSION subroutine of SPSS. SPSS (Statistical package for social scientists) (Nie et al., 1975) has been used extensively for analysis of data in this study and will be further referred to only by its initials.

Bivariate scattergrams of pairs of variables are illustrated in figures 98-104 (appendix 3). Correlation coefficients (r) are given for each pair of variables. The significance of these correlations can be judged from the fact that for 94 degrees of freedom an r value of .2 is significant at the .05 level. The distribution of correlation coefficients over the set of characters is illustrated in figure 42.

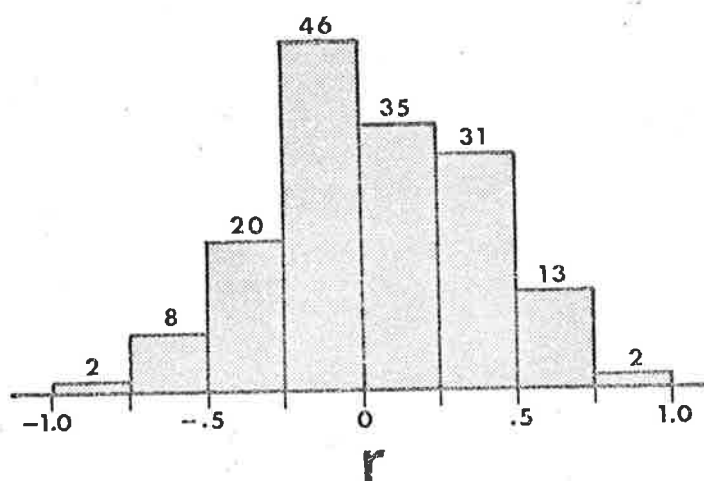


FIGURE 42. Distribution of correlation coefficients (r) for 13 characters scored on 94 species of Sauraula.

Clearly if the only criterion for accepting or rejecting a character on the basis of correlation was the significance of the correlation coefficient then most of the characters of the thirteen tested would have to be rejected. Correlation values can be seen as expressing the noncircularity of a scatter plot of OTUs on characters. Thus high r values indicate a very linear or elongate scatter, low values indicate that the scatter is essentially circular. Therefore as the sample size is increased even slight and visually nonapparent deviations from circularity become significant. A less rigorous method of rejecting correlated variables must therefore be found.

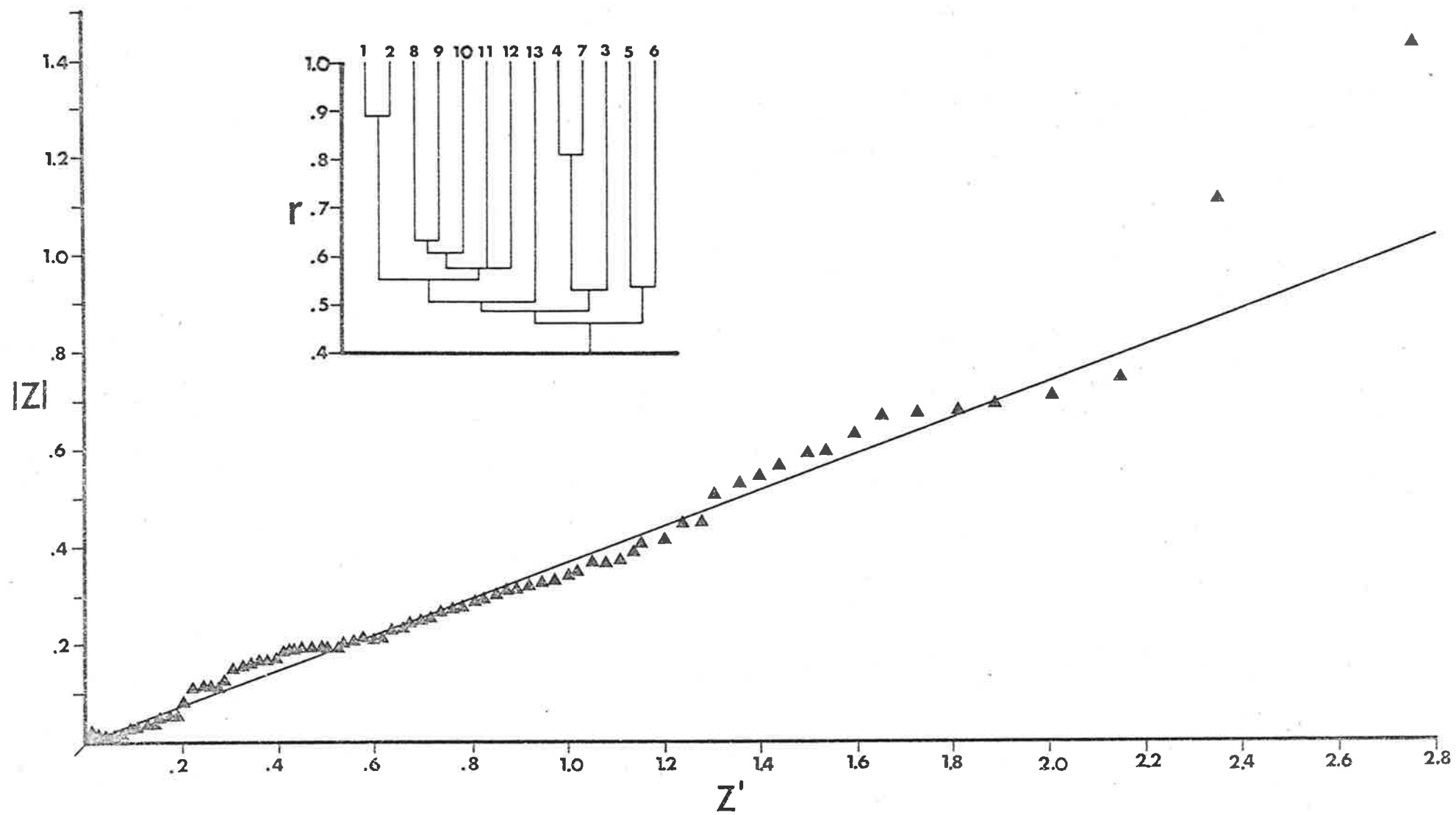
An approach to determining whether correlation coefficients could have come from the same population i.e. whether some values deviate markedly from the general trend of values, is to use the Z transform procedure of Hills (1969). This approach makes the assumption that the r values represent a population value of zero, and graphical methods are used to illustrate those values too large to make the assumption that they come from the same population. The technique of Z -transform plotting involves firstly transforming values of r to z using relevant statistical tables e.g. table 16 of Arkin & Colton (1963). Z values are then ranked ignoring sign as $|z_1| \leq |z_2| \leq \dots \leq |z_{78}|$, where 78 is the number of nonredundant correlation values in a pairwise comparison of 13 characters. The proportion of z 's numerically less than $|z_1|$ is approximately $(i-\frac{1}{2})/78$ if one assumes that z values are independent and therefore represent a normal population. That is z 's are from a distribution $N(0, \sigma^2)$, where each z_1 approximates $(\sigma z'_1)$ and the area of the normal curve for each z'_1 is equal to $(i-\frac{1}{2})/78$. Values of z'_1 can be read from a table of areas of the normal curve using $\frac{1}{2}(i-\frac{1}{2})/78$ to read off z 's.

Values of $|z_i|$ are plotted against z'_i . The points should lie about a straight line through the origin with slope σ . Figure 43 illustrates such a plot. The slope of .3702 deviates from the value of σ , (.3309) because clearly z 's are not independent. The plot illustrates that correlations with z values larger than .7 are too large to make the assumption that they come from a zero population value. The two widely divergent points represent correlations between variables 1 and 2, and 4 and 7, where the first pair have the highest z . Other correlations are ordered on the z plot by the level at which they separate on the phenogram of r values inset in figure 41. Note that the correlation of variable 8 and 9 represents the next point down the z plot. Therefore based on correlations there are reasons for rejecting one character from each of the variable pairs 1-2 and 4-7. These characters are respectively leaf length and width and apex angle and % acuminate.

The rejection of correlated variables must be considered in terms of other criteria before the choice of actual rejected character from a correlated pair can be made. F and r values have already been considered. The next criterion for judging the value of a variable is to determine the contribution of each variable to overall variance of the character set. The reasoning behind this is that variables contributing little to the overall variance are less important in defining the set of OTUs. The analysis of the contribution of each character is known as factor analysis.

Factor analysis on the set of thirteen characters was accomplished using the FACTOR subroutine of SPSS. The essence of this procedure is to use observed correlations between variables, to reduce the data to a smaller set of factors, which can be seen as the variables accounting for the observed correlations in the data. The

FIGURE 43. Z transformed correlation coefficients for 13 characters. Points deviating from the fitted regression line cannot be considered to come from a population $N(0, \sigma)$. Explanation of the plotting method is given in the text. The inset phenogram is plotted from correlation coefficients (r) between pairs of characters.



amount to which a particular character contributes to the factors is called common and the part of the character which is not influenced by other characters is called unique or residual. The amount of residual variance of a character is important in determining minimally redundant characters.

Where no a priori assumptions can be made about the underlying structure of the data, principal components analysis should be used. However, in the case of the variable set in hand some knowledge of the structure of the correlation matrix has already been obtained.

There are many alternatives available to a factor analysis. These involve estimating communalities (complement of residual variance), factor axis rotation and the choice of iterative or non-iterative procedures. The consideration of which final approach should be used is very involved and is considered in detail in both the SPSS documentation (Nie et al., 1975) and Sneath and Sokal (1973). The final procedure chosen for the present analysis will be briefly justified. Firstly factor analysis rather than principal components analysis was chosen for reasons given above. The second choice of method relates to estimates of communalities. Two procedures available in SPSS are PA1 and PA2 factoring where the first of these is non-iterative and the second uses iteration. The PA1 method uses an initial estimate of communalities to determine factors. The second method begins with an initial estimate and uses this to determine factors from which it extracts new estimates of communality and so on. This procedure is continued until stability is achieved in succeeding estimates. This latter procedure was chosen because it provides a more exact estimate of communality.

Other criteria over which the user of SPSS factoring has control, are rotation of factor axes and choice of the number of factors to be

extracted. Rotation of factor axes is performed to maximise the contribution of the first factor. Where no a priori assumption concerning correlation between factors can be made, orthogonal factor axes should be used. To illustrate, the first factor is often found to be primarily a size factor. That is size characters often contribute most to the first factor, while shape characters often contribute most to the second factor. Orthogonal and oblique rotations assume respectively no correlation and correlation between factors. Since in the present study there are no prior grounds for assuming correlation, orthogonal rotation was chosen. Of the techniques within SPSS for orthogonal rotation of factor axes, VARIMAX was used. This procedure maximises the variance of the squared loadings in columns of the factor matrix. This method tends to maximise the separation of variables on each factor to avoid having a number of variables loading heavily on the same factor.

The final choice governing the factoring procedure is limitation of the number of factors. If the number of factors is allowed to equal the number of variables there is no simplification advantage in using the factor analysis. Therefore a criterion should be found which allows the choice of a subset of factors. When the contribution of a factor to the total variance is no more than that of a single character then retention of the factor is of questionable value. The method of estimating this relationship is to use the eigenvalues of each factor. Where the eigenvalue of an initial factor is less than one, it can be said to account for less than the total variance of a single character and such a factor is therefore rejected.

Factor analysis using PA2 factoring with VARIMAX rotation and MINEIGEN equal to 1, was performed on data for 94 species scored on

13 characters. Five factors with significant eigenvalues were extracted and the factor scores of each variable on the five factors are given in table 5.

Character	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
1	.93005	-.04075	.07159	.29777	-.01934
2	.89451	.08852	-.26758	.26410	.09538
3	-.13313	-.15948	.87400	.02189	-.13707
4	.03612	.82718	-.48751	.13203	-.16500
5	.12867	.32791	-.37568	.09731	.69161
6	.00210	.09325	-.02398	-.13579	-.81978
7	-.28581	-.72329	.26164	-.02438	.12356
8	.17504	.18907	.59195	.76295	.10211
9	.32590	.05525	-.12065	.82769	.06876
10	.09497	.09216	-.00076	.69205	.13125
11	.16158	.55904	.22315	.35395	.39629
12	-.01341	-.55702	-.13553	-.07584	-.29271
13	-.53522	-.24970	.00000	-.02515	-.04793

TABLE 5. Factor scores for leaf characters using principal factoring with iteration. Characters are defined in table 4.

Alphanumeric labels refer to characters and are defined in table 4. Eigenvalues of the reduced factor set and their contribution to total variance are included in table 6.

Factor	Eigenvalue	Pct. of Variance	Cum.Pct.
1	3.98906	41.4	41.4
2	2.21119	23.0	64.4
3	1.52446	15.8	80.2
4	1.20839	12.5	92.7
5	0.70158	7.3	100.0

TABLE 6. Five extracted factors with eigenvalues, percent total variance accounted for by each factor and cumulative percent of variance. Three iterations used to achieve estimates of communality.

The plot of characters on the first three factors is shown in figure 44. Only a single plot is shown firstly because subsequent plots can be inferred from the factor matrix, and secondly the first three factors are significantly larger than subsequent factors. The location of characters on these first factors can be understood in terms of the correlation coefficients between pairs of characters. Characters one and two lie close together a long way out on the first factor, reflecting their high positive correlation. Characters four and seven are almost diametrically opposite at the ends of factor two reflecting their high negative correlation.

Referring back to the discussion of choice of minimally redundant high discriminatory power variables, factor analysis can now be put in context. Characters scoring high on the factors contribute a large amount to the total variance of the OTU set. Where this high variance contribution is accompanied by a high F value the character will be efficient at distinguishing OTUs. Accordingly figures 46 and 47 illustrate the relationship between factor contribution and variance ratio.

Overall contribution of characters to the five factors have been determined as follows. The columns of the factor score matrix (table 5) were multiplied by the corresponding percent of total variance that each factor contributed (table 6). Sums of squares of factor contributions (rows of the factor matrix) were then computed for each character. The reason for finding sums of squares is to remove the effect of sign on the factor contributions. Figure 45 illustrates a plot of the sums of squares ($\sum S$) against the order number of characters ordered by their S . It can be seen that all but two variables are linearly related with respect to S . The characters 1 and 2 (leaf length and width) deviate markedly from the fitted regression

FIGURE 44. Plot of the first three factors (third factor values in brackets) for 13 characters over 94 species using principal factor with iterations. Factor scores are after rotation using VARIMAX. Numbers refer to character codes given in table 4.

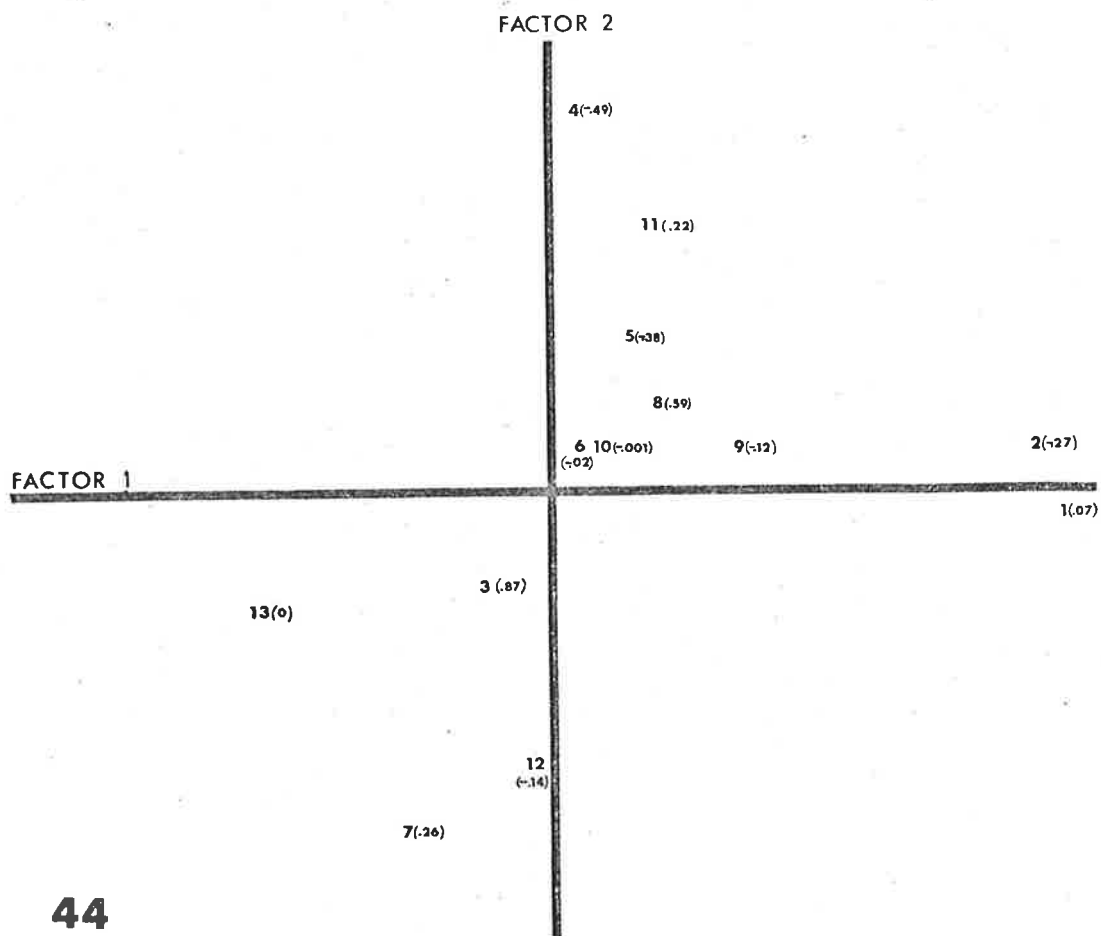


FIGURE 45. Leaf characters ranked by sum of squares of factor contribution (vertical) and plotted against actual sum of squares values, with regression line fitted to non-divergent values. Numbers refer to individual characters. For character names refer to table 4.

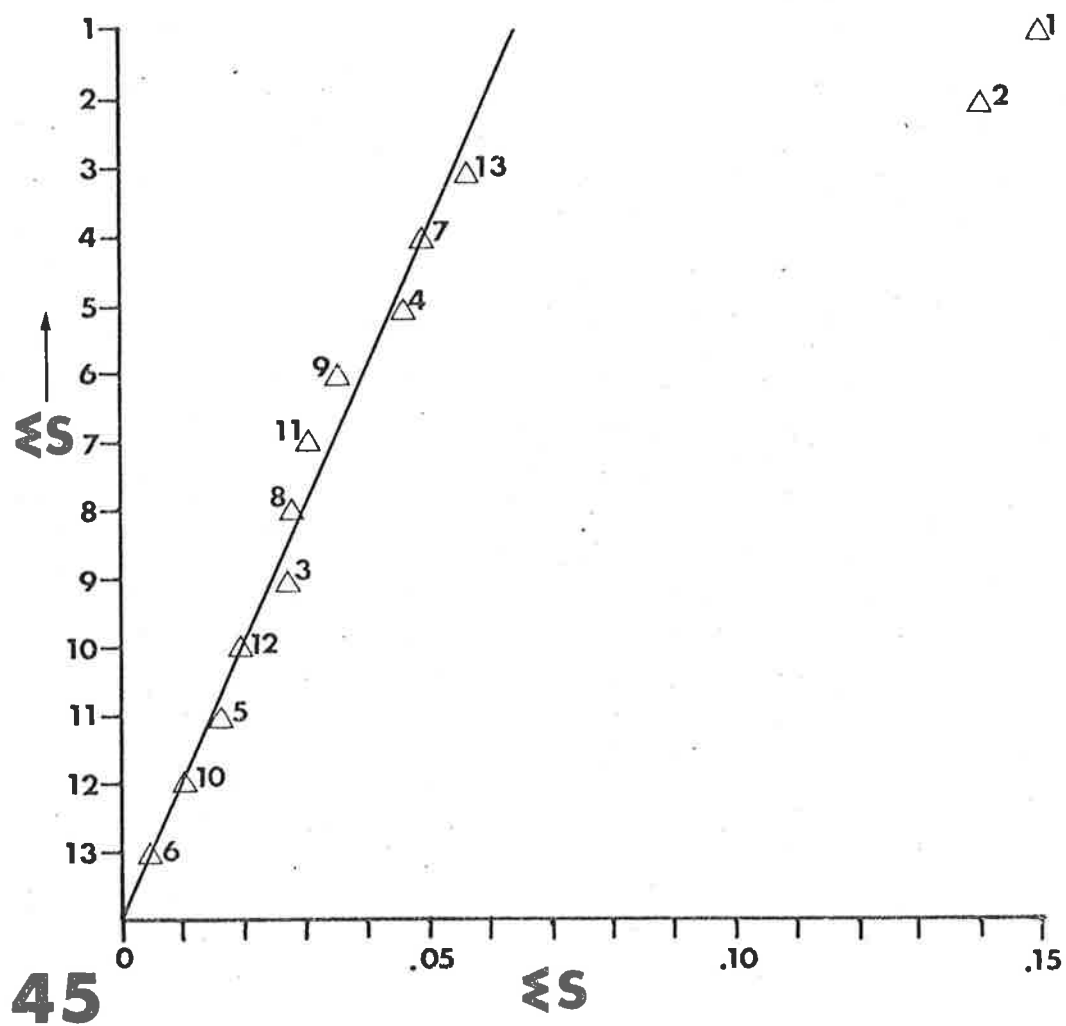
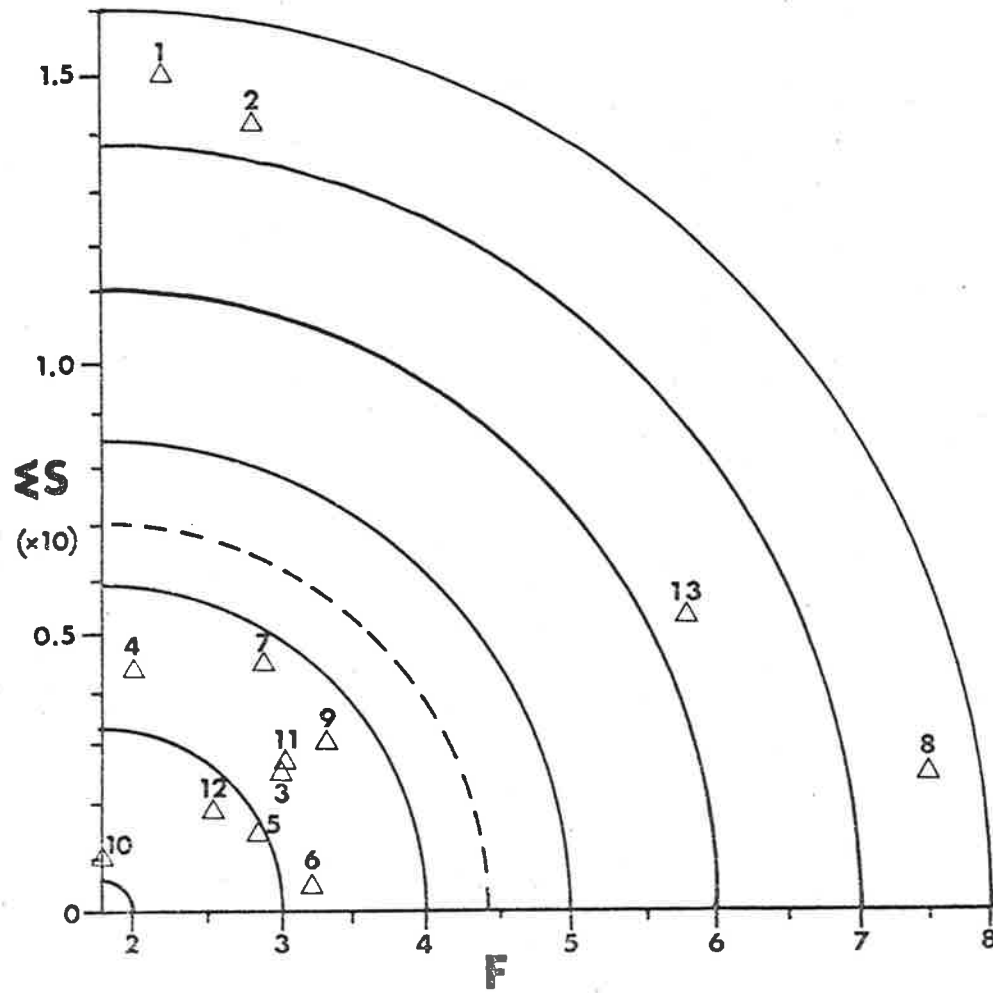


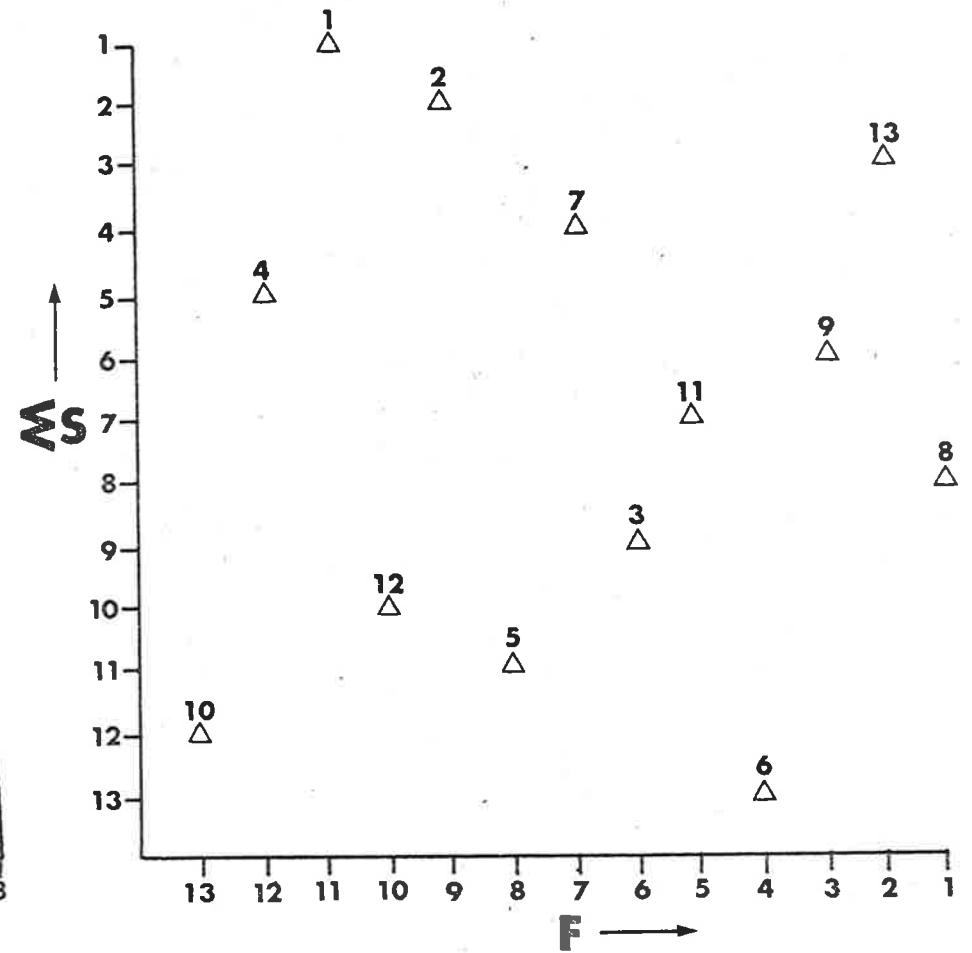
FIGURE 46. Leaf characters plotted with sum of squares of factor contribution (vertical) and F value (horizontal). Concentric circles illustrate how the value of a character is a balance between these two criteria. Numbers refer to character code numbers as given in table 4.

FIGURE 47. Characters ordered by both sum of squares (vertical) and variance ratio (horizontal). Point close to the origin represent least reliable and discriminatory characters. Numbers refer to character codes.

46



47



line. This extreme deviation supports the idea that leaf size is the most variable character in the set of species. Figures 46 and 47 put this variability in context with variance ratios. Figure 46 is a plot of F against S and it shows how the four characters of leaf length (1), leaf width (2), dentition frequency (13) and number of secondary veins (8) lie well clear of a cluster of the other characters around the origin. The high variance ratio and moderate factor contribution of character 8 supports the fact that this character was a deciding factor in the choice of the genus for study (see chapter 1). Figure 47 orders the characters by both sum of squares and variance ratio.

It is now possible to decide which characters can be excluded from further analysis. For reasons of correlation, one of characters 1 and 2 were removed. Each of these contributes approximately the same to the total variance so that the final choice must be made on F value. Concentric circles on figure 46 illustrate how character choice at this stage is a payoff between factor score and variance ratio. Figure 47 clarifies the relative positions of the characters ordered on these scores. On this basis, character 1 (leaf length) has been excluded from the character set. The reason for this is that leaf width has a much more satisfactory F value. Leaf length is also often not measurable on fossils. The second decision relates to the exclusion of one of apex angle and % acuminate (4 and 7 respectively). On the basis of figure 47, character 4 was excluded. Although no other correlations were found to be very high, it was decided that character 10 (2° vein straightness) contributed so little to each factor and had such a low F value that there was little reason to retain it. The final set of architectural characters is therefore as given in table 7. The coding prefix A will be added here to

distinguish character code numbers.

Character code	Character
A2	Leaf width
A3	Length to width ratio
A5	Base angle
A6	Position of maximum width
A7	Percent acuminate
A8	Number of secondary veins
A9	2° vein intercostal shape
A11	2° vein angle
A12	3° vein angle
A13	Dentition frequency

TABLE 7. Final character set with character codes.

At this stage some final comments regarding the character set are in order.

Characters which are ratios of other features are very commonly used in numerical taxonomic studies. The present study is no exception. Ratio characters in the list above are A3, A6, A7, A9, and A13. That is, some absolute measurement, either length or frequency, is standardised for all these characters by dividing by an absolute length value. There have been objections raised by Atchley et al. (1976) who suggest theoretical reasons for why ratios are unsuitable in a statistical analysis. The main objection is that ratios are generally used to remove size effects so that absolute measures of shape for example can be compared. Atchley et al. contend that this does not in fact remove the influence of size and indeed may enhance it. In the present study it was considered that if a ratio exhibited no significant empirical correlation with the size character whose effect was intended to be removed, then the ratio could be justifiably used. In only one character were the correlations of the ratio with the contributing variables similar in magnitude to the

correlations between contributing variables. That character was A9, secondary vein intercostal shape. It was felt that except for this slightly questionable character the use of ratios reduced correlation within the character set.

Lastly some consideration of inter-relationships i.e. correlations between characters, has been given to the architectural features of Hickey (1973). Dolph (1976) noted that a number of Hickey's characters were correlated in a set of leaves from 24 species over 20 genera of extant angiosperms. Of the characters that Dolph used, four were two state, ten were non-hierarchic multistate, and the remainder were hierarchic multistate, though not necessarily with equal spaced classes. The analysis used to elucidate the relationships between the characters was firstly computation of distances between pairs of variables and clustering using the UPGMA algorithm and then a principal components analysis. I have discussed in chapter 2 the influence of non-metric and in particular binary and non-ordered multistate characters on analyses which assume linear metric characters. It is interesting to note that of the five most closely related characters in Dolph's analysis of distance, four are binary characters and one is a three state character.

Even though the characters were standardised in that study, the fact that binary characters are almost certain to not have equal frequencies in the two classes, means that on the whole, distance will be reduced. This is due to the fact that more than half of the comparisons will be on scores within the greater frequency class.

The need for a careful selection of characters by the various criteria discussed in this chapter cannot be overstressed. It is also essential to select a set of characters applicable to the analyses that one wishes to conduct. It is my belief that a relat-

ively small set of characters whose inter-relationships are well understood and whose discriminatory power is maximal is superior in many cases to a large character set. This is especially true of those cases where the character set has been chosen purely for its "largeness," with no attention to details of individual characters. Even though a large character set may produce a stable classification, that classification may be quite different from the underlying natural grouping of the OTUs because of poor selection of characters. The work of Dolph (1976), although influenced by misapplication of numerical procedures, illustrates an important fact. That is, the set of architectural features described by Hickey (1973) should not be allowed to form the basis of a leaf data banking system without a detailed appraisal of their suitability for classifying angiosperm leaves. Dolph demonstrates that the resolution achievable with these characters is negatively correlated with sample size. In the present study it has been shown that a combination of metric characters and a multistate distance measure has increased resolution with increasing sample size. This matter will be considered further in a later section.

CHAPTER 6

TAXONOMIC PARTITIONING OF THE SET OF SPECIMENS

Introduction

Detailed leaf architectural data for the present study has been obtained from a set of approximately 330 herbarium specimens over 94 species of Saurauia. Because there is no general classification of the Old World species in a form suitable for identifying specimens, species names on herbarium specimens were for the most part accepted. Where a set of specimens was clearly morphologically heterogeneous the specimens were divided into homogeneous subgroups. Attempts were then made to identify these groups to the correct species. This was generally not possible, so that only general affinities were deduced. In such cases the species have been designated "aff."

Type descriptions were obtained for most species seen as herbarium material. In such cases the specimens were compared with the description to verify if the name applied was reasonable.

Splitting of specimens described as the same species if necessary, was performed subjectively. Only in one complex of species does there appear to be serious doubt as to identity of specimens. That group includes the species S.andreana, S.bifida, S.dielsiana, S.dufaurii and S.schumanniana, all members of the section *Setosae* and all from the New Guinea, Solomon Islands, Australian region. The final list of species is described in Appendix 3.

Sample sizes for species

The question of sample size has been considered in the previous chapter. Figure 41 illustrates the distribution of number of specimens per species against number of species. The rapid decline in frequencies indicates that most species were represented by only one or two specimens. This reflects in part my inability to obtain

collections from many herbaria, but also reflects the state of collections of Saurauia. Many of the specimens seen at herbaria represented the only collections of species.

Although a large sample of specimens for each species was desirable for a reliable numerical study, it was felt that a large sample of species was also important. Many specimens per species stabilise the location of individual species within a classification. Many species, however, are necessary to produce the general ordination needed before individual species can be put in context. For these reasons it was decided that small numbers of specimens could be tolerated. There is also a technical reason for not being perturbed by limited specimens. The metric leaf characters used in this study are generally not easy to measure. For the most part measurements were taken from photographs. If one includes photographic processing time, then approximately one quarter of the total time spent on the project was involved in data collection. This appears to be a reasonable figure especially if one considers that the time spent in analysis of data increases in proportion to the square of the number of data items.

The sections of the genus

Mention has been made in several places in this thesis of the sections of the genus defined by Diels (1922). Diels' classification of generic subgroups is based on Old World species. A classification of New World species was proposed by Buscalioni (1912) but will not be considered further here.

Diels recognised divisions between groups of species primarily on features of the inflorescence. He described ten sections, but these have been resolved into four major sections, Uniflorae, Ramiflorae, Calyptratae and Pleianthae, which are defined solely on

the inflorescence (Gilg and Werdermann, 1925). The last of these sections is itself subdivided into seven sections, based mainly on features of the indument. The sections of the genus as defined by Gilg and Werdermann are as follows (my translation from the German).

Section 1 Uniflorae

Flowers more or less solitary on the leafy branches.

Section 2 Ramiflorae

Flowers single or in simple dichasia in panicles, usually on leafless branches.

Section 3 Calyptratae

Flowers single on leafy branches; flowers before anthesis enclosed in a cup-shaped bract.

Section 4 Pleianthae

Flowers in simple or compound (sometimes strongly reduced) dichasia, which are sometimes united in large panicles.

A - Flowers not enclosed in a bract (however there are a few species in which leafy bracts sometimes enclose the buds so that there is a transition with Bibracteatae and Obvallatae).

i Undersurface of leaf not tomentose.

(a) Squamulosae - Leaves and young branches bearing obtuse scales or tufted hairs.

(b) Setosae - Leaves and young branches bear lanceolate scales or setae.

(c) Armatae - Leaves and young branches beset with prickles.

ii Undersurface of leaf tomentose.

(a) Obtectae - Leaf tomentum interwoven in a close appressed layer.

- (b) Rufae - Leaf tomentum more or less woolly.
- B - Flowers enclosed in a leafy bract.
- (a) Bibracteatae - Bracts and sepals finely tomentose.
- (b) Obvallatae - Bracteoles and sepals beset with straw-like chaffy scales.

For convenience these sections have been coded with numeric or alphanumeric labels. Pleianthae is considered as its separate subdivisions. The coding used is as follows:-

Section	Label
Uniflorae	1
Ramiflorae	2
Calyptratae	3
Pleianthae	
Squamulosae	4A
Setosae	4B
Armatae	4C
Obtectae	4D
Rufae	4E
Bibracteatae	4F
Obvallatae	4G

There has been some criticism of Diels system on the grounds that it is largely artificial (Burt, 1936), and that groups of species in sections are often heterogeneous. In some cases this has been found to be true. There are a number of species whose inclusion in one of the Uniflorae or Setosae is questionable. The composition of the Ramiflorae is also questionable since some species have flowers both in the axils of leaves and on the old wood. In general, however, it has been found that once one becomes familiar

with the range of variability in the genus, it is possible with a high degree of consistency to assign species to the sections. The sections with the greatest degree of overlap are the Squamulasae-Setosae, Uniflorae-Setosae, Ramiflorae-Setosae and Obtectae-Rufae. It is not surprising in the light of evidence offered in chapter 1, to find Setosae common to several of these pairs. The relationships between geography and generic sections discussed in chapter 10 support the fact that correct assignment of species to generic sections can be accomplished. Diels' sections as modified by Gilg and Werdermann are therefore retained for this study. For the sake of biogeographical discussion in chapter 10, two sets of alternate divisions are recognised however. The first of these groups species by a set of geographical regions delimited by natural barriers, such as seaways or lowlands. The second grouping recognises a major division in the genus in the number of styles that species possess. The question of variability of style number within species has been considered by Hoogland (1977). However, the ordination of species by style number is shown in chapter 10 to be more or less consistent with that based on generic sections. It is accepted, therefore, that some species regularly exhibit variability in style number and that all species may vary to some extent. However, it is also pointed out that essentially, species can be grouped into those with three or five styles.

Specific assignment of specimens

The question of grouping specimens to species has already been briefly mentioned. It is not the purpose of this introductory chapter to consider in detail the grouping of specimens. That detail follows in chapter 7. To remove the subjectivity of groupings or at least to minimise it, I decided to conduct a prior analysis of the

whole set of about 330 specimens, recognising the individual specimens as separate taxonomic entities. Therefore in the following chapter, unless specifically stated, the term OTU will refer to a single herbarium specimen, independent of its taxonomic affinities.

CHAPTER 7

APPLYING MEASURES OF RESEMBLANCE TO THE SET OF OTUs

The Multistate similarity/distance measure

The development of a multistate resemblance measure has been discussed in detail in chapter 2. Either as a similarity coefficient (I_D) or as a Euclidean distance metric (D_I), this measure is able to accommodate single or multiple scores for character states. In chapter 8 its application to single and multiple value, normal and randomly distributed data will be demonstrated. It is relevant, therefore, to consider some influences on values of resemblance.

Missing values

Commonly in biological data one finds that OTUs have one or more characters for which no data can be scored. This generally comes about in one of two ways. Firstly a character is inapplicable to the OTU under consideration. For example, hair type is an inapplicable character if one is investigating fish. Where a character is applicable to only some of the members of a set of OTUs, it is considered as a secondary character and should be excluded from an overall analysis. The second type of missing data occurs when that organ or part thereof, on which the character is scored, is missing from the specimen in question. Where an OTU is represented by a collection of specimens this can often be overcome by sampling. However, particularly in the case of palaeobotanical studies, OTUs are often single organs such as leaves, about which no prior taxonomic judgements can be made. In such cases where an OTU has a missing character, that character should not be included in comparisons involving the OTU.

Although there are theoretical reasons for the approaches to

handling missing data, it is worth demonstrating the empirical influence of missing values. Therefore a set of hypothetical OTUs has been chosen to illustrate the extreme situations. Both pairs of OTUs are considered as having twenty characters, with ten states for each character. In both examples OTUs can possess only single values on each character.

In the first example, two OTUs are as different as possible on the set of characters i.e.

OTU i	1	1	1	.	.	.	1	1	1	1
OTU j	10	10	10	.	.	.	10	10	10	10

where numbers refer to character state scores.

To assess the effect one must compute the value of D_I for this situation and then progressively delete one character at a time from OTU j. If D_I values for each step of this process are computed without excluding the characters with missing data from consideration, then the curve with positive slope in figure 48 results. The vertical axis of figure 48 represents the difference between the number of states possessed by the two OTUs. Thus where j has no missing data the difference is 0 and the corresponding D_I value is 10. As the amount of missing data is progressively increased D_I becomes smaller in value and eventually when OTU j has no character scores, becomes 1. If characters with missing values are deleted from comparisons then the value of D_I remains stable at 10 (marked by an X).

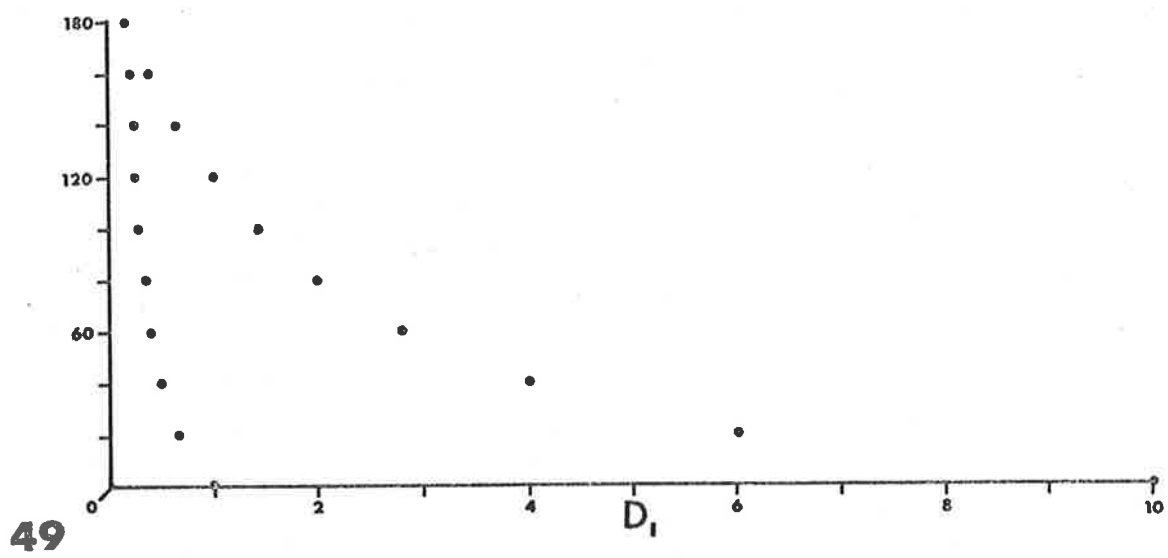
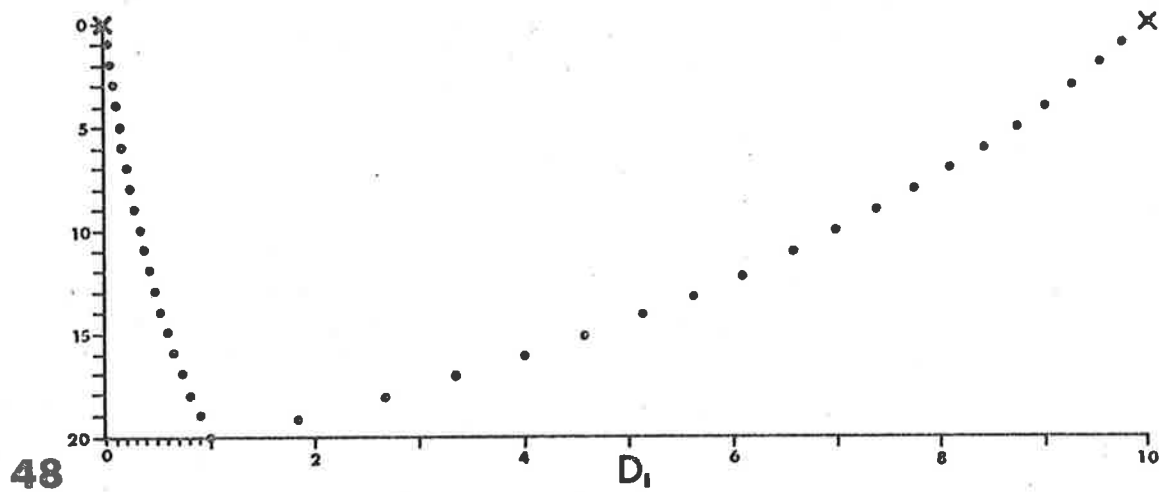
In the second example, two OTUs are identical on all their character states, i.e.

OTU i	1	4	3	4	7	1	2	2	6	.	.	.	6	7	3	1
OTU j	1	4	3	4	7	1	2	2	6	.	.	.	6	7	3	1

The distances between state scores for the two OTUs in this example are zero. As the amount of missing data for j is increased, the D_I values increase from the base value of zero to a maximum of one.

FIGURE 48. Influence of missing data on distance values (D_I) computed for two hypothetical pairs of OTUs. The number of missing characters is plotted on the vertical axis. OTUs scored for twenty characters of ten states each.

FIGURE 49. Influence of using ranges rather than single values for computing distances for two pairs of hypothetical OTUs. Difference between total ranges for the OTUs is plotted on the vertical axis. OTUs scored for twenty, ten state characters.



The curve of negative slope in figure 48 indicates this situation. When characters with missing data are excluded from comparisons the value of D_I remains at zero.

The rapid shift in computed distance values as the proportion of missing data is increased and when missing values are incorrectly handled demonstrates the importance of the correct treatment of missing data. This can be simply stated as:- when an OTU has missing data for one or more characters, those characters must be deleted from all comparisons involving the OTU.

Characters scored as ranges

In general OTUs are scored as single values for each character. Where multiple values are scored for characters, D_I can be used to compute distances between OTUs. Two extreme examples of the situation where one OTU (i) possesses only single values for each character while the other can possess any number of states will be considered.

Two test OTUs are considered where both possess twenty characters of ten states each. OTU j is considered in all situations from ten states per character to one state per character, whereas OTU i has only a single state per character. In the first example, OTUs i and j are matched on every character. As the range of j values is decreased incrementally the computed value of D_I follows the steeper curve of figure 49. Where j has the maximum number of values for each character, the value of D_I is at a minimum. Where j has, like i, only a single score per character the D_I value is one. In the second example, the range of j values is incrementally decreased. However in this case the values of character scores for i and j are separated as widely as possible. The steps of this procedure are illustrated in the following example.

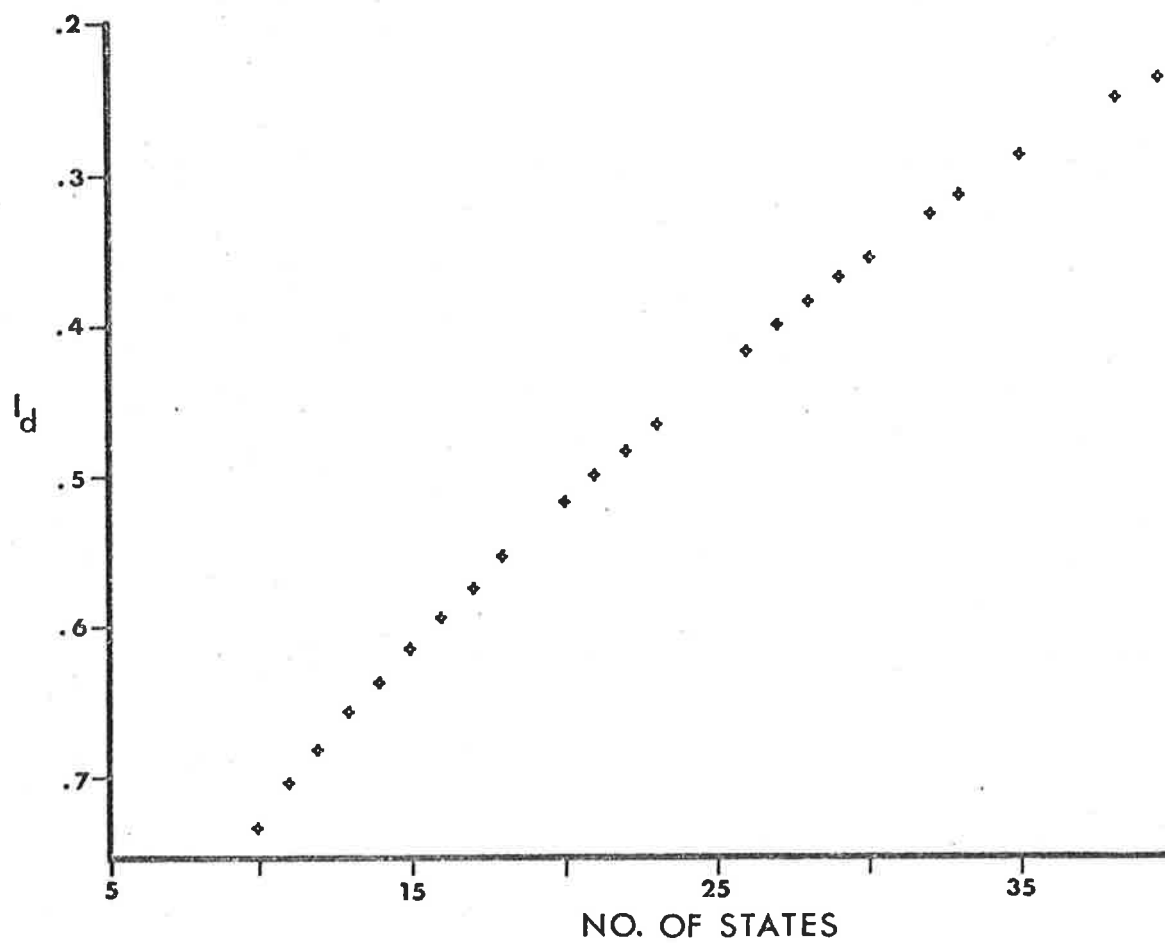
$$\begin{array}{ll}
 (1) \quad \begin{array}{c} j \quad ++++++++ \\ \hline i \quad +----- \end{array}, & (2) \quad \begin{array}{c} j \quad -+++++++ \\ \hline i \quad +----- \end{array}, \\
 (3) \quad \begin{array}{c} j \quad --+++++++ \\ \hline i \quad +----- \end{array}, \dots, & (10) \quad \begin{array}{c} j \quad -----+ \\ \hline i \quad +----- \end{array}
 \end{array}$$

The upper curve of figure 49 is plotted for this last series of examples. The vertical axis of figure 49 represents the difference in number of states for the two OTUs. When each OTU is allowed only a single value for each character and the separation of the scores on characters are increased incrementally, D_I takes integer values along the horizontal axis. The deviation of computed D_I for multiple scores, from that for single scores, can be seen to be related to the magnitude of the difference in the number of scores for each OTU. Where the number of states scored for OTUs on a character is likely to differ, the use of single values is recommended in general for analyses of similarity.

The range dependence of similarity values was tested empirically on a set of fifty-seven Saurauia species, scored for leaf data. These species were tested against a hypothetical OTU which ranged over all states of every character i.e. it exhibited sixty-three states over the nine characters. The distribution of I_D values is illustrated in figure 50, which graphically demonstrates the profound influence that ranges may have on apparent similarity.

The handling of variability of OTUs has been considered by McNeill (1974), who recommends that instead of using single values, such as means of "chance exemplars" i.e. single specimens chosen at random to represent the species, one should use measures of dispersion. Such measures include ranges, or character state frequencies. The latter method has already been incorporated into the present study by using randomised or equal frequency class intervals in the

FIGURE 50. Dissimilarity index values computed for 57 species of Saurauia scored as ranges on leaf data and compared to a test OTU which has a range over all states of all characters, to illustrate the influence of range on similarity.



final analyses. The use of range, whether a simple range or a measure such as the standard deviation, is not recommended for general taxometric studies.

The use of multiple measurements for characters of individual OTUs should be restricted to the following cases. (1) When an OTU is found to have a metric character which lies close to an arbitrary class boundary, then the error involved in placing it in a single class may be quite pronounced. For example, if another OTU lies barely on the other side of the boundary, then the two will be considered more different than a pair which lie at opposite extremes of the same class interval and which may be numerically further separated. In such a case the perturbing effect of using a range of values with a distance measure may be more desirable than the gross misclassification that may otherwise result. (2) When OTUs are scored for qualitative characters with non-exclusive states, multiple assignments may occur. For example, a leaf may possess two hair types. Sneath and Sokal (1973) and other workers have considered the question of using such characters as a set of two-state characters where each state is given character status. The effect of two-state characters on distance and similarity measures has already been considered in this thesis. It is recommended that the effect of multiple values is likely to be less serious than the use of a set of intersecting two-state characters. (3) When a range of values of states for an OTU expresses some feature of the whole species variation, its use is recommended over single values such as a mean, or that derived from a "chance exemplar." Meaningful ranges of states can be defined in terms of the variance, or a studentised range, or a confidence interval of the mean, or even an absolute range. It should be noted that such multiple measurements must be estimates of species para-

meters. They should therefore be based on an adequate sample of OTUs so that a specified degree of confidence can be placed upon them. Where the range of values is not consistently a good estimate through the set of OTUs, such as when samples sizes vary markedly, the use of ranges should be treated with caution. The influence of differing ranges on the distribution of similarity values is convincingly demonstrated in figure 49. Even where population parameters are used to express the variability of OTUs, it is necessary to be aware of two influences. Firstly using multiple values reduces the variance of similarity values. This may be acceptable since it can be contended that similarities based on means have unreal variances. Secondly, the similarity of OTUs tends to be related as much to their absolute variability as it is to the values that they take on characters. This is especially true as the relative "within-species" variability becomes large.

Distribution of data

The underlying distribution of the data set may also have a profound influence on similarity values. For example, in the situation where character states are independent, then the distribution of distances between OTUs based on such characters approaches normality (central limit theorem). The assumption of independence of character states cannot generally be made with biological data. It is often found that biological measurements follow a more or less skew normal distribution. It is easy to see that distance measurements on such a data set will have considerably reduced variance. Thus for a data set with an underlying non-random, unimodal distribution the distance values for pairs of OTUs will generally be underestimated. In later sections considerable attention is given to the relative ordinations of OTUs by D_I on random and normal data. Further

discussion of randomising data is given in chapter 2.

Recommendations for minimising distortion of distances

The following recommendations are made for using any measure of similarity or distance.

(a) Where an OTU has missing data on one or more characters, those characters must be deleted from any comparisons involving the OTU.

(b) Where OTUs are optionally scored for single or multiple values for characters, single values are to be preferred when the number of scored states is likely to vary. This is qualified in those cases discussed above.

(c) The effect of the underlying distribution of the data should be appreciated. When prior weighting of character states is intended, character states should be transformed, to achieve equal frequencies of OTUs in each state.

TAXON, a computer program for evaluating distance/similarities between pairs of OTUs

Defining the treatment of various data parameters is the first step in generating an algorithm for computing measures of resemblance. The computational loads involved in evaluating the algorithm for distance are, however, very great in a large data set. For this reason a computer program has been prepared which will compute alternatively, similarity or distance values for a set of OTUs. The current form of the program also allows mixed data sets and multiple scores on characters. The program is written in FORTRAN 4.6.

I wrote the initial form of the program, which computes only similarities. In collaboration with C. Annells, Botany Department, University of Adelaide, a more complete version of the program has been written. The updated version, TAXON, computes distances and

similarities between pairs of OTUs on the following types of characters, (a) binary, as two state or alternate, (b) ordered multistate and (c) metric. Non-ordered multistate and matrix characters cannot be presently adequately handled by the program except by transforming them by methods discussed in chapter 2. Such characters can be input without transformation, but will cause some distortion in values of resemblance.

At present the program is limited to eight discrete states per character i.e. eight states per multistate character or a range of a maximum of eight values for an integer metric character. The reason for this limit is twofold. Firstly, the subdivision of continuous metric characters into discrete states has been recommended in chapter 2. Secondly, to simplify the manipulation of data matrices both by the user and the program, it was decided to use a standard data coding format for all classes of characters. This has been accomplished by giving a score of 1 when an OTU exhibits a character state, and a score of zero when the converse is true. By this means characters are scored by a string of 1's and 0's where the number of digits represents the number of states of the character. This method also allows multiple entries in each character. In practice the limitation to eight states has not proven to be a problem. Extension of the number of states allowed could be accomplished quite simply by expanding the computed GOTO statements (see program listing).

Size limits on the data set

There is no real limit to the size of data matrices that can be handled except that imposed by storage and time limitations on individual systems. In actual runs on a CDC Cyber 173 computer, it has been found that a set of 300 OTUs scored as single values per

character over fourteen characters with a total of sixty-five states, use about fifty seconds of central processor time and approximately 1.5 million words of storage. The time and storage requirements are increased if ranges rather than single values are used, but this is generally marginal. Loading and compilation of the program and a data set of 300 OTUs takes about one second and 115000B words. To minimise computation time and the quantity of output, only the non-redundant half of the similarity matrix is computed. The size of the output file is the most serious limitation to a user of the program. The print file produced without program listing for 300 OTUs is 44850 lines or approximately 700 pages long. This figure will exceed most installation defined print file line limits.

In general even the non-redundant half of a similarity matrix contains a high degree of marginally useful information. The reason for this is that similarity values are generally used to achieve some sort of grouping of OTUs, whether by a clustering algorithm such as UPGMA, or a phenogram generating routine, or by the plotting of nearest neighbour trees. In all of these graphical approaches, the data is reduced to a smaller number of dimensions than that needed to completely express the relationships between OTUs. The number of dimensions needed for plotting OTUs with Euclidean distance values, is one less than the number of OTUs. Clearly any plot of more than three OTUs is going to be a distortion of the multidimensional situation, by virtue of its being limited to three dimensional space. The most meaningful method of representing relationships between OTUs is to use a subset of dimensions, where distances on those dimensions are defined by the closest pairs of OTUs. This is the concept beneath all graphical approaches to illustrating relation-

ships. In a non-independent set of OTUs, such as species in a genus where there is a high degree of morphological homogeneity, it will be generally found that most of the species are linked into a cluster of phenogram etc. at a considerably higher level than the maximum distance value. Therefore, an option for deleting surplus values from the output file has been included in the program.

The DISCARD option is used to delete unnecessary cases from an otherwise excessively long output file. DISCARD can be set to any value between 0 and the maximum possible value of distance. All comparisons with a distance between OTUs greater than the DISCARD value will not be output. In the simplest case, where all characters have the same number of states, the maximum value of distance is one less than the number of states. In general it has been found that DISCARD value of $(0.2 \times \text{max. distance})$ gives a preliminary idea of the overall distribution of similarities. In most biological situations this value should include a high proportion of the first linking matches of OTUs.

Output

Output can be obtained in two ways. Either it is printed on a hard copy device or written onto a magnetic storage device. Where the output is relatively small e.g. from about twenty OTUs, hand processing of the data is possible in a reasonable time. In such cases only hard copy may be required. Larger outputs however, may require machine processing and so it is recommended that the output be written onto magnetic storage. Such output is then in a form suitable for inputting into various other programmes such as those available through SPSS.

The format of printed output is as follows.

OTU1 x OTU2, States 1, States 2, No.Chars., Assoc.Chars., Assoc. Coeff., Distance.

Each pair of OTUs occupies a single line. The meanings of the output headers respectively are, (1) names or labels for the OTUs being compared, (2) number of states of each OTU, (3) number of characters on which they have been compared, (4) number of characters on which they have at least one state in common, (5) value of the association (similarity) coefficient, (6) value of the distance measure.

The other output option writes output onto a file called TAPE 3. To retain this after execution of the program, it is necessary to include a set of control cards to catalogue or store TAPE 3. Output onto TAPE 3 contains only OTU labels and distance values.

Using the program

Program limitations

Up to eight states per character for eighty states over 200 OTUs can be handled by the program, in its present form. To alter these values it is necessary to modify two dimension statements within the program. The arrays NAME and ISAMP specify the maximum number of OTUs. The arrays NSTATE, ICON and ICAR specify the number of states and characters and should be set to at least the maximum number of states. The matrix KAR must be equal to (No. of OTUs x no. of states). Two other statements need to be altered if the dimension statements are changed. These are the read statement (READ 1001 etc.) and the DO statement (DO 100 etc.) at respectively lines 7 and 19 of the program (not counting comments). Both of these should be set to the value of NSTATE. Lastly, format statements 1001 and 1003 near the end of the program should be set to NSTATE.

Setting the DISCARD value

The DISCARD value is set to a proportion of the maximum distance value. This maximum is equal to one less than the average number of

states per character. The DISCARD value is set as the D parameter in a statement of the form $P = Dx((NS/NC)-1)$. D should be a positive real number less than or equal to 1. If D is set to 1, all values will be output. If it is set to 0, no values will be output. For large (greater than about 30) OTU lists, D should be set to a low value such as 0.2 for the first run to ascertain the distribution of similarity values, before committing to a large output.

Data format

The format of data is controlled by the first card of the data file. This card specifies the length of OTU identifying labels (names, numbers etc.), number of characters in the particular data set, and the number of states of each character. The length of OTU labels is set within the program at six characters. OTU labels can be any combination of up to six characters. If fewer are used, the remaining spaces are blank filled. If longer labels are used, the surplus symbols will be read as data and errors will result. The format of the data header card is:- AAAAAA1122222233111122 etc. The first character string of all A's defines the label field. Following this is the variable field, where each variable is defined by a numeric string of identical digits. The number of digits (1's, 2's or 3's) equals the number of states in each variable. After the third variable the numeric string returns to the sequence of 1, 2 and 3. Thus in the example above, the first variable has two states, the second has six states, the third has two states, the fourth has four states and the fifth and last has two states. Any number of variables can be entered on one or more lines, as long as the program statements mentioned in the section on program limitations are modified correctly. Data should fill all eighty columns of cards before proceeding to the next card. Data is entered as firstly the

six character label and a numeric string. The numeric string consists only of 1's or 0's where 1 represents the presence of a particular character state and zero, its absence. The following example illustrates the form of a line of data, corresponding to the header card in the previous example.

SAUR..1010000010100010

The OTU, SAUR.. has a single state in each variable, in this example all of them scored on state 1.

Run name

Particular runs may be labelled with a unique identifier up to any length within the FORMAT statement 1000. An example may be seen in the program listing.

Deck structure

The structure of the program and data deck is as follows.

Control deck and job card

End-of-record

Program

End-of-record

Data header card

Data

End-of-record

Within the program deck all cards to be altered for the various options and dimension changes are delimited by lines of "\$\$\$\$" and explanatory comments.

Program listing - TAXON

A Fortran listing of the program is given as Appendix 5.

Ordination of specimens by similarity

The study of the whole set of OTUs, represented by herbarium material, was commenced with no prior assumptions regarding the

status of individuals. However, so that the ordination of OTUs produced by the various analyses could be compared to that produced from type description data, only those specimens with species names or at least affinities noted on herbarium labels were incorporated.

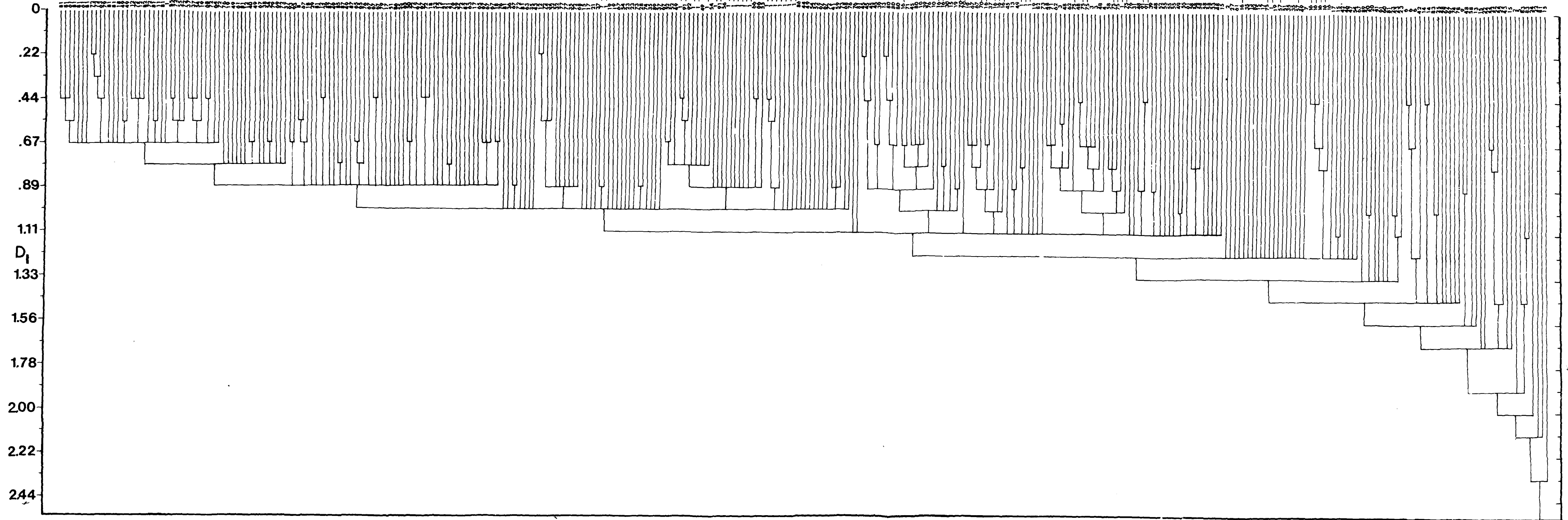
For the initial analysis, 332 specimens scored for the final set of 9 characters were analysed using the program TAXON. Data was not entered as continuous metric information however, but divided into a fixed number of class intervals. These were selected using the method described in chapter 2 (Sturges, 1926). For 332 OTUs, nine intervals are computed by this method. To avoid having to transform the whole data file for later analyses on about 90 OTUs, the number of intervals was selected for 90 values. This gives 7 class intervals.

The data for each of the 332 OTUs was recoded into seven class intervals. The class intervals were of equal width, and were computed using a simple range. The data matrix was input into the program TAXON for computation of distance values for pairs of OTUs. These were then used to generate a phenogram for the whole set of OTUs (fig. 51). A reduced matrix of similarity chosen by setting DISCARD to .5, was ordered by decreasing similarity, using the SPSS subroutine SORT. No suitable phenogram generating program was available at the time so that figure 51 was generated by hand from the ordered distance values. In the phenogram, OTUs are labelled with the sequence number of the species to which they were identified on herbarium specimens. This numbering differs from that used in the species descriptions in appendix 6. The actual species to which the numbers refer are not important in the present context, however. OTUs are also labelled with the generic section to which they belong. These sections are defined in chapter 6.

FIGURE 51. Single linkage phenogram generated from distance matrix computed for 332 Saurauia specimens. Numbers of OTUs refer to species to which the specimens have been assigned according to herbarium label identifications. Specimens are also grouped by generic section.

Uniflorae
Ramiflorae
Calyptratae
Pleianthae

a
b
c
d
e
f
g



The phenogram indicates that the whole set of OTUs are very homogeneous on the set of characters for which they have been scored. This is indicated by the general level at which OTUs are linked to clusters. The theoretical maximum distance value for the data set is 6. Most OTUs are linked above 1.1. The poor separation of OTUs is a consequence partly of the fact that the data is approximately normally distributed. This means that the variance of distance values is quite low, so that only those OTUs that lie to the extremes of the multivariate distribution, will be clearly distinguished in the phenogram. The second reason for the poor separation relates to the correlation between the first linkage information relevant to constructing the phenogram and the total information for all linkages of the similarity matrix. It is possible to test this by computing the cophenetic correlation coefficient between the distance matrix and the phenogram. Because the distance matrix contained 54,946 nonredundant entries, it was decided to reserve such testing for the reduced set of OTUs combined from specimens grouped to species.

A measure of correct ordination by the phenogram can be gained from the extent to which OTUs are grouped into "runs." If all specimens were correctly grouped by species, they would occur as runs of numbers where all the specimens of a species were adjacent in the phenogram. For figure 51, in 34.3% of cases, the OTUs are grouped by runs of two or more. Because the numbers of specimens in each species are not equal (fig. 41) such a measure of ordination does not give a completely accurate judgement of the phenogram. Of the 93 species incorporated into the phenogram, 60 have been scored on two or more specimens. If the number of specimens occurring in runs is expressed as a fraction of the number of those specimens for each species, the mean fraction gives a measure of the overall grouping of

specimens to species. In this case 38.2% are in runs. This still does not give the complete situation, as specimens separated by a single intermediate specimen of a different species are not counted as being in a run. However, it is clear that probably less than half of the specimens are correctly grouped according to the label names. It is worth noting that 61.4% of specimens are correctly grouped by generic sections.

Details of the phenogram

The most obvious feature of the phenogram is its homogeneity. However, a number of clusters of specimens are visible. For the most part these consist of only a small number of specimens. If the phenogram is studied in association with the generic sections to which the specimens are assigned, some regularity in the clustering becomes apparent. The first species to be linked to clusters lie on the left hand side of the figure. This area of the phenogram contains many small indistinct clusters, each of which tends to contain runs of species numbers. Most significant, however, is the fact that the most similar (first linked species) on the left hand side of the plot are dominantly members of the generic section 4b (Setosae). It had been noted that members of this section were generally the hardest to distinguish on herbarium sheets. They showed a high degree of morphological homogeneity and appeared to be frequently misidentified as one of the other species in that section. This region of clustering includes specimens of the species Saurauia andreana, S.bifida, S.dielsiana, S.poolei, S.purgans and S.schumanniana. These species in particular have proven difficult to separate. They have been the subject of an S.E.M. study to try and find features for distinguishing them.³

3. Green, P.S. Leaf hair morphology of the Saurauia schumanniana Diels complex. Unpublished honours project, Dept. Botany, University of Adelaide.

Several of the major clusters throughout the phenogram are dominated by members of the Setosae. Near the centre of the phenogram a set of specimens of the species number 1, S.conferta of the Obvallatae (4g), form a distinct cluster. The size of this cluster is certainly an artifact of the large number of specimens representing this species in the study. It is encouraging to find such a high degree of correct ordination (18/26 specimens). Clearly, however, there are a large number of specimens which do not show such a distinct grouping by species. Where only a single specimen of a species has been seen in this study, nothing can be said about its correct identification on the basis of the phenogram. Where two or more specimens of a species have been included, the ordination of those specimens can be used to make judgements on their identification. Distance values can be used to give a measure of the morphological homogeneity of groups of specimens.

Defining species limits

The instability of classifications of leaves based on single specimens has been demonstrated both in the previous chapter and in the literature by Dolph (1976). Dolph attempted to classify a sample of extant leaves by various clustering methods, using the character set of Hickey (1973). Although not directly comparable with the present study, it is worth making some comparisons between the results of Dolph and the results illustrated in the phenogram (fig. 51). In the present study, whole herbarium specimens were used. In general these bore several leaves. Dolph, however, used single leaves in his classification. In that respect one would expect a better clustering based on herbarium specimens if the same set of taxa were used. The reason for this is that mean values computed from a number of leaves remove the effect of within individual variance. The reliability of Dolph's classification was tested

by constructing alternate ordinations by various clustering algorithms and by computing cophenetic correlation coefficients to assess the relationship between the distance matrix and the phenograms. Using arbitrary significance levels a phenon line was constructed for each phenogram. This line represents the cutoff above which OTUs are considered to be the same species and below which they are separated. By this means the degree of correct identification of OTUs can be calculated. I consider that this defines a major source of error in assessing the reliability of phenograms. Such an approach makes the assumption that all OTUs in the set have equal within species variances. That is, clusters of specimens of each species are presumed to form the same size and shape clusters in attribute space. Subjective assessment of any set of leaves from a number of species will indicate that this is not the case. Both environmental and genetic factors influence the within species variance of leaf characters. Simply stated one must recognise that there are different species limits for different species. Figure 51 gives a good illustration of the errors that might result from using a rigid phenon line. The phenon line chosen for a rather average member of the Setosae (4b) S.andreana lies at just below 1.11 on the phenogram. This choice is discussed later in this section (see fig. 53). This species is number 81 in the figure 51. The first linked of these specimens are actually among the first linked to the whole phenogram. If one chooses the level of 1.11 as the general phenon line for all OTUs, the phenogram contains only 58 species. Inspection of the phenogram further reveals that several species, notably 6, 22 and 67 are linked to the phenogram below that level, yet the specimens of these species are considered to be correctly identified. Therefore, a flexible phenon line based on individual species is justified. It

is unfortunate that the subjectivity in this choice cannot be overcome in the present study. It was decided, however, that the errors resulting from such choice were considerably less than the absurd classification resulting from an inflexible phenon line. The value of a subjective assessment of a classification, lies particularly in the detection of incorrect assignment of OTUs which may otherwise be accepted.

A great deal of consideration has been given in the literature to the delimitation of biological species. The most interesting numerical approaches to this problem have been those of Legendre and Vaillancourt (1969), Rogers and Appan (1969), Crovello (1976) and Sneath (1976). Of these only the last mentioned distinguishes the position of morphological species in taxonomy. The other papers are primarily concerned with genetically defined species and so are not very significant where the aim is primarily morphological discrimination. In the present study, acknowledgement of "real" biological species is made in comparing alternate classifications achieved with leaf and type description (mainly floral) data.

Sneath (1976) considered the value of morphologically oriented approaches in delimiting species and applied various similarity based clustering procedures to illustrate this. No discussion of the kind of species philosophy that I have followed has been given, however. In fact, Sneath and Sokal (1973) make specific mention of the value of strict phenon lines in interpreting species discrimination by phenograms.

If one is working with a heterogeneous set of taxa such as the sample that Dolph (1976) used, chosen from genera from different families, the differentiation of species can be related to morphological disjunctions within the set of specimens. The prior defi-

nition of species should be of the form: A species is defined in terms of the set of characters under study, by a significant disjunction over the set of characters such that in terms of the total variation of those characters over the whole set of OTUs a distinct cluster can be defined. This simply means, where a set of OTUs consists of specimens of species, in an analysis which compares OTUs on the basis of the whole character set, disjunctions which arise in the distribution of values of resemblance can be used to define OTU limits. The significance of the disjunction with respect to the distribution of all values in the resemblance matrix will decide the status of OTUs so defined. Thus where there are real morphologically definable limits to species these limits should appear, if the relevant characters are chosen, in an analysis using a similarity or distance or correlation algorithm. Where OTUs form broad clusters these in general should be seen as morphologically variable taxa. A complication arises in cases where a group of taxa are closely related in the evolutionary sense. This is the case within groups such as large genera like Saurauia. The presence within this genus, of local broad clusters, more or less distinct from others in A-space suggests that there exist species complexes. These may be the result of current speciation within the genus, or the retention of a stable morphology from previous speciation, or a case of convergent morphology. The first case is probably the most reasonable when floral characters are also considered for those complexes.

The philosophy of seeking the morphological disjunctions has been applied in the present study to grouping specimens to species. It does not warrant consideration in specimen by specimen detail in the present context however. The ultimate result of grouping specimens by this procedure was that 94 distinct taxonomic entities could

be recognized on the basis of both leaf and floral characters. Decisions regarding the clustering of specimens were necessarily weighted by identifications given on herbarium labels. It is considered that a survey of the leaves of Saurauia using individual leaves as a starting point, would require considerably better sampling than that in the present study. Because sample sizes for the species differed very considerably in the present study, the ordination of specimens in the phenogram (fig. 51) was confused by those differences. No claim of the stability of the classification illustrated in the phenogram can be made.

The second method of improving ordination, after better sampling of OTUs, would be to improve sampling of the morphology of individual OTUs. This is achieved by using an increased character set. The choice of characters for application to classifying Saurauia has been considered in detail in chapter 5. It was not possible to justify the retention of a larger set of characters than the nine that were finally used. Therefore, to achieve an improved ordination of OTUs by leaf characters, the only approach that could be followed was to work with grouped specimens, clustered to maximise morphological disjunctions. The grouped specimens were then analysed using the program TAXON to assess their morphological relationships.

Before committing any further time and effort to an analysis based on the multistate measure of resemblance, using TAXON, it was felt that an empirical assessment of the distribution of resemblance values across several sets of OTUs was justified. The effect of variation in leaf characters in the set of OTUs on the values of resemblance, expressed either as dissimilarity index values (I_D) or as distances values (D_I) was not well understood. Experience with the phenogram of individual specimens suggested that these measures

might, on the character set in use, be inordinately sensitive to morphological variation within OTUs. To test the empirical distribution of these measures, a species of Saurauia which was both well represented in terms of number of specimens, and was regarded as a homogeneous species by the clustering procedure for defining species limits, was chosen. Fifteen specimens of S.andreana were therefore input into the computer program and values of similarity and distance were calculated. The relationship between these measures is illustrated in figure 52. The reason for this plot can be defined in terms of what each of the two measures can express with regard to morphological similarity.

The dissociation coefficient measures effectively only those situations where two OTUs have states in common on a character. In a multistate character all degrees of non-overlap of states of two OTUs are given the same value of dissociation when there are no states in common. This coefficient is therefore sensitive to morphological differences between OTUs only insofar as the OTUs have common character states. The distance measure conversely accounts for any degrees of overlap or non-overlap on characters of two OTUs. Therefore, the greater the degree of similarity between OTUs, the greater the correlation between I_D and D_I . This correlation has been measured on S.andreana and is demonstrated in figure 52. The value of the correlation coefficient is .941, a value of such significance that the plot could be used as a fairly reliable model for estimating either I_D or D_I from the other measure. The reliability of either of these measures as an estimator of the other is illustrated in a pair of phenograms constructed for S.andreana from alternatively I_D and D_I (fig. 53). The ordination of specimens within the phenograms are almost identical, with only specimen 14 in different positions.

FIGURE 52. Correlation between association coefficient (I_D) and distance measure (D_I) for specimens assigned to Saurauia andreana. Limit lines refer to theoretical limits to the relative values of the two measures.

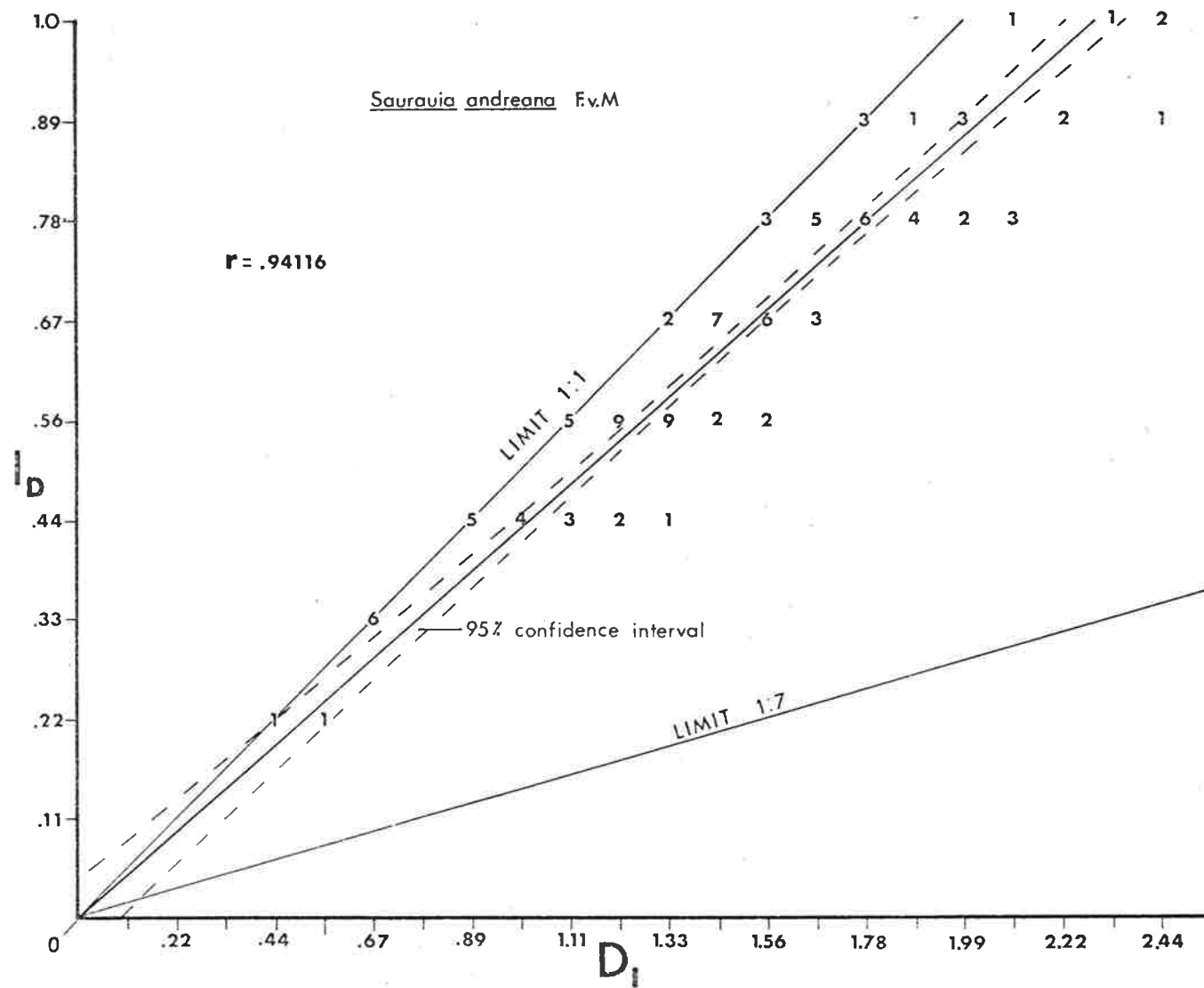
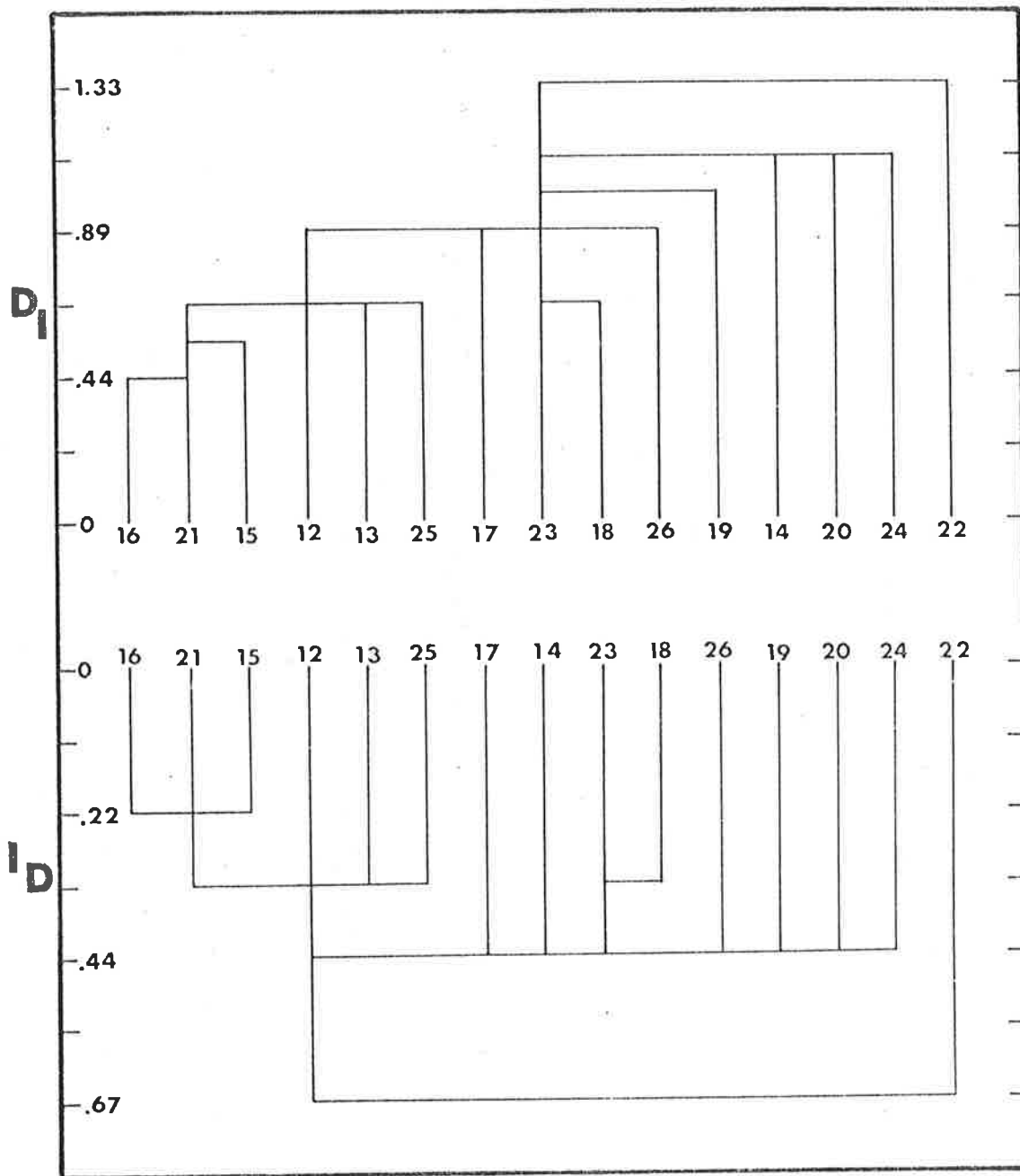


FIGURE 53. Single linkage phenograms for fifteen specimens of Saurauia andreana produced alternately by dissociation coefficient values (I_D) and distance values (D_I) on nine metric leaf characters. Specimen numbers refer to sequence numbers for the whole set of 332 specimens.

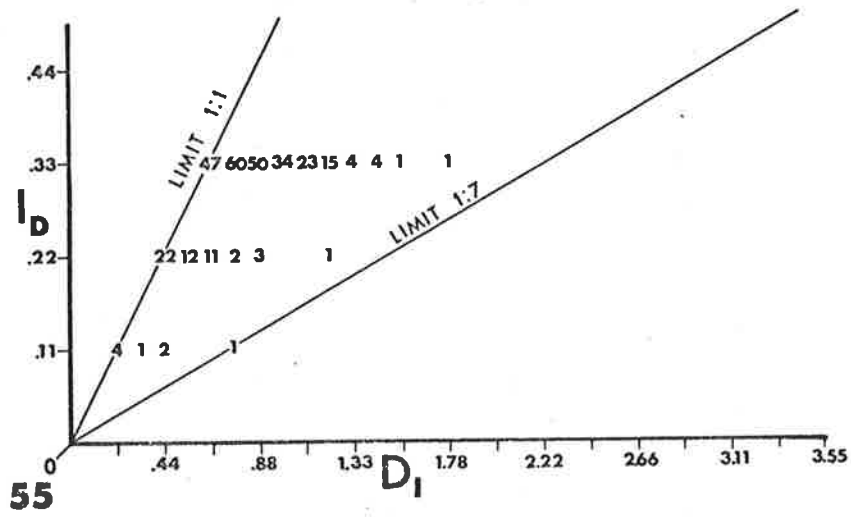
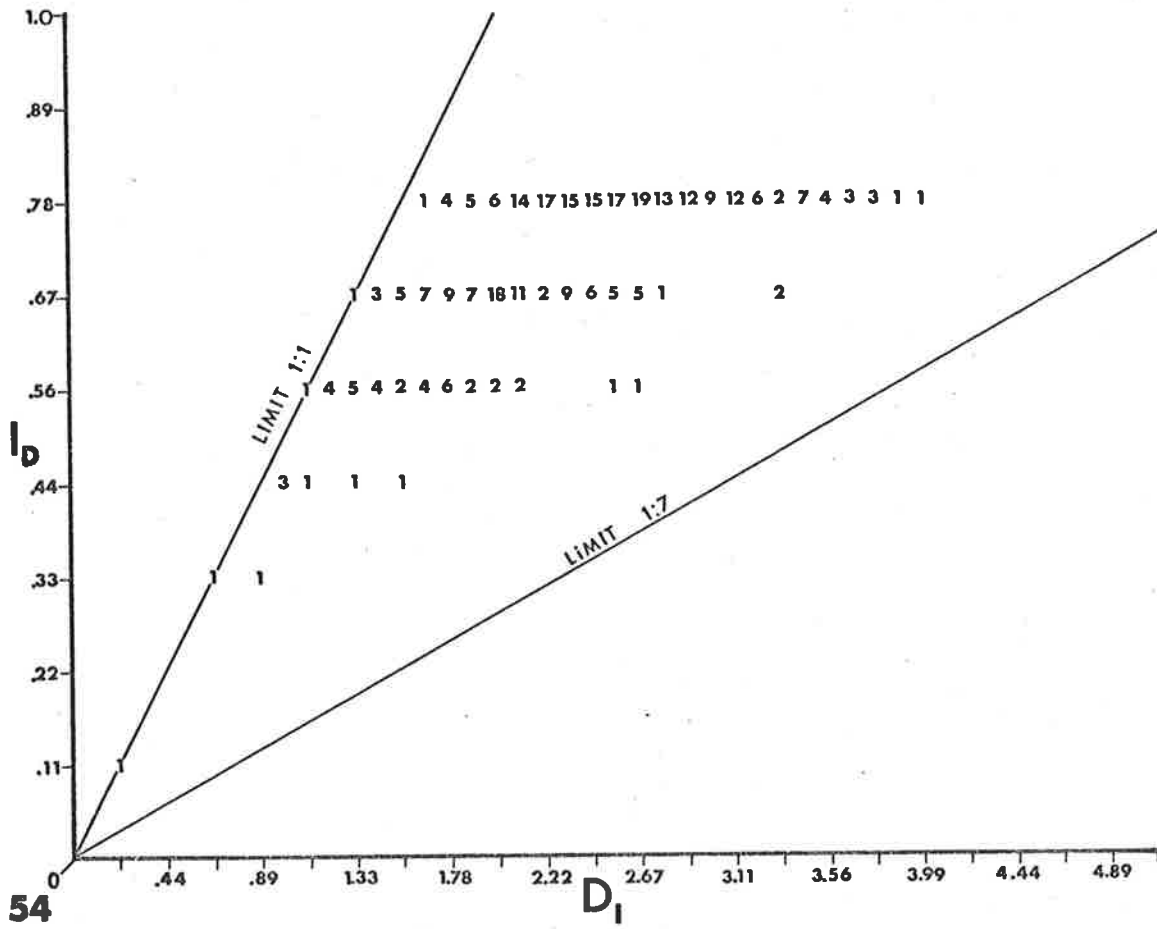


Specimen numbers refer to the sequence numbers of the specimens within the whole set of 332. Although the absolute values of distance and dissimilarity are different, the relative levels at which specimens are linked are very close. For this case then the computationally simpler I_D values could be used in place of D_I . To test whether this can be generalised to a larger OTU set, similar plots were produced for both 57 species scored on leaf data, and the set of 332 specimens (figs. 54 and 55). Limits shown on these and the previous such plot (fig. 52) refer to the theoretical limits of I_D and D_I values. Limit (1:1) refers to the situation where, when OTUs are mismatched on a character, the scores for each OTU lie in adjacent states. The limit (1:7) occurs when OTUs are mismatched such that they have scores on a character at opposite extremes of the character. Therefore, when the empirical values of I_D plotted against D_I approach the upper limit (1:1) this implies that the pairs of OTUs, when mismatched are only just so. Points lying at high values of I_D are also mismatched on a large number of characters. The limits of the plots refer to the situation in the present study where characters are divided into seven states.

Plotting relative I_D and D_I values for 57 species, figure 54 results. Only values of I_D less than .8 have been plotted. It is clear that in this figure, there is a greater spread of D_I values than in figure 52. This is to be expected since the morphological variation within the whole set of species is much greater than within a single species. The number of species chosen for this plot is less than the total number of 94 species scored for leaf characters. This subset has been chosen because of a need to compare leaf based classifications with those based on characters derived from type descriptions of the species. Of the 94 species scored for leaf characters,

FIGURE 54. Correlation between association coefficient (I_D) and distance measure (D_I) for 94 species of Saurauia, scored for leaf characters. Values of I_D greater than .78 have not been plotted.

FIGURE 55. Correlation between I_D and D_I for 332 specimens of Saurauia. Values of I_D greater than .33 have not been plotted.



only 57 were scored for type data. This was either because in a small number of cases the descriptions had not been obtained, or more commonly, because they were too inadequate to allow extraction of the relevant data.

The distribution of values in figure 54 is such that using I_D to estimate D_I , in only 5.3% of cases will the estimate be correct. Compare this to figure 52, for which 45% of the estimates will be correct. The value of correct estimation D_I for the plot of 332 specimens (fig. 55) is 49%. This very high figure illustrates the difference in overall morphological continuity between the set of individual specimens and the set of species, derived from the specimens and expressed only by mean values. The effect of grouping specimens to species and then truncating their variability by using only mean values is to increase the disjunction between OTUs. This is well supported by figures 54 and 55 and indicates that at least for the case of species data, a dissociation index is inadequate. Taking this inference a step further, it is reasonable to suggest that an analysis of specimens grouped to species will produce a classification in which the OTUs are more distinct than they are in the classification on specimens.

This chapter concludes the discussion of preliminary analyses. All the analyses discussed in the rest of the thesis are considered to be steps towards defining a final classification. Henceforth the term OTU will refer to specimens clustered into groups of morphologically similar entities which have been given species status. In all, 94 species have been defined for the final aspects of the analyses and 92 of them are included in leaf descriptions in Appendix 6.

Two general recommendations for leaf studies can be made on the basis of this chapter. Firstly, it appears that errors of assignment

of individual leaves or specimens to species are less, when subjective assessment of species limits, coupled with seeking morphological discontinuities within clusters is performed, than when single specimens are considered as OTUs. This may be particularly important in palaeobotanical studies for which a little subjectivity at ground level is better than a classification which cannot be applied beyond the set of specimens on which it was based. Secondly, unless there is some prior knowledge of the homogeneity of a sample, or else the character set is dominated by two state characters, it is better to use a distance measure than an association index for defining morphological entities.

CHAPTER 8

CLASSIFICATION OF SAURAUIA SPECIES

In this chapter the procedures which were followed to achieve the ultimate leaf based classification of the species of Saurauia will be elucidated. These procedures will also be discussed in terms of their application to a more classical taxonomic classification of the species of the genus. This was attempted on the basis of data extracted from type descriptions of 193 species. The relative ordinations of species on leaf and type description data will also be compared and discussed. To remove the need for repetition of standard phrases the following terms are defined. These terms are used throughout the text and also in figure legends.

1. TYPES

Species for which type description data has been obtained for a set of characters to be defined later.

2. SPECIMENS

Species which have been defined by the methods considered in the previous chapter. It will include only those species for which herbarium specimens have been obtained, and which have been scored for the set of metric leaf characters.

3. NORVARS

The underlying distribution of characters across the set of species, both TYPES and SPECIMENS, has been found to be more or less normal. This term refers to character data when it has been divided into class intervals such that the underlying distribution is preserved. In the case of metric characters, this has been achieved by using equal class intervals. In the case of non-metric characters no recoding of states has been carried out.

4. RANVARS

To minimise distortion of the distance measure, characters have been recoded such that there are equal scores in each character state. For metric characters this has been accomplished by the method described in chapter 2. Where possible, non-metric characters have had states combined to approach randomness.

5. RANGE

Ranges of states for characters have been used.

6. MEANS

This is applied where only mean values for species have been used in an analysis.

7. I_D

The symbol for the multistate dissimilarity index as defined earlier.

8. D_I

Multistate distance measure

Throughout this section a number of species of Saurauia will be referred to frequently, especially on figures. Therefore, a standard set of abbreviations has been used. These abbreviations were chosen mostly with data processing in mind and so have been kept as short as possible, concomitant with their readability.

Abbrev.	Name	Abbrev.	Name
ACTIN	<u>S.actinidifolia</u> Stapf.	ACUM	<u>S.acuminata</u> Merr.
ALPIC	<u>S.alpicola</u> Smith	BIFID	<u>S.bifida</u> Warb.
BRAC	<u>S.bracteosa</u> DC.	BRASS	<u>S.brassii</u> Diels
CAPIT	<u>S.capitulata</u> Smith	CAULI	<u>S.cauliflora</u> DC.
CLEME	<u>S.clementis</u> Merr.	CONFE	<u>S.conferta</u> Warb.
CONGE	<u>S.congestiflora</u> Smith	DIELS	<u>S.dielsiana</u> Smith
DUFAU	<u>S.dufaurii</u> (F.v.M.) Diels	EBURN	<u>S.eburnea</u> Smith
EGREG	<u>S.egregia</u> Diels	EMARG	<u>S.emarginata</u> White & Francis

Abbrev.	Name	Abbrev.	Name
FRAG	<u>S.fragrans</u> Hoogl.	GJELL	<u>S.gjellerupii</u> Lautbch
HOLO	<u>S.holotricha</u> Diels	LEPRO	<u>S.leprosa</u> Korth.
MACRA	<u>S.macrantha</u> Smith	MALAY	<u>S.malayana</u> Hoogl.
NAPAU	<u>S.napaulensis</u> DC.	NOVOG	<u>S.novo-quineesis</u> Scheff.
NUDIF	<u>S.nudiflora</u> DC.	OCCUL	<u>S.occulta</u> Smith
PANAY	<u>S.panayensis</u> Merr.	PENDU	<u>S.pendula</u> Bl.
PLURI	<u>S.plurilocularis</u> White & Francis	POOLE	<u>S.poolei</u> White & Francis
PUNDU	<u>S.punduana</u> Dyer	PURG	<u>S.purgans</u> Burt.
ROXB	<u>S.roxburghii</u> Wall.	RUBEN	<u>S.rubens</u> Ridl.
RUBIC	<u>S.rubicunda</u> Seem.	RUBRI	<u>S.rubrisquamata</u> Smith
SCABE	<u>S.scaberrima</u> Lautbch	SCHUM	<u>S.schumanniana</u> Diels
STERR	<u>S.sterrolepida</u> Diels	STICH	<u>S.stichophlebia</u> Diels
TAFa	<u>S.tafana</u> Smith	TRIST	<u>S.tristyla</u> DC.
VULCA	<u>S.vulcani</u> Korth.	XIPHO	<u>S.xiphophylla</u> Diels

Leaf based classification of species

The leaves of 94 species of Saurauia were scored for the character set described in chapter 5. No raw data matrix of these species will be included as all data used in analyses of leaves is included in descriptions in Appendix 3. For analyses only the final reduced character set of nine metric architectural characters described in chapter 5 was used. Metric characters were initially divided into seven intervals of equal length and the species scores in the raw data matrix were transformed to these class values. Tests of normality were conducted on each of the character distributions and the histograms were plotted. The plots have not been included here however. In all cases no significant difference from normality was noted. This procedure was important for two reasons. Firstly, the underlying distributions had to be known to allow equal frequency class intervals to be generated. Secondly, several of the statistical procedures that were later applied to the data require a multivariate normal distribution in the data set. Univariate normality

testing does not guarantee this. However, currently no test for multivariate normality is available. The fact that the distributions were not divergent from normality in one dimension is at least a partial guarantee that this requirement is likely to be met.

Using the normally distributed data (NORVARS) a phenogram to a reduced set of species whose leaves had been scored was constructed. Two criteria were used to reduce the number of species for analysis. Firstly because it was considered important to compare both leaf and floral classifications, only those species for which type descriptions had been obtained were included. Secondly, to produce as reliable a classification as the sample of species would allow, only those species for which two or more herbarium specimens had been seen were used. This produced a set of 44 species which were analysed in detail.

The first stage in analysing the 44 species on leaf data was the construction of single linkage phenograms to see if these produced any discernible taxonomic structure. The following eight phenograms were constructed on SPECIMENS.

SPECIMENS, NORVARS, RANGE, I_D & D_I

SPECIMENS, NORVARS, MEANS, I_D & D_I

SPECIMENS, RANVARS, RANGE, I_D & D_I

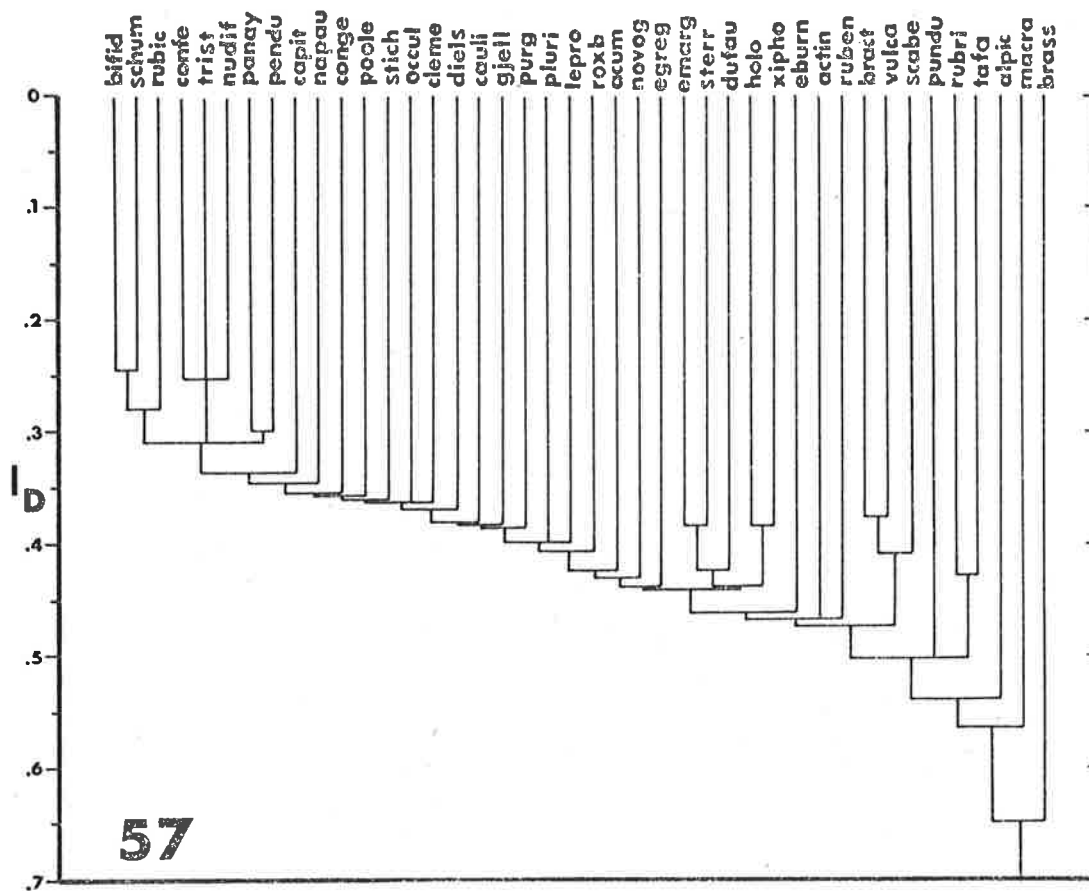
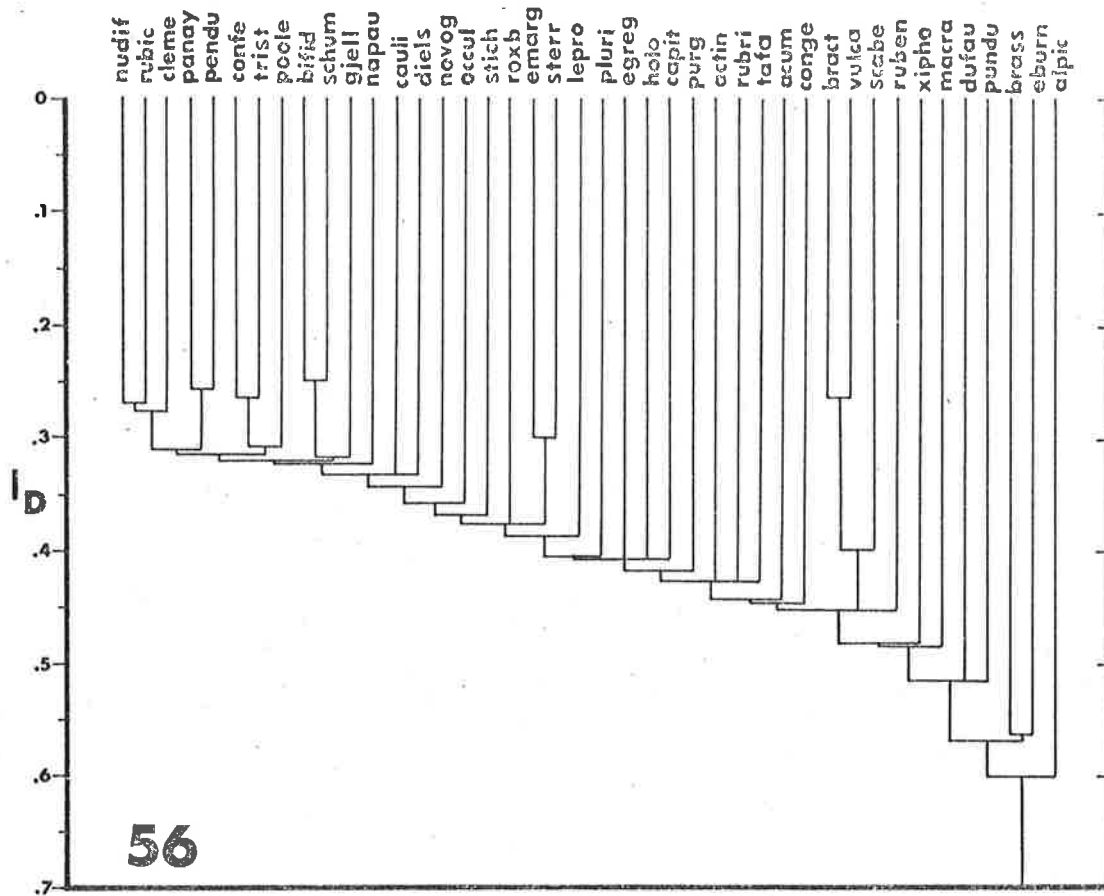
SPECIMENS, RANVARS, MEANS, I_D & D_I

According to recommendations given in the previous chapter with regard to generating classifications using measures of resemblance, the most reliable of these could be expected to be the final one in the list. This last phenogram, however, has been constructed in a special manner and will be considered after some of the intermediate plots have been illustrated.

Firstly, phenograms of SPECIMENS, NORVARS, RANGE, I_D (fig. 56) and SPECIMENS, RANVARS, RANGE, I_D (fig. 57) were prepared. It was

FIGURE 56. Single linkage phenogram for Saurauia species using SPECIMENS, NORVARs, MEANS, I_D . See text for further explanation.

FIGURE 57. Single linkage phenogram for the same data set as figure 56, but using RANVARs. See text for details.



apparent at this stage that nothing so simple as a single linkage phenogram was likely to be of value in defining a classification. Examination of the figures supports this. Both of them show a distinct lack of structure. Very few clusters are clearly discernible and the ordinations produced by them differ quite considerably. To compare relative ordinations, the number of positions in the phenogram that each species was shifted by RANVARS as opposed to NORVARS, was computed. On average, species moved 5.33 places. Although this is a shift of only 12.7% of the number of possible positions, a visual assessment of the phenograms reveals little meaningful information. The only easily discerned consistency between the two is that a number of species are paired in both. The species BIFID and SCHUM, EMARG and STERR, and BRACT, VULCA and SCABE form fairly distinct clusters. From a subjective assessment of the leaves of these species, it was decided that such clustering was at least reasonable. However, in many cases the juxtaposition of species made no sense at all.

The next analysis involved generating phenograms using MEANS. These plots will not be included here as they form only a further intermediate step in the analysis. However it was very apparent that the most satisfactory discrimination was achieved using RANVARS, MEANS, D_I . This plot produced a number of distinct clusters of species, and the variance of distance values was considerably increased. It had been found that the levels at which species separated in phenograms using NORVARS lay mainly between the phenogram levels 1.0 and 1.6. RANVARS, RANGES however increased this so that most species linked between 1.0 and 2.0. This produced much clearer clustering and also meant that fewer species linked at identical levels, thereby making the phenograms easier to interpret.

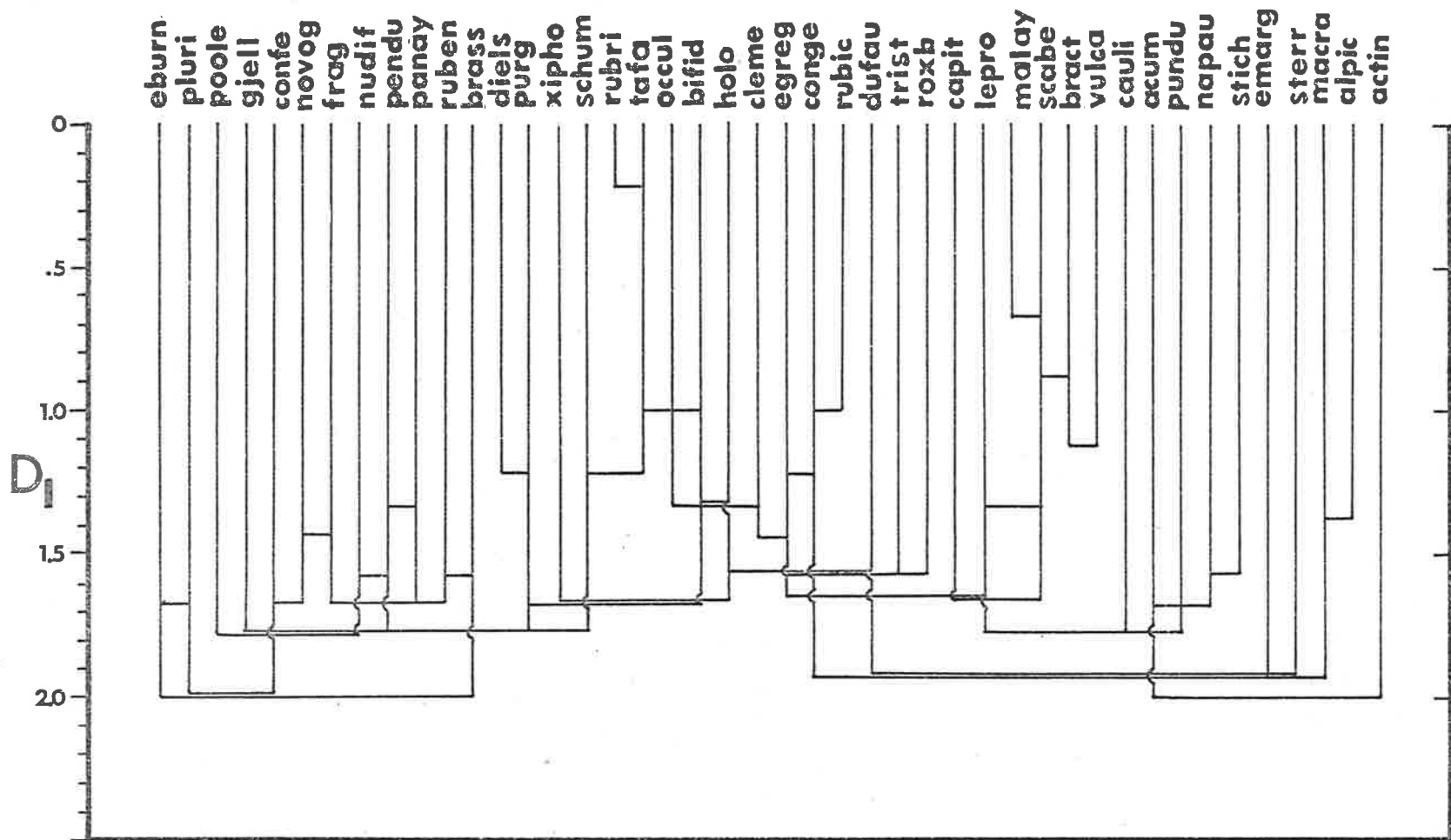
Single linkage phenograms have minimal information content as graphical representations of the whole resemblance matrix between OTUs. A discussion of such plots, their information content and the distortions they are subject to as two dimensional representations of a multi-dimensional situation, are discussed in the following chapter.

An improvement in phenogram information was achieved by using a method of plotting nearest neighbour relationships. The method of producing such figures is explained in the following chapter and illustrated for SPECIMENS, RANVARs, MEANS, D_I in figure 58. Careful examination of this figure reveals some distinct clustering of species. Especially notable is the cluster at the centre of the phenogram, containing the species SCHUM, RUBRI, TAFA, OCCUL, BIFID, HOLO and CLEME. These species are all New Guinea species in the generic section Setosae. A second very distinct cluster contains the species LEPRO, MALAY, SCABE, BRACt and VULCA, mostly members of the Ramiflorae of Indo-China and Malaysia.

Nearest neighbour phenograms of the type illustrated in figure 58 are very difficult to interpret without careful study of the linking lines. Therefore a further graphical method derived from these phenograms was applied.

This involves using the phenogram levels or dissimilarity values to arrange the OTUs in a two dimensional nearest neighbour tree. Such a plot is illustrated for SPECIMENS, RANVARs, MEANS, D_I in figure 65. The distinct clustering of species on this figure is the same as that in figure 58. Distances between species are equal to the value of D_I for nearest neighbour pairs. Angles between the connectives are maximised unless constrained by a closed loop. On figure 63, three such loops can be seen. These arise where a species has two nearest neighbours. That is, it is equally linked to two

FIGURE 58. Strict nearest neighbour phenogram
for Saurauia species using SPECIMENS,
RANVARs, MEANS, D_I . See text for
details.



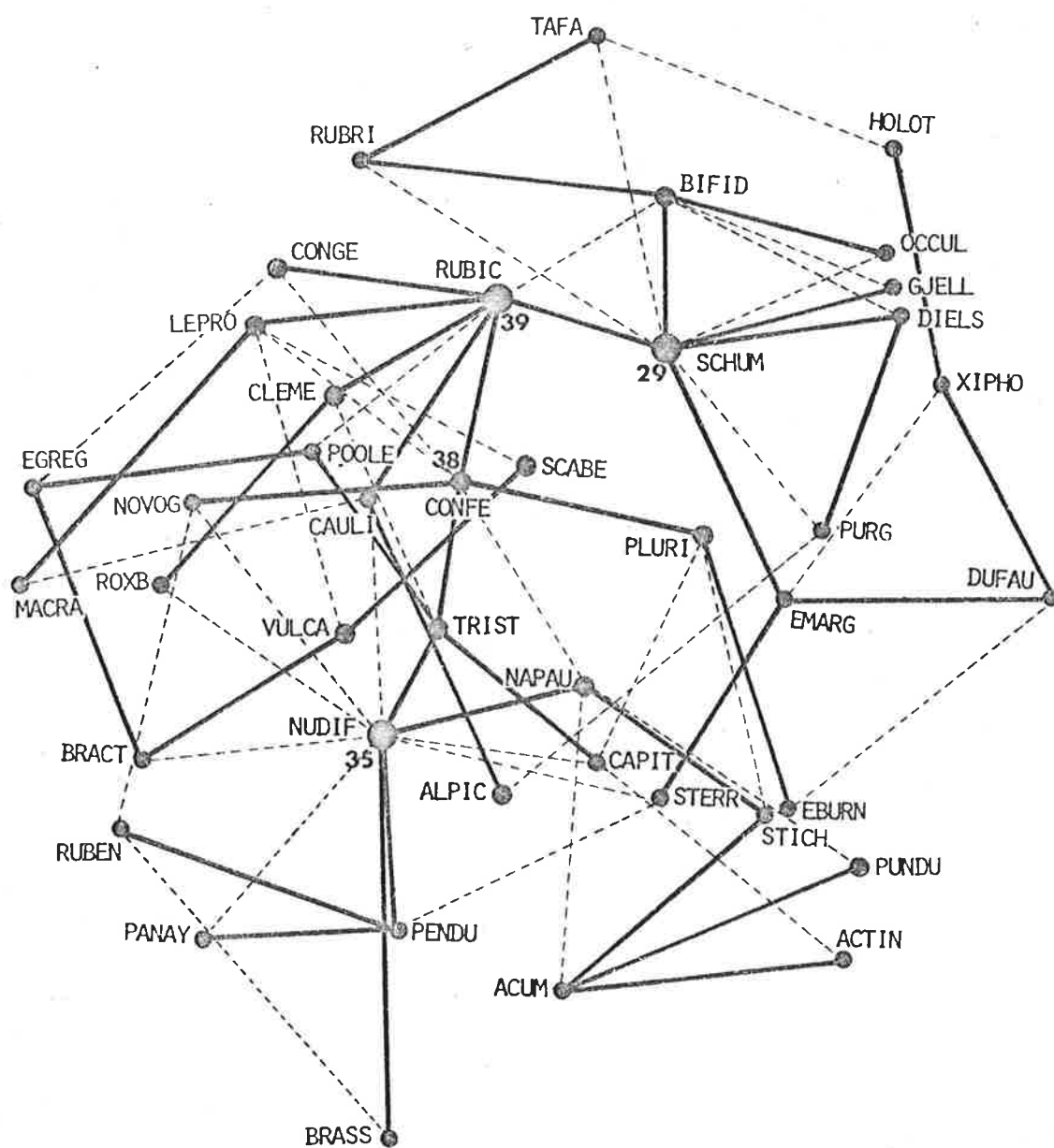
species e.g. GJELL to SCHUM and PURG, or PURG to GJELL and BIFID. In the figure, the cluster of New Guinea Setosae forms a closed loop near the centre. The Ramiflorae clearly clustered in figure 58 occupy the lower right region. To clarify the ordination of species in this figure, species from outside the New Guinea-Australian-Solomon Islands region are enclosed by a broken line.

The ordination of species on a nearest neighbour tree has been demonstrated in figure 65. However, this plot does not contain any information regarding higher order relationships between species. Therefore a nearest neighbour plot incorporating higher order neighbour relationships is illustrated in figure 59. In this figure second nearest neighbour links are defined by dashed lines. A number of interesting features which can be related to both the underlying taxonomic structure and to various other influences can be discerned. Firstly it is obvious that there are a number of species which form local centres around which other species are clustered. The particular species forming centres are SCHUM, RUBIC, NUDIF and CONFE.

The presence of local centres on the plot is very easily explained. The influence of range measurements i.e. multiple values per character for OTUs, has already been considered in detail. If one remembers (see example illustrated in figure 50) that similarity is proportional to range, then determining the number of states for OTUs on the plot (figure 59) reveals the reason for the centres. The four species involved are those in the set of OTUs which have the largest number of states. The actual values are included on the plot adjacent to the species involved.

Beyond the presence of these centres, there is further structure discernible. It will be noticed that the plot as a whole is not particularly equilateral (for comparison see figure 64). In fact

FIGURE 59. First and second nearest neighbour tree for SPECIMENS, RANVARs, RANGE, I_D for 42 species of Saurauia. Solid lines represent nearest neighbour links and dashed lines, second nearest neighbour links. Three large dots indicate local centres of similarity (NUDIF, RUBIC, SCHUM). The fourth centre is the species CONFE. Numbers adjacent to centers indicate the total number of states, over a set of 9 characters of 7 states each, that the species possess.



most of the equilateral structure appears to lie about the centres. Interpretation of the overall heterogeneity of the similarity matrix is probably not justified in this example however because of the range influence.

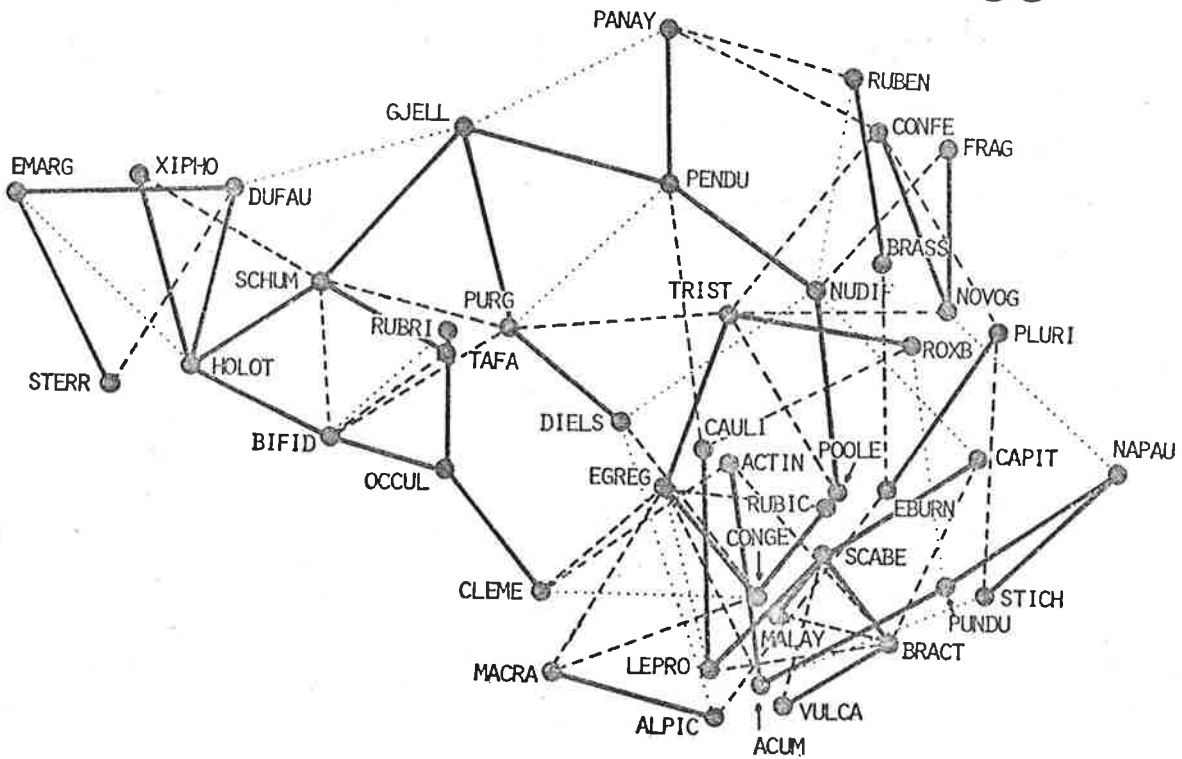
The final structure which will be considered however appears to have an underlying taxonomic determinant. It is apparent that the upper right hand side of the cluster is fairly well separated from the remainder. In fact all the links joining it to the main cluster are, except for the SCHUM-RUBIC link, second nearest neighbour links. If this subsidiary cluster is studied in detail, one finds that it contains only New Guinea species of mainly the section Setosae. These are the same species which form the central cluster of the nearest neighbour phenogram (fig. 58) and the nearest neighbour tree of figure 63. This relationship is completely masked in the phenogram of figure 57. This single example of a consistent and unambiguous ordination of species is enough to justify further investigation into nearest neighbour graphs.

To remove all possible influences and produce a plot whose structure was determined solely by taxonomic relationships, I decided to produce a final ordination using SPECIMENS, RANVARs, MEANS, D_I . This is illustrated in figure 60. In this figure it has been possible to include a high proportion of third nearest neighbour links as well. A very distinctive structure results. To the left a "sheet" of species can be seen. These have a very pronounced equilateral structure indicating the homogeneity of the set of species involved. Note that this region of the plot is dominated by members of the New Guinea Setosae. These are in fact the same group of species that were differentiated in figure 61. The species RUBRI and TAFA are very close together and doubtfully distinct. The spatial relationships of these

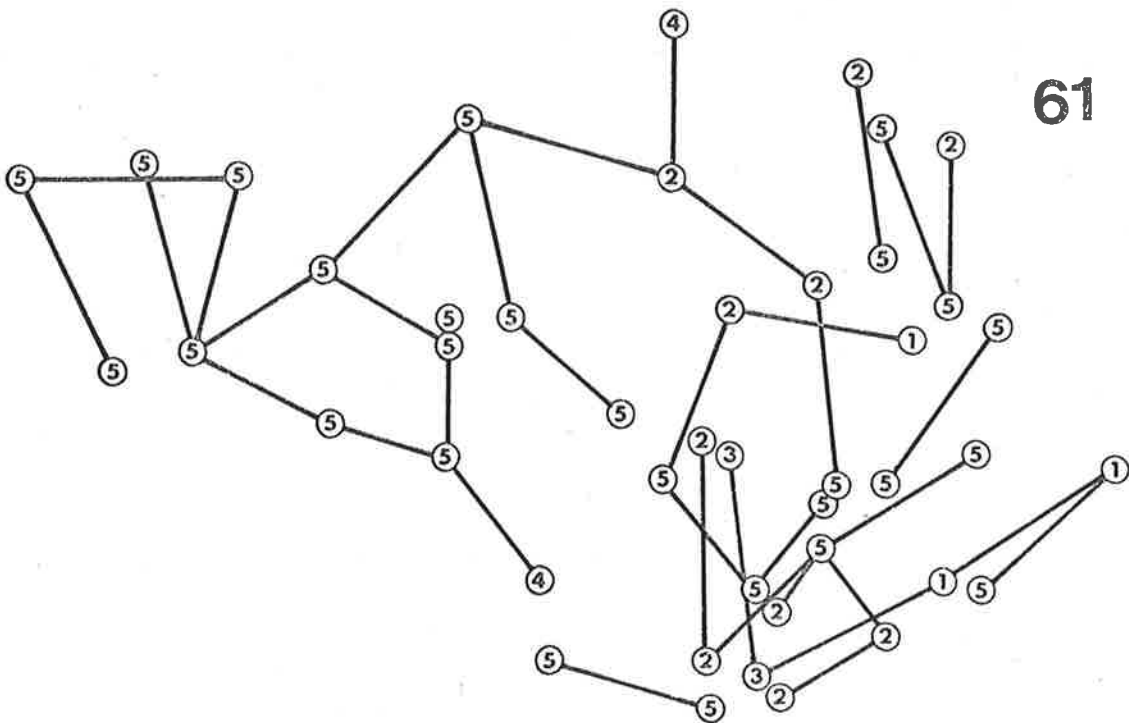
FIGURE 60. First and higher order nearest neighbour plot for SPECIMENS, RANVARs, MEANS, D_I . Solid lines represent 1st neighbour links, dashed and dotted lines represent respectively 2nd and 3rd nearest neighbour links.

FIGURE 61. First nearest neighbour links from figure 60, with species marked with geographical regions.

60



61



species and the species BIFID, HOLOT, SCHUM, OCCUL and PURG are almost identical in this plot and in figures 58 and 65. The consistency of ordination and the apparent homogeneity suggest that these species form a distinct complex. Their relationships will be investigated further in a later section. Another consistently ordered group contains the species VULCA, BRACT, MALAY, SCABE, LEPRO and CAULI. The chain containing these species is visible in the lower right corner of figures 60 and 61. The same set is clearly linked in figure 59 and also forms a distinct cluster in figure 58. This group contains mostly species of the section Ramiflorae from mainly Indo-China and Malaysia. The geographical influence on ordination in the plots has also been noticed on figure 59 and can be detected in figure 58. To illustrate the geographical clustering of species, figure 61 has been plotted. The species are given numbers which correspond to the five geographical provinces defined in chapter 12. Only first nearest neighbour links are shown to illustrate the backbone of the tree. Although there are a high number of species from the geographical region 5 (New Guinea and Solomon Islands), the ordination of mainly region 5 species to the left of the plot and mainly region 1, 2, 3 and 4 species to the right, is very clear. This geographical ordination will be considered in detail at the end of this chapter and is explained in chapter 12.

The general impression of the taxonomy of Saurauia based on leaves is that the genus is morphologically fairly continuous and does not have distinct clusters of species, except in one or two cases. Any clustering is fairly diffuse, and contains very homogeneous species complexes. A consistent general ordination however has been demonstrated by the analysis of nearest neighbour relationships. Detailed analysis of speciation will be reserved for after a discussion of numerical analysis of type description data.

Classification of species on type description data

The second aspect of generating a leaf based classification of Saurauia was the comparison of this classification with that produced from more classical taxonomic features by the same analyses.

To obtain this data it was decided that the only feasible method was to extract information from the type descriptions. This was done because a larger number of species were available as type descriptions than as herbarium specimens. This was also the form in which most taxonomists would encounter the species. Finally, it was interesting to attempt a classification on type descriptions because many type specimens of Saurauia species, notably those of Diels, have been destroyed. Diels described many of the species scored for leaf characters in the present study.

Type descriptions acceptable for data retrieval were obtained for 193 species. Several other descriptions were excluded because they were so abbreviated that they were considered to be of almost no use. A survey of the 52 most complete descriptions was conducted, scoring each species for every distinguishable character. The characters scored for each species were then compiled and where necessary equalised. That is where characters were given different terms in descriptions but were clearly the same, they were given a standardised term in the data matrix. The frequency with which each character was scored over the data set was also determined. This was done to try and minimise the quantity of missing or inapplicable data such as secondary characters. Initially 76 characters were defined from the data set. I think that it is worth including the list of characters and their scoring frequency so that the reduced character set can be seen in context with those that have been excluded. The ordering of characters within the character list follow as closely as possible

the general order in which they have been used in descriptions.

The characters and frequencies are given in Table 8.

Character	Percent of type descriptions in which character has been scored. Actual no. of species to which characters are appli- cable is bracketed
Habit	90.4 (52)
Branch indument	88.5 (52)
Arrangement of leaves on branches	11.5 (52)
Petiole length	94.0 (52)
Petiole thickness	30.8 (52)
Petiole indument	90.4 (52)
Leaf shape	100 (52)
Leaf length	100 (52)
Leaf width	100 (52)
Base type	96.0 (52)
Petiole structure (e.g. winged)	21.2 (52)
Leaf apex type	100 (52)
Margin type	98.1 (52)
Tooth/seta density	28.8 (52)
Midrib indument	92.3 (52)
Number of pairs of 2° veins	100 (52)
2° vein course	53.8 (52)
Occurrence of inter-2° veins	3.8 (52)
2° vein indument-upper surface	88.5 (52)
2° vein indument-lower surface	100 (52)
Intercostal indument-upper surface	86.5 (52)
Intercostal indument-lower surface	96.2 (52)
Inflorescence position	82.7 (52)
Inflorescence type	75.0 (52)
Number of flowers per inflorescence	86.5 (52)
Peduncle length	94.3 (35)
Peduncle thickness	64.5 (31)
Number of bracts	67.5 (40)
Bract shape	80.0 (30)
Bract length	80.0 (30)
Bract width	50.0 (30)

Character	Percent scored
Bract indument-outer surface	90.0 (30)
Bract indument-inner surface	36.7 (30)
Number of bracteoles per flower	54.8 (42)
Bracteole length	61.9 (42)
Bracteole width	21.4 (42)
Bracteole shape	57.0 (42)
Bracteole indument-outer surface	57.0 (42)
Bracteole indument-inner surface	28.6 (42)
Pedicel length	86.4 (44)
Pedicel width	45.5 (44)
Bud diameter	21.2 (52)
Number of sepals (e.g. 2+3)	100 (52)
Sepal indument-outer surface	96.2 (52)
Sepal indument-inner surface	67.3 (52)
Sepal shape	100 (52)
Sepal length	96.2 (52)
Sepal width	82.7 (52)
Sepal apex shape	84.6 (52)
Number of petals	94.2 (52)
Petal indument-outer surface	86.5 (52)
Petal indument-inner surface	84.6 (52)
Petal length	80.8 (52)
Petal width	73.1 (52)
Petal shape	78.8 (52)
Petal apex shape	73.1 (52)
Number of stamens	95.2 (42)
Number of series of stamens	69.0 (42)
Filaments free or fused	57.1 (42)
Filament total length	69.0 (42)
Filament, length of fused part	28.6 (42)
Anther length	59.5 (42)
Anther apex type/shape	71.4 (42)
Anther dehiscence	16.7 (42)
Ovary shape	78.0 (50)
Ovary height	28.0 (50)
Ovary diameter	44.0 (50)
Ovary indument	86.0 (50)

Character	Percent scored
Ovary lobing	16.0 (50)
Number of styles	94.1 (51)
Styles free or fused	78.4 (51)
Style length	74.5 (51)
Style, length of fused part	37.3 (51)
Style indument	23.5 (51)
Stigma type/shape	51.0 (51)
Number of ovules	3.8 (52)

TABLE 8. Frequencies of species for which applicable taxonomic characters have been scored. Determined from a set of 52 *Saurauia* species.

Scoring the character set was not a simple task of extracting the characters directly from the descriptions. In many cases characters were inapplicable, or uncertainly so and therefore the actual number of scores of a character was expressed as a percentage of the total number of species to which it was unquestionably applicable. This latter number is given in brackets in table 8. Some characters were therefore computed over as few as 30 descriptions e.g. the characters of bracts. A number of characters such as stamen features were only counted as applicable if the type specimen possessed flowers at the correct stage to score them. In this manner the frequency of scores of secondary characters were equalised somewhat. This latter case is an example of "being fair" to the floral characters because it soon became apparent that these were on the whole scored less frequently than leaf characters. The distribution of character frequency scores was fortunately skew towards high frequencies e.g. 44 characters were scored in more than 71% of descriptions. Several interesting statistics can be recovered for the relationships between leaf and floral characters in the descriptions.

Firstly it is possible to compute the average within species range of characters as compared to the overall range. This was done by firstly standardising the metric characters to six states. This number was chosen because it was close to the average number of states of qualitative characters. Species were then scored for the number of character states they exhibited for each character. By finding the sum of the number of states exhibited by species for all species on each character the total frequency distribution of states of each character was obtained. The number of values in this distribution was then divided by the product of (number of species scored for the character) and (number of states for the character). This value was then converted to a percentage. Average values of these were then computed for leaf and floral characters. By this means the average percent of the total range of each class of characters that a species could be expected to show, was calculated. The relative values of these percentages for leaf and floral characters are:

Leaf characters average percent of total range = 24.75%

Floral characters average percent of total range = 27.03%

Finally the average number of leaf and floral characters scored for species was calculated.

Leaf characters scored on average for a species 95.41%

Floral characters scored on average for a species 81.48%

These figures produce a startling interpretation. Of 14 leaf characters and 28 floral characters scored in more than 70% of the 52 type descriptions, leaf characters were slightly less variable on average than floral characters and they were scored much more frequently in the descriptions. The first finding is surprising in the light of what critics of leaf studies would have us believe. The second finding is also very surprising considering that the absence

of a structure was considered as a case where a character was inapplicable in computing scoring frequencies. Many descriptions were scored in which all or part of the floral structures were missing. If this had not been accounted for in the computation of frequency then it could be explained simply in terms of the flowers being incomplete or lacking due to the season in which they were collected. This is not valid in the present case. I have no explanation of why taxonomists who traditionally ignore leaves should have scored them more often in the set of type descriptions. Taking into account the fact that flowers are more likely to be missing from any one collection than leaves, leaf characters must surely be more reliable than floral characters from the aspect of being generally present. Whether leaf characters are better discriminators of species will be considered in the last section of this chapter when leaf and type description classifications are compared.

For reasons of missing data, secondary characters, probable character correlation and invariance in the set of species, a subset of the 76 characters was chosen for analysis of the species set. The three criteria that were applied to choose the final character set were as follows. (1) The characters should have been scored frequently. This was arbitrarily set so that only those characters scored in at least 80% of descriptions were considered. (2) The characters should, where possible, be multistate metric or non-metric. (3) The character set should consist of mostly floral characters so that the classification produced was as far removed from the leaf based classification as possible. This was done so as to check the two essentially independent classifications. The final character set that was chosen after exclusion of a few clearly correlated or highly variable characters was as follows:

Character	No. of states
Leaf width	Metric
Leaf base type	5
Number of secondary veins	Metric
2° vein indument - lower surface	5
Sepal indument - outer surface	5
Sepal shape	5
Sepal length	Metric
Number of styles	6
Percentage of style fused	Metric
Number of flowers per inflorescence	Metric
Petiole indument	5
Bract length	Metric
Petal length	Metric
Ovary indument	2

Of a few characters excluded at the last step, leaf shape was omitted because the terminology was ambiguous. Habit was also excluded because there was no guarantee that this had any meaning as a specific character, and petiole length was excluded because of its high within species variance demonstrated from the leaf study. Number of stamens was excluded because it was generally expressed as an inexplicit range or with an adjective such as "many."

Because of the criterion of equal frequency class intervals, the distributions of the characters across the set of species were studied and tests of normality conducted where applicable. The frequency distributions for each of the characters across the full 192 species set are given below.

Character	Character states and frequencies
Leaf base	
	Attenuate (49) Acute (62) Obtuse (35) Rotund (22) Cordate (10)
Vein indum.	
	Glabrous (24) Tomentose (45) Setose (72) Scaly (45) Thorny (5)
Sepal indum.	
	Glabrous (40) Toment. (34) Setose (75) Scaly (31) Thorny (5)

Character	Character states and frequencies
Sepal shape	Linear (1) Oblong (28) Elliptic (21) Ovate (73) Orbicular (56)
No. styles	2(1), 3(68), 4(20), 5(92), 6(3), 7(2)
Petiole ind.	Glab. (20) Toment. (30) Setose (77) Scaly (60) Thorny (9)
Ovary indum.	Glabrous (103) Tomentose (53)

The character "number of styles (or locules)" has an interesting distribution. Hoogland (1977) makes note of the variability of style number within some species. The ranges of style numbers from type descriptions support this, but also demonstrate that probably most species are mostly 3- or 5-styled. Of all species, only 6 range from 3 to 4, 2 range from 3 to 5, 5 range from 4 to 5 and 2 range from 5 to 7 styles. A discussion of style number in dividing the species is given in the biogeography chapter.

For the analyses metric characters were subdivided into eight class intervals. The class limits and class frequencies are given below.

Character	Class limits and frequencies of species in each class (bracketed)
Leaf width (cm)	1(59) 3.9(69)4.5(68)5.1(65)5.6(69)6.1(66)7.2(71)9(62)25
No. 2° veins	5(33)9(48)11.5(56)12.5(42)13.5(36)14.5(39)16.5(31)19(35)60
Sepal 1th. (mm)	1.5(39)4.5(34)5.7(42)6.7(40)7.7(42)8.7(44)10(31)12.7(16)26
No. flowers	1(61)2(56)3(51)5(50)many
Bract 1th. (mm)	0(17)2.5(18)5(23)7(23)8.5(18)9.5(19)11.5(22)15.5(17)40

Petal 1th. (mm)

1(19)5.7(26)7(20)8(19)9.5(21)10(22)10.5(21)13.5(22)

Class intervals for metric characters were set to equalise frequencies as much as possible.

Figure 62 is a frequency plot for equal class intervals for the character leaf width. Very obvious on this plot, is the presence of local peaks on the underlying distribution. The peaks always lie at integer values of the scale. The scale is plotted for both centimetres and inches. The obvious explanation for these peaks is that they represent a measuring bias on the part of the taxonomist describing the plant. This is either due to a preference for reading integer values off the measuring scale, or due to rounding the values to whole numbers when expressing ranges. This latter is probably the case, though the former cannot be discounted especially when the measurements are at a relatively small scale e.g. bract or sepal lengths. This effect was noticed in distributions of all metric characters. In some cases such as sepal length the integer bias was so pronounced that the frequencies in integer peaks were approximately double those in adjacent non-integer classes.

Type data phenograms and nearest neighbour trees.

In this section the TYPES, RANVARs, etc. terminology will once again be applied.

Using the computer program TAXON, dissimilarities and distances for all OTUs were computed. Initially, phenograms and nearest neighbour trees were constructed for only the 44 species that were treated by these procedures for leaf data (see leaf classification section, this chapter).

Phenograms were constructed for the following.

TYPES, NORVARs, RANGE, I_D & D_I
 TYPES, RANVARs, RANGE, I_D & D_I
 TYPES, NORVARs, MEANS, I_D & D_I
 TYPES, RANVARs, MEANS, I_D & D_I

FIGURE 62. Frequency plot of leaf widths derived from 193 type descriptions of Saurauia species. Scale in inches and centimetres to demonstrate the bias towards integer values. See text for discussion. Crosses mark polynomial curve fit to underlying distribution.

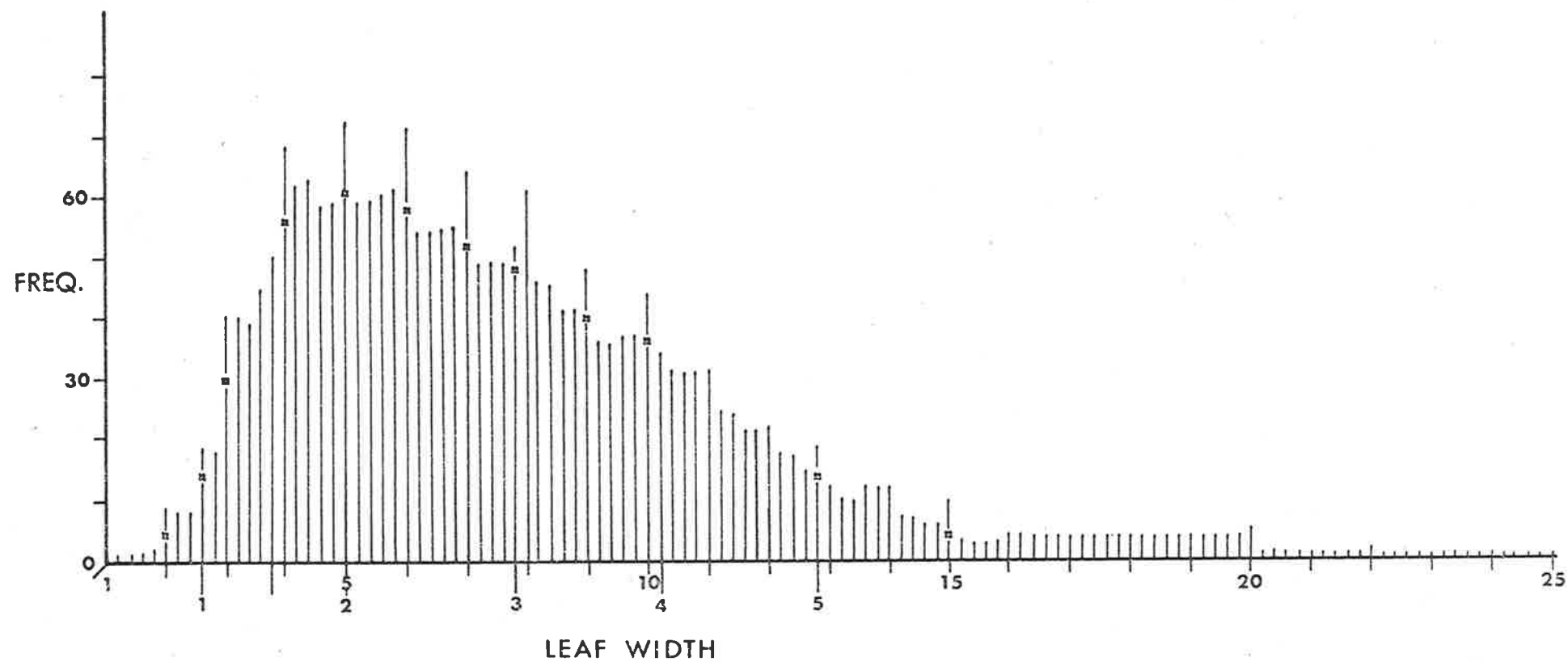


Figure 63 illustrates the similarity based ordination of TYPES. This plot is reminiscent of the SPECIMENS phenograms in presenting a distinct lack of structure (see Figs. 55 and 56). A few indistinct clusters such as that containing ROXB, VULCAN and NAPAU are in evidence. This plot and the equivalent plots of species scored for leaf data (figs. 56 and 57) have no apparent structure in common. Even the strong Setosae clustering is lacking. To improve the information content of the phenogram, figure 64 was plotted. This is a nearest neighbour phenogram constructed for TYPES, RANVAR, MEANS, D_I and is therefore equivalent to figure 58. Comparing these two plots, the most obvious difference is that the variance of distance values for TYPES is much lower than it is for SPECIMENS. This can be seen in the fact that the phenogram levels at which species are linked are much more clustered about a narrow band in the region of D_I equal to 1.5, than they are in the SPECIMENS phenogram. Examining details of the phenogram it can be seen that there is some similarity to the equivalent plot for SPECIMENS (fig. 58). For example the species EMARG and STERR are linked in both. The species RUBRI and TAFA are also linked in both, though this is far less significant on the TYPES plot. POOLE and PENDU are fairly close in both, as are SCHUM, HOLO and TAFA. To overcome the difficulty of reading the nearest neighbour phenogram, a nearest neighbour tree was constructed. This is illustrated in figure 65 with the equivalent tree for SPECIMENS.

Inspecting the nearest neighbour trees for TYPES and SPECIMENS many more similarities can be found. For example STERR and EMARG in both are on an outlying branch. In both they are fairly closely linked to the branches including SCHUM, TAFA, RUBRI, DIELS and OCCUL, species which belong to the New Guinea Setosae. It is also possible to see that in both trees the species VULCA, NAPAUL, MALAY, PUNDU, BRAC and CAULI are closely associated. These species are mainly

FIGURE 63. Single linkage phenogram for TYPES,
RANVARS, RANGE, I_D for 42 Saurauia
species.

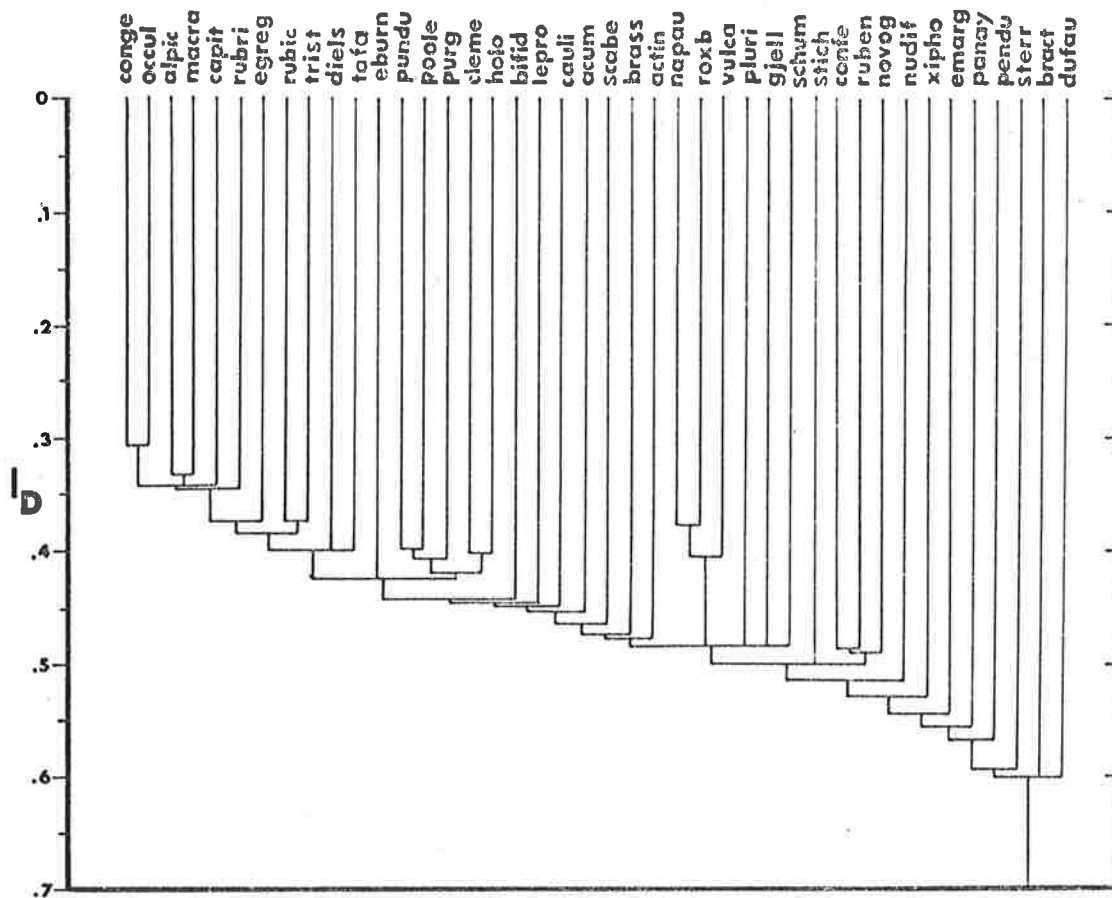


FIGURE 64. Nearest neighbour phenogram for
TYPES, RANVARs, MEANS, D_I for the
same set of species as in figure 63.

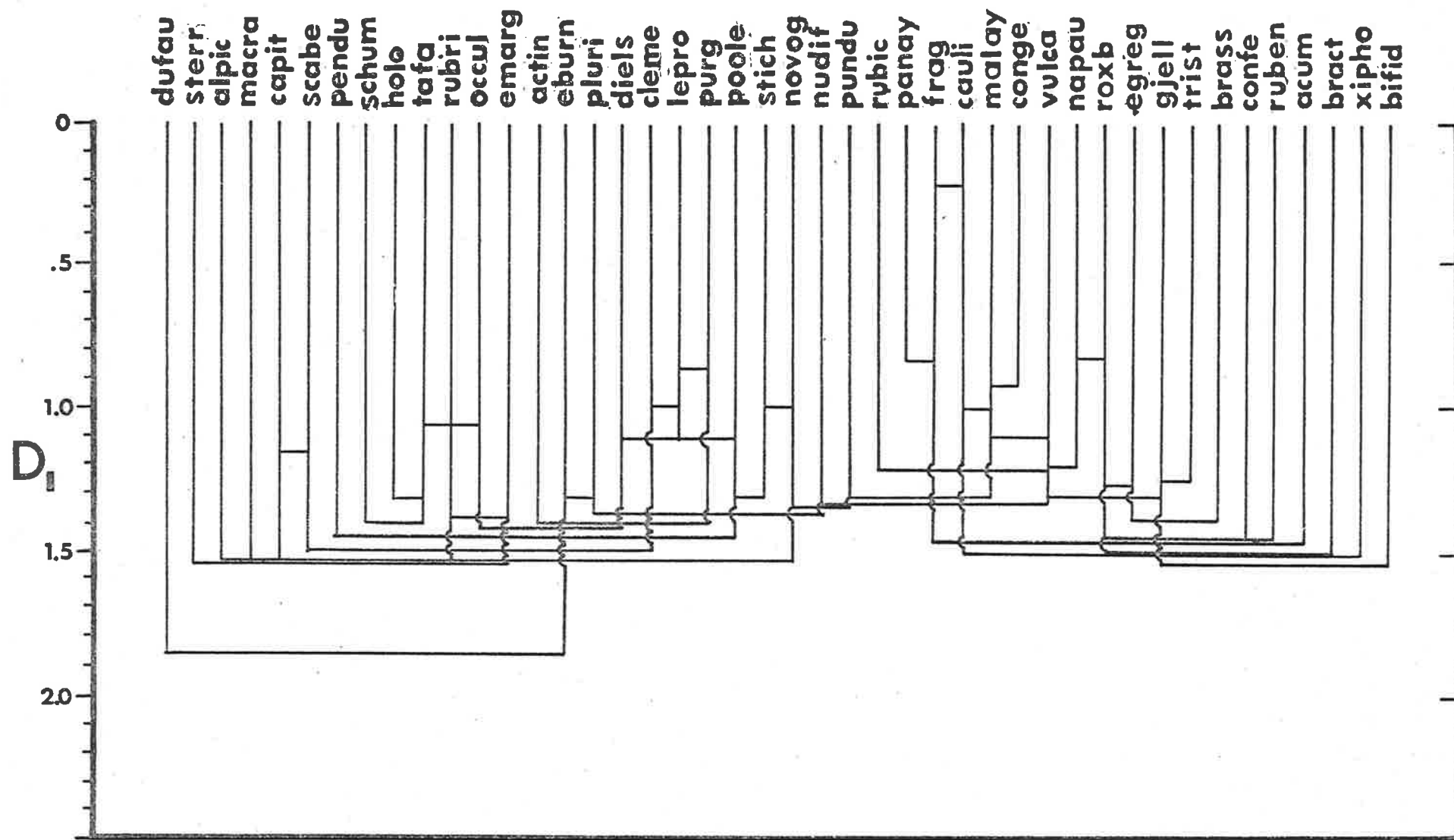
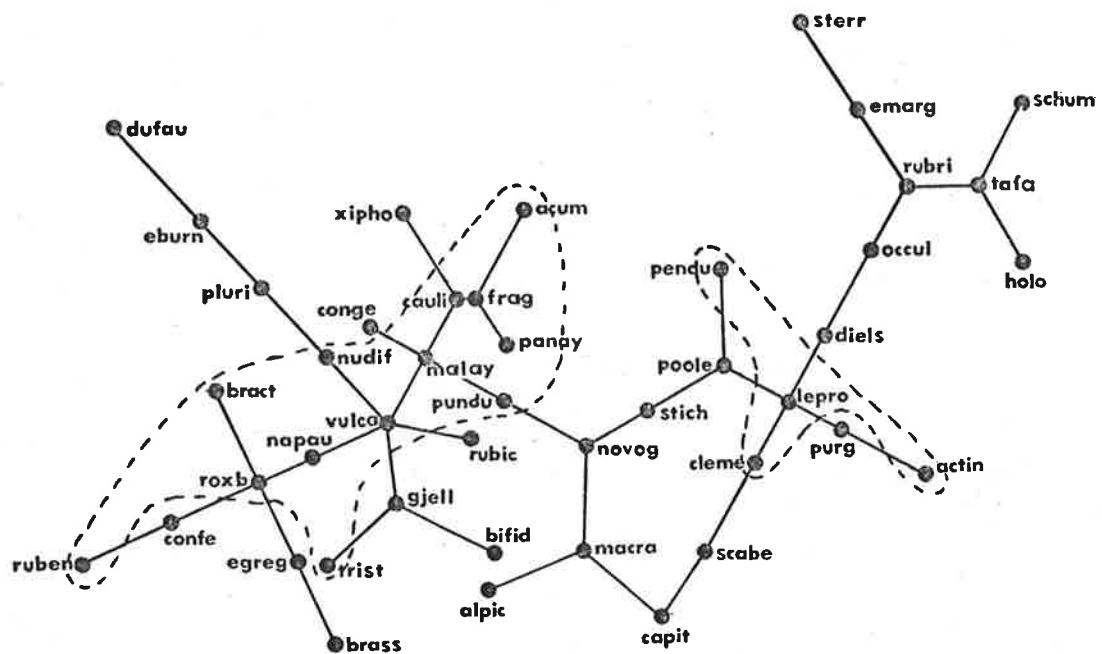
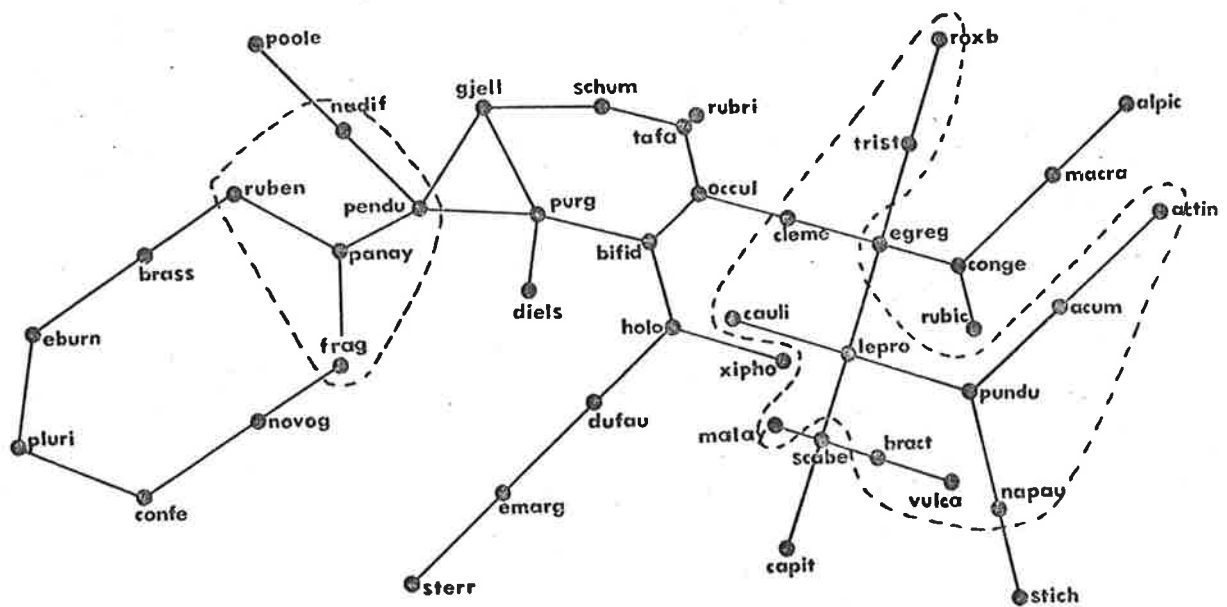


FIGURE 65. Nearest neighbour tree for Saurauia species scored for leaf data (top) and type description data (bottom). Outlined areas include species from regions outside New Guinea.



members of the Ramiflorae from Indo-China and Malaysia. In both plots the pairs EBURN-PLURI, ALPIC-MACRA, PANAY-FRAG and SCABE-CAPIT are maintained. POOLE is close to PENDU in both plots. In all cases but the last, the ordination is supported by my subjective appreciation of the relationships between the species. The last relationship is puzzling but appears to be mainly influenced by the fact that they both have relatively large flowers and similarly shaped leaves. They are, however, quite distinct on features of the indument and inflorescence. This appears to be the only case of extreme misclassification. POOLE is a member of the Setosae and hardly distinguishable from species like SCHUM and andreana (not included in the plot), while PENDU is a member of the Ramiflorae.

To see if the ordination of species was still discernible on other dimensions of the distance matrix, higher order neighbour graphs were drawn for TYPES, RANVARs, RANGE, I_D and TYPES, RANVARs, MEANS, D_I . The first of these is illustrated in figure 66.

This figure is extremely equilateral, more so than its SPECIMENS counterpart (fig. 59). Note that although RANGE has been used, there is not the degree of local centre clustering that was in evidence in figure 59. This is because the variance of ranges is much smaller in the case of TYPES. That is, the sums of ranges over all characters are not so widely dispersed over all possible values as they are for SPECIMENS. Because the plot is so equilateral it could be expected that it is very distorted from the multi-dimensional situation (see reasons for this in the following chapter). This can be checked by plotting the higher order neighbour net. If the ordination of species is drastically affected this will be evidence for distortion. Before continuing with the higher order neighbour plot, it is worthwhile considering the few superficial similarities between

FIGURE 66. Plot of 1st and 2nd nearest neighbours
for TYPES, RANVARs, RANGE, I_D .

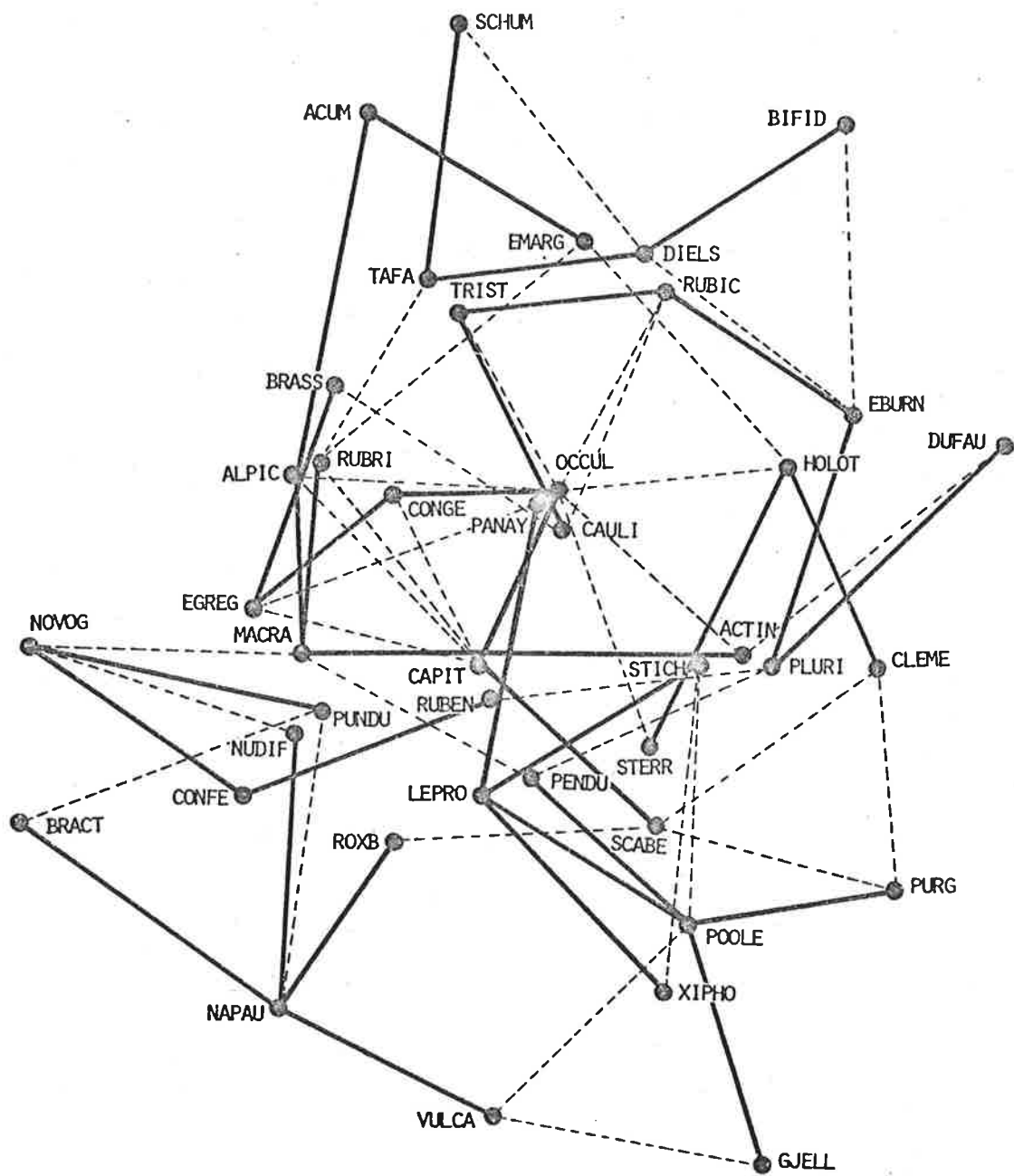


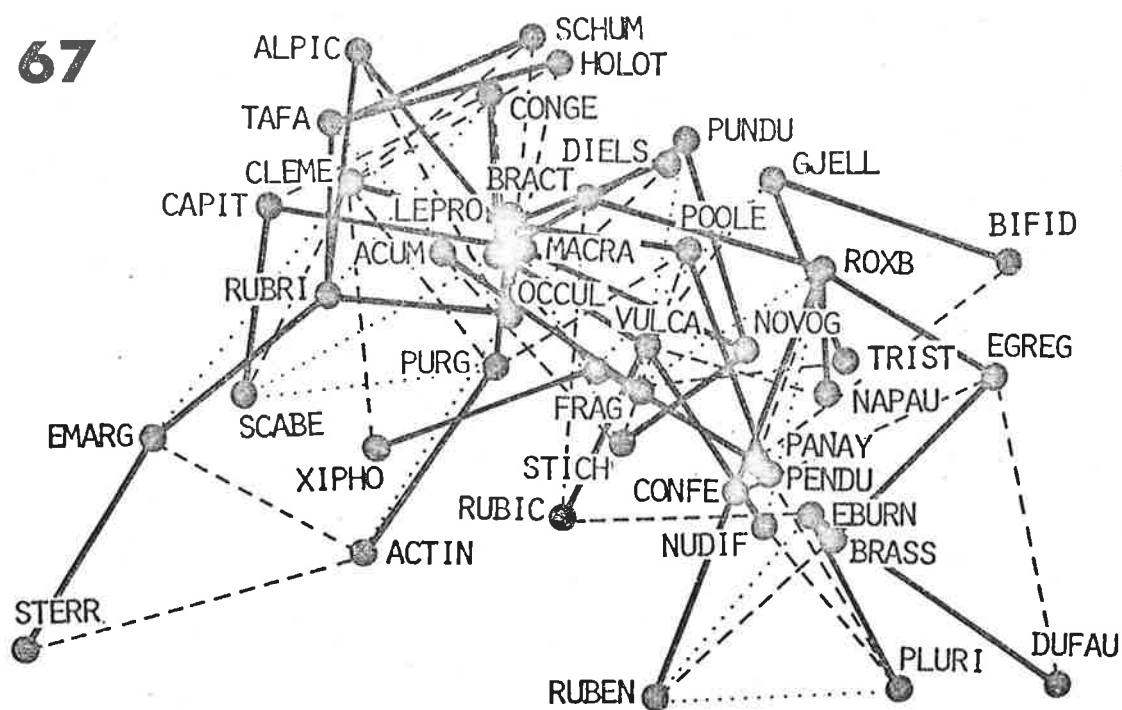
figure 64 and all of the previous nearest neighbour plots. The species SCHUM, BIFID, DIELS and TAFA are clustered together at the top of the figure. This is consistent with the relationships between these species that have been recurrently demonstrated throughout this chapter. At the bottom of the plot the species BRAC, NAPAU, VULCA and ROXBU are clustered. These species have also been more or less consistently ordered by the various graphical methods. The pairs EBURN-PLURI, POOLE-PENDU and ALPIC-MACRA have once again been maintained.

Figure 67 is the first and higher order neighbour tree for TYPES, RANVAR, MEAND, D_I . By convention, nearest neighbour links are represented by solid lines, second nearest links are dashed and the dotted lines represent the few third nearest neighbour connectives. The change in the spatial organisation of species from figure 66 to figure 67 is drastic. The distortion of the plot is very much more in evidence than that occurring in the transition from figure 59 to figure 60. Since all the information of figure 65 is retained, the ordination is essentially the same but confused by the second neighbour links. If one sees the higher order plots as folded versions of nearest neighbour trees then figure 60 is figure 65 for SPECIMENS with the two lateral clusters folded upwards about PENDU and CLEME. Figure 65 is much more difficult to relate to figure 67. The infolding of branches has taken place about numerous pivot points. This implies a high degree of cross linking between alternate branches of the figure. That is, there is greater likelihood that a member of a cluster linked by its first nearest neighbour relationships will link to a different cluster as a second nearest neighbour than is the case with the SPECIMENS plots. A higher degree of misclassification of species is expected therefore in the type description based classification than in the leaf based one. Geographical ordination of

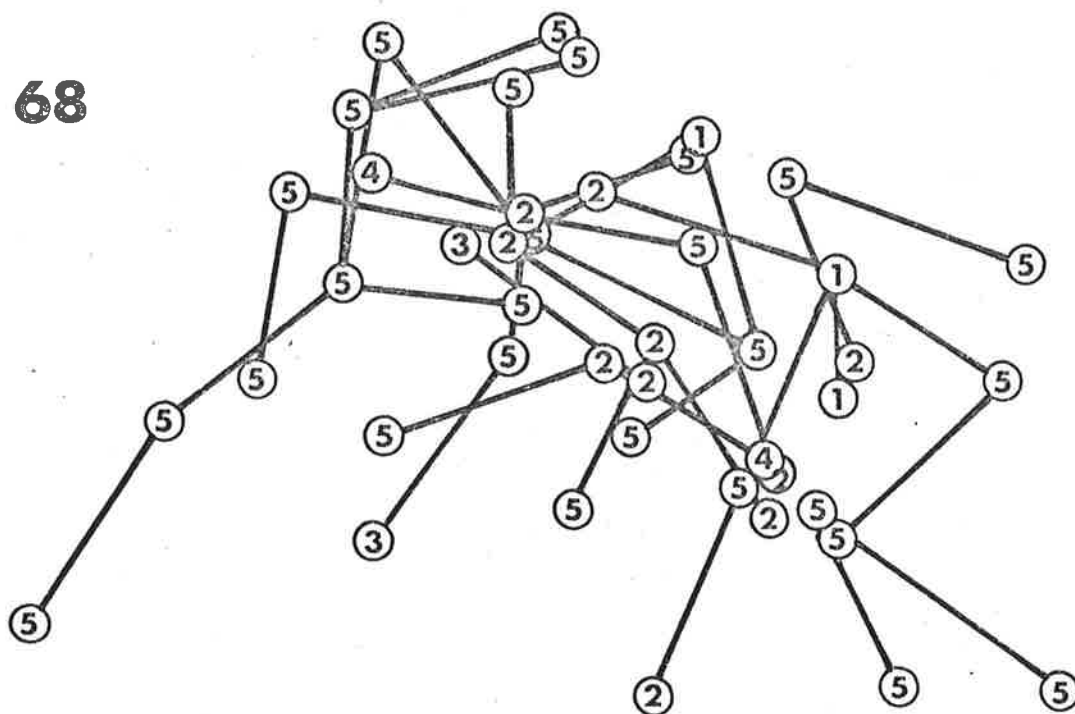
FIGURE 67. First and higher neighbour plot
for TYPES, RANVARs, MEANS, D_I .

FIGURE 68. First neighbour links only plotted
for species in figure 65. Species
are marked by numbers representing
geographical regions.

67



68



TYPES is illustrated in figure 68. This is equivalent to the SPECIMENS plot of figure 61. It is clear that there is a much poorer geographical ordination on TYPES. The question of geographical trends in morphology will be discussed at the end of this chapter. Note however figure 65 where the species have been divided into two broad geographical groups. Those species contained within the outlined areas are from Indo-China, Malaysia and Indonesia, while those outside the borders are from New Guinea and eastwards. The leaf classification tree of figure 65 has a much clearer separation of species into these two regions than its type data counter part. New Guinea species on the leaf plot form basically a central core, while on the type data plot they are spread across the figure as a number of short branches. The outlined regions on this latter figure are also diffuse. The conclusion that must arise from all comparisons between leaf and flower based classifications is that the former is more consistent, shows a greater agreement with subjective taxonomic appraisal and indicates a greater separation of species into internally homogeneous but well separated clusters.

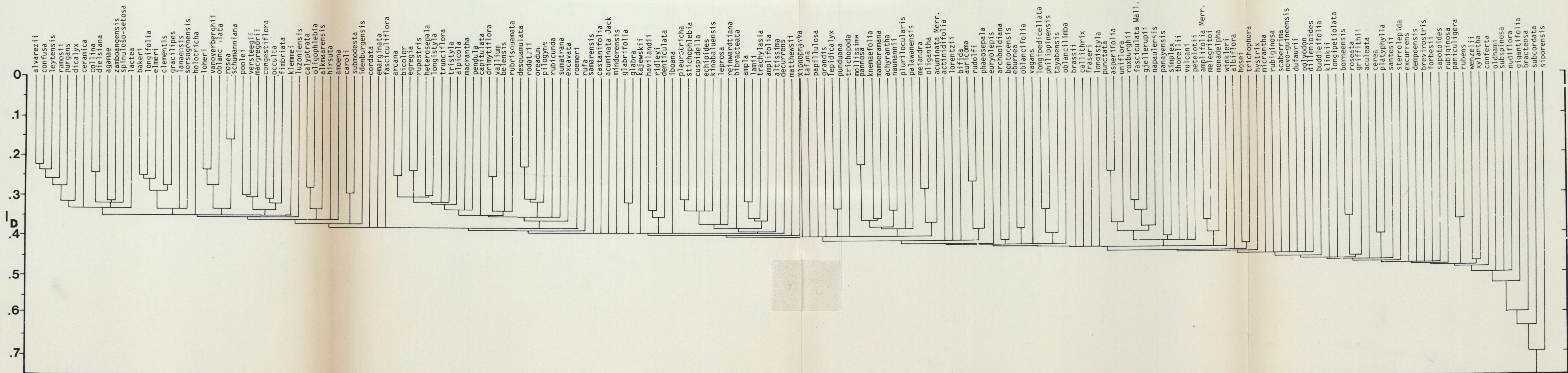
Final treatment of TYPES using the program TAXON was the construction of a phenogram and nearest neighbour tree for all species whose type descriptions had been obtained. This includes a total of 193 species. The value of doing this relates to the stability of classification. In a small set of species there is a high probability of misclassification, simply because the real linking species are not present. It would be expected therefore that an ordination which was apparent on a small species set should be even clearer on a large species set.

The phenogram, figure 69, has been produced from similarity values for 193 species scored on type descriptions. It was decided that similarities should be used for analysing type data because of

FIGURE 69. Single linkage phenogram for 193 species
of Saurauia using RANVARs, RANGES, I_D .

Uniflorae
Ramiflorae
Calyptratae
Pleianthae

a
b
c
d
e
f
g



an unexpected reversal in variances of similarities and distances. Similarities computed from range data for type descriptions had a higher variance than distances computed for the same data. This is at odds with findings for leaf data. The nearest neighbour tree for TYPES, RANVARs, RANGE, I_D (fig. 66) has been shown to be free of expected range induced influences that were so evident on the SPECIMENS equivalent (fig. 59). That fact, the reduction of variance of distances and therefore a reduced discriminatory power, has meant that RANGES and I_D were superior to D_I for classificatory purposes.

Figure 69 shows some distinct clustering of species. These clusters have been analysed carefully by comparing descriptions and specimens where available, of the species involved. In general this ordination shows a high degree of agreement with subjective taxonomic criteria. Although subjectivity is incompatible with the philosophy that I have extolled in this thesis, the choice of species cutoffs by objective disjunction based criteria is in general agreement with my "feeling" for the taxonomy of the genus. A classical taxonomic approach is totally subjective. Therefore I feel that it is valuable after an objective survey to make some classically oriented statements regarding the classification. This should be done to make the system more acceptable to taxonomists in general. The choice of a phenon line is a good place to start. The discussion of this in chapter 7 indicated that no one phenon line could be acceptable to all taxa. On the phenogram (fig. 69) however, it is possible to find a phenon line which is an average of all the individual cluster phenon lines. This line is fairly acceptable on a subjective assessment of taxonomic relationships between species and can be placed to minimise error at or near the .3 level. A more accurate assessment of the position of this line is given in the chapter on estimating the number of species in the genus (page

What it means is that based on type data alone I would have no hesitation in lumping those species that lie above the critical level. In order across the phenogram therefore the following groups are considered highly similar at the specific level.

- Group 1. alvarezii-confusa-leytensis-ramosii-purgans
- Group 2. collina-dielsiana
- Group 3. bakeri-longifolia-elmeri
- Group 4. clementis-gracilipes
- Group 5. loheri-vanoverberghii-oblongeolata
- Group 6. rosea-schumanniana
- Group 7. calyptrata-oligophlebia
- Group 8. arcana-bicolor
- Group 9. egregia-rupestris
- Group 10. drimytilflora-vallium
- Group 11. desquamulata-rodatzii
- Group 12. mollissima-pannosa
- Group 13. meiandra-oligantha
- Group 14. auricoma-rudolfi
- Group 15. punctata-asperifolia
- Group 16. cerea-platyphylla
- Group 17. paniculigera-rubens

Note that groups 16 and 17 are linked below the .3 phenon, but I consider that they are sufficiently disjunct clusters to be included in the list.

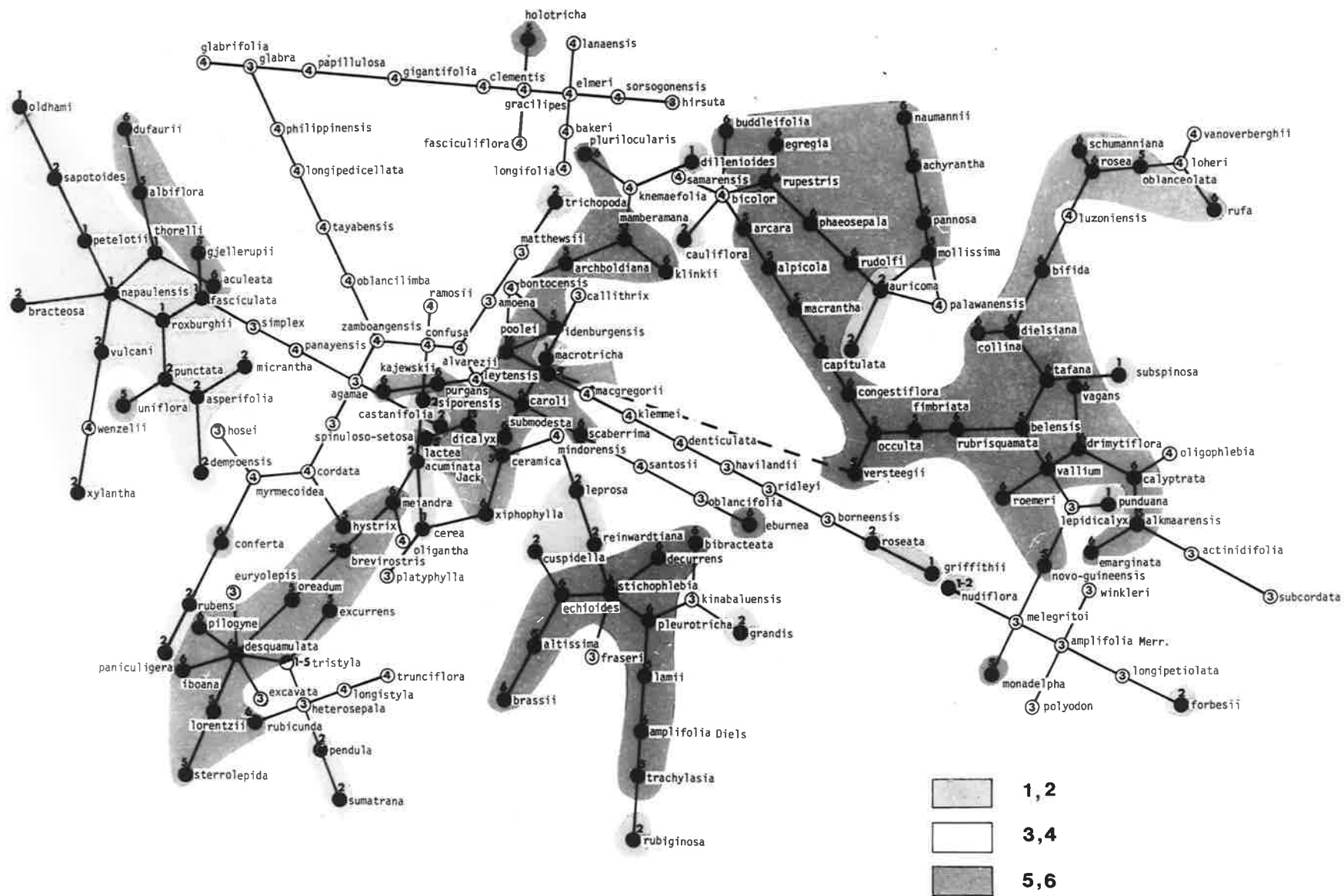
Group 1 contains mostly species from the Philippines. The species Saurauia confusa and S.leytensis were both compared in their type descriptions to S.ferox, a species which has been excluded from the phenogram because of the incompleteness of the type description. The species S.purgans is questionably placed here as it is a species from the Solomon Islands. Group 2 contains a pair of New Guinea

species, the first of which was incorrectly compared to S.brassii in its description. Group 3 contains species from the Philippines or Celebes (S.longifolia). S.bakeri is compared closely to S.clementis in its type description and likewise S.elmeri is compared to S.gracilipes from which they are both virtually indistinguishable. S.clementis and S.gracilipes form group 4 and were barely separated from group 3 by the arbitrary level of the phenon line as the two groups are linked at .3. Groups 3 and 4 could probably be combined, but will be left separate here because of some differences in length and density of hairs. Group 5 species are from the Philippines and Irian Jaya (S.oblanceolata). S.loheri has been compared in its description to S.clementis to which it is quite close. Group 6 contains two species from New Guinea and the Solomons which have been combined into one species, S.schumanniana by Diels in his description of that species. Group 7 is made up of species from New Guinea (S.calyptrata) and the Philippines. Both are distinguished by solitary flowers or almost so and are bracteate only in bud. Group 8 contains species of the Philippines and Irian Jaya (S.arcana) which are very similar on features of the inflorescence and indument. Distinguishing features of bracts are not given in the description of S.bicolor however, so that their grouping is open to some question. Group 9 species are both from New Guinea and indistinguishable by any features of specific importance. The same localities and taxonomic statements also hold for species of group 10. Species of group 11 are both from New Guinea and although virtually identical in many characters, appear to differ in features of the inflorescence and style number. As both of these features have been known to vary within species their separation into two species is questioned. Group 12 species are from New Guinea. The specimen on which S.mollissima was described appears to have been aberrant in possessing only two styles, otherwise the

species are indistinguishable on the descriptions. Species of group 13 are from the Philippines and New Guinea (S.meiandra). Apart from the shape of the leaf base, a character which appears to be environmentally influenced, the species are very similar. Species of group 14 are from Sumatra and the Bismarck Islands. The type description of the Sumatran species (S.auricoma) is very incomplete. Its placement close to S.rudolfi appears to be an artifact of this. However, based on the total though inadequate type data, these species are very similar. Species of group 15 are both from Sumatra and are considered to be the same. Although they were both described by Stapf, in the same publication a remarkable degree of non-standardisation of terminology is shown. That is, equivalent features are given different terms in the descriptions. This observation raises the question of how many species are artifacts of non-standard terminology. Species of group 16 are from India and the Philippines (S.platyphylla) but otherwise appear to be identical. Group 17 consists of two species from Malaya (S.rubens) and Sumatra (S.paniculigera) which appear to be identical.

The final ordination diagram produced for this section is a nearest neighbour tree for all 193 species scored for type data (fig. 70). If one recalls the ordination achieved with SPECIMENS and TYPES as summarised in figure 65, it will be seen that leaf data has consistently produced a better classification than type description data. This can be projected to a data set for a larger number of species. Because the number of species scored on type descriptions was a large proportion of the total number of species of Saurauia in the Old World, it was felt that a nearest neighbour tree constructed for those species would be a significant contribution and would justify the amount of work involved. Only 92 species were considered in detail for leaf data and only 44 of those had a sufficient number

FIGURE 70. Nearest neighbour tree for 193 Saurauia species scored for RANVARS, RANGE, I_D . Shaded areas indicate geographically defined clusters of species. The key gives region numbers. The regions are defined in the text. One cluster has been laterally shifted to avoid overlap. Movement and breaking point for the shift are indicated by a dashed line.



of specimens to place any reliability on the classification. Therefore the nearest neighbour tree in figure 65 stands as the basic framework on which an expanded leaf classification can be built. Completion of this aspect was impossible within the present project. Despite this, the type data nearest neighbour tree of figure 70 forms a basis for a classification of the genus in the Old World. It is a purely objective assessment of relationships between species based on all presently available information.

Clustering of species in figure 70 is strongly geographically related. This is consistent with the ordinations produced for subsets of species. Geographically defined clusters are delineated by shading. These regions are defined and discussed in the chapter on biogeography, but will be briefly mentioned here to facilitate discussion of the figure. Regions 1 and 2 encompass India, Burma, China and South-East Asia including peninsular Malaysia, Sumatra and Java. Region 3 contains Borneo and Celebes, with region 4 encompassing the Philippines. Regions 5 and 6 extend from Irian Jaya through Papua-New Guinea, the Solomons, Australia and Fiji.

Clusters have also been studied with regard to other characters such as generic section, style number, indument type and so on through most characters of the TYPES list. The general ordination is rather more weakly evident in the features of style number, indument and generic section. These are the traditional characters for dividing the genus. However by far the most consistent differentiation of clusters in figure 68 is by geographical region. Theories for why this may be so are presented in the biogeography chapter. It is still worth discussing some details of the plot here. Starting at the left hand side, one can see a large cluster of species dominated by species from region 1. It is interesting that some species,

S.vulcani, S.bracteosa and S.roxburghii, consistently clustered by leaf data are grouped here. To the right of this cluster is a long chain (bent at S.glabra to fit the page) of dominantly region 4 species but with several region 3 species. Note that in almost all cases, species linking this cluster to others are region 4 species when linking to region 5 or 6, and region 3 species when linking to region 1 or 2. The centre of the plot is dominated by region 5 and 6 species. This arrangement is reminiscent of the SPECIMENS nearest neighbour tree of figure 65. The very large cluster of region 5 and 6 species to the upper right of the diagram has been shifted laterally to avoid its overlapping other clusters. The linking line joining this to the rest of the diagram has been broken at S.versteegii. Direction and quantity of lateral displacement is indicated by a dashed line. Many species near the centre of this cluster are members of the Setosae, including S.dielsiana, S.purgans, S.bifida and S.schumanniana. The last major cluster lies between two groups of region 5 and 6 species and is represented by a straight chain of 3 and 4 species. Even here linking to other groups has region 4 linking to 5 and region 2 linking to 3.

Throughout this chapter it has been demonstrated that there are consistent ordinations of species defined by both leaf and type description data and that the ordinations of both show a strong geographical dependence. It also has been demonstrated that as discriminatory characters, classificatory characters and frequently scored characters in type descriptions, leaf characters are superior to a predominantly floral character set. This completes one of the aims of the present project, and that was to compare leaf and floral characters to demonstrate that leaf characters were comparable to floral ones in terms of yielding useful taxonomic characters. In appendix 4 the application of several other statistical procedures

to the present data are discussed and results compared with those achieved by the multistate similarity and distance measure. The following chapter contains a discussion of some concepts related to the information content of graphical representations of resemblance matrices, used in this chapter.

CHAPTER 9

INFORMATION CONTENT OF GRAPHICAL REPRESENTATIONS
OF RELATIONSHIPS BETWEEN OTUs

In the previous chapter, various graphical methods of presenting resemblance matrices were considered. These can be divided into two general classes, phenograms and connected graphs or nearest neighbour nets. The information content, information distortion and methods of construction of these plots will now be considered. Frequent reference will be made to figures illustrated in the previous chapter.

Phenograms

A single linkage phenogram such as those illustrated in figures 56 and 57, is dependent for its structure only on the level at which species first link to clusters. Generally the actual order of OTUs in a phenogram has very little significance and usually there is no information regarding where in the phenogram newly entered OTUs are linked. The procedure of constructing such phenograms involves finding the closest matched pair of OTUs and then adding successively more distantly matched individuals, adding them to a consistent side of the existing clusters. Newly added OTUs are generally linked to the centre of the horizontal links of the previous cluster. By this means the familiar one sided phenogram with symmetrically balanced clusters is produced. In such a plot not even the whole of the information for the first dimension of the similarity or distance matrix for each OTU is included. In an attempt to improve the information retention of such plots, the concept of nearest neighbour phenograms was developed. In such plots, each OTU is linked directly to the OTU with which it is most closely matched. This approach has been partly used before, by plotting newly linked OTUs so that they are tied by phenogram lines to a point on the preceding horizontal

link, as close to vertically beneath the actual linking OTU as possible. Such phenograms are presented by Gilmartin and Harvey (1976) and Small et al. (1976). This latter paper terms such a phenogram, "nearest neighbour." I cannot agree, however, that this represents a true nearest neighbour plot. The term "nearest neighbour" was first used in a taxonomic sense by Lance and Williams (1967) in reference to single linkage clustering. Although such clustering (the type used in the present study) uses nearest neighbour values to generate clusters, the inability of phenograms to demonstrate the nearest neighbour relationships graphically means that within the usage of the term in the present study, such plots cannot be considered "nearest neighbour." It is recommended that nearest neighbour be a term reserved for graphical procedures which can demonstrate such relationships and that the term single-linkage clustering be used for the clustering strategy that uses nearest neighbour values.

The first step in producing nearest neighbour plots in the present context was to overcome the traditional limitations of phenograms. Plots of Gilmartin and Harvey, and Small were therefore taken a step further by allowing not only the position of linking to define the link OTUs, but by allowing the direction of branching, combined with allowing link lines to transgress existing lines in the plot, to express the pairs of linking OTUs. Such a plot becomes more or less symmetrical about the best matched pair of OTUs which lies at the centre of the plot. This is illustrated in figure 58 for SPECIMENS, RANVARs, MEANS, I_D . Starting with the species RUBRI and TAFA, it can be seen that the next species to link is OCCUL, via TAFA. Following this BIFID links via OCCUL. The direction of branching then changes as SCHUM is linked to TAFA. In such a manner a phenogram which contains the whole of the information relating to nearest neighbour relationships within the set of OTUs can be constructed. The direction

of branching is determined by the side of the first linked pair that the linking species of additional species is found to lie. In this phenogram very clear clustering of species can be seen. Note the distinct cluster of the species MALAY, SCABE, BRACI and VULCA. A graphical convention which must be followed in producing such plots is the means of handling cross overs and horizontal links which would be superimposed. For cross overs I have used a fairly standard electrical circuitry convention, that of little "bridges" for indicating cross over without joining. These should be used on the vertical line of a crossing pair since the verticals are usually easier to interpret than the horizontals which would be confused by not being straight. To avoid the problem of horizontal lines being superimposed when OTUs are found to link at the same level, I have found that drawing the two lines parallel and close together usually suffices. Although such lines could be misinterpreted as meaning that the OTUs link at different levels, the spacing of the lines can be kept to such a small distance that such a difference in linking level would hardly be significant.

The major problem with such phenograms is that they are hard to interpret at a glance. Although a requirement such as ease of reading may seem unimportant, I decided to find a less ambiguous method of presentation. The first solution to this involved abandoning the constraint of a unidirectional plot of similarity. The new approach can be likened to swinging the OTUs in the nearest neighbour phenogram out of the plane of the page on hinge lines represented by the verticals of the phenogram. In the simplest case the criterion for choosing the final relative positions is simply that distances between OTUs linked to the same OTU via the same vertical should be as far apart as possible. Thus if three OTUs are linked to a vertical

then they should end such that they are 120° apart. Visually such a plot would resemble a "mobile" sculpture with names attached to the ends. The final step to producing a two dimensional representation of nearest neighbour relationships involves letting the length of the horizontals equal the value of resemblance. Such a plot is illustrated in figure 65. Comparison of this figure with figure 58 reveals that they are identical in expressing the first dimension of OTU relationships. Such a plot differs from the connected graphs and nearest neighbour plots described in Sneath and Sokal (1973) and applied by Lance and Williams (1967) and Rohlf et al. (1971), in that it seeks to maximise the distance between OTUs not linked directly. In the plot, closed loops arise when a member of a branch of the "tree" is linked equally to two species. In such cases an angular distortion may arise in the plot. Angular distortions have also been allowed in those cases where a new branch is likely to cross an already constructed one. Where there is a choice of direction in branching such as for the species RUBEN and FRAG arising from PANAY (fig. 65) this can generally be arbitrary unless there is some other constraint like interfering branches or multiple simultaneous linkage producing closed loops. An example of this last case is the direction of branching of PENDU from PURG being constrained by its link to GJELL (also fig. 65).

The nearest neighbour plots described here differ from the shortest spanning trees of Kruskal (1956), Whitney (1972) and Rohlf (1970, 1973) in that these latter graphs are constructed on OTUs already ordered in a finite space by such methods as principal component analysis or discriminant analysis. Shortest spanning trees are produced by joining the closest OTUs in a defined multidimensional space. Such plots are constrained by being plotted on axes which are

defined in terms of the overall relationships between all OTUs.

The existence of graphical methods which rely on the relationships between OTUs in the full multidimensional space raises the question of the accuracy of plots constructed on only nearest neighbour values. It is important to be aware that more than a single dimension of information is demonstrated in the nearest neighbour graphs. This can be demonstrated by considering a situation where OTUs form a distinct cluster with a few outliers. The nearest neighbour clustering procedure begins with seeking the most closely similar pair of OTUs. These are represented as points on a page such that their linear separation is scaled to represent the actual value of similarity. Either dissimilarity or distance values are used here so that similar OTUs are closely clustered. The two member cluster formed at the first iteration is then added to by successive iterations such that at the next step a new OTU closely linked to one of the existing ones is added. As the cluster is built up, it becomes evident that many of the comparisons are redundant. That is, they are between OTUs already linked by other OTUs. After many iterations through mostly redundant values of the similarity matrix the outlying OTUs are added. For the outliers this linkage represents a nearest neighbour link. However, to the member of the cluster to which the outlier is linked it may be a high order neighbour link. Because of this a nearest neighbour graph is unlikely to be severely distorted from the real multi-dimensional situation because in effect it has sampled through the dimensions. This argument can also be applied to phenograms. It would be expected that a phenogram for a very heterogeneous set of OTUs would be less distorted from the multi-dimensional space than a phenogram for a very homogeneous group. A cophenetic correlation coefficient for a heterogeneous set of OTUs

will be higher than that for a homogeneous "complex" of OTUs. This can be stated in another way. An ordination of homogeneous OTUs will be more sensitive to the addition of further individuals or to slight changes in similarity values than would an ordination of distinct entities. When entities are distinct the pattern of their relationship is less easily masked than when OTUs are more or less equally similar to a large number of other OTUs. This circuitous buildup to the question of the ordination of Saurauia species can be illustrated by a hypothetical example. Let us say that there exist a set of OTUs which cluster into three distinct groups, morphologically homogeneous within each cluster and with large distances between clusters. The three clusters can be seen to behave like three individuals. It will be possible to represent their relationships on two dimensions with very little distortion. The distances between individual OTUs of one cluster and another will be close to the distances between cluster centroids. This example is illustrated in figure 71. The angle subtended by lines joining OTUs between clusters gives a measure of this approximation. Such an example can be seen to be presentable in fewer dimensions than the full $(n-1)$. This is because most of the dimensions represented in the matrix of distances will be insignificant. The structure is essentially flattened in hyperspace.

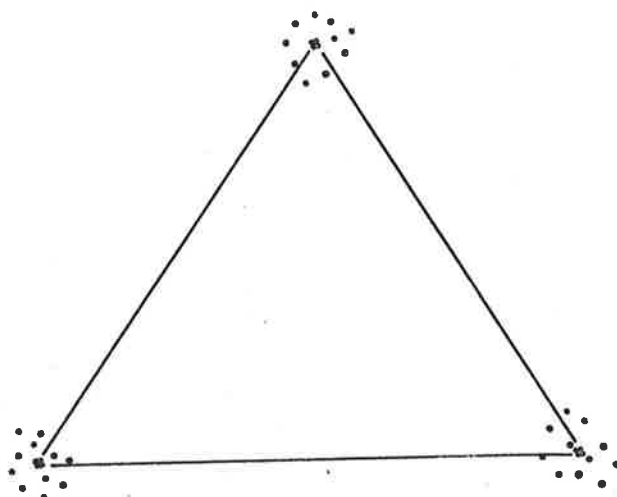


FIGURE 71. Three hypothetical clusters of OTUs, widely separated in attribute space. Cluster centroids are joined.

The converse argument is true for a very homogeneous set of OTUs. Such a group form essentially a single cluster which approaches a hypersphere in shape. The two dimensional representation of such a set of OTUs will involve many more serious distortions because the choice of nearest neighbour links over other links involves a small numerical difference. The variance of distance values for a homogeneous set of OTUs will be low compared to that for a heterogeneous set. In the case where this variance is zero, the relationships between any three OTUs will be perfectly equilateral and the hyperspace will be isotropic. Given that the character scores for a set of OTUs define a hyperspace then the simplest statement of the function of numerical taxonomy is that it seeks to quantify the anisotropy of hyperspace.

The revelation that homogeneous sets of OTUs approach the equilateral situation allows critical interpretation of the nearest neighbour plots illustrated in the previous chapter (figs. 59-61, 65, 66-68).

Higher order neighbour plots

Strict nearest neighbour plots can be produced with zero distortion in two dimensions. The method does not seek to illustrate higher order relationships, however. Second nearest and higher order neighbours present greater difficulties in their two dimensional representation. There are no real limitations to producing two dimensional plots of first and second nearest neighbour relationships on two dimensions except the greatly increased work load. It is also important to be aware of the possibility of alternate ordinations with nearest neighbour plots. In the case of first nearest neighbour trees, the criteria for deciding branching direction have already been discussed. The geometrical form of the plot shown in figure 65 is,

however, clearly not unique. There are many alternatives that could be produced by simple rotations at branching points. It is important to emphasise that all of the reliable information contained in such an ordination is contained in the lengths of linking lines. The spatial relationships of OTUs beyond the level of the distance to the adjacent OTUs is not reliably expressed. To test how much more information was available from the second nearest neighbour relationships, plots of first against second nearest neighbour values for 42 OTUs for both TYPES and SPECIMENS, RANVARS and MEANS for I_D and D_I were produced. These are illustrated in figures 72 and 73. Figure 72 is plotted for dissimilarity coefficient values for both TYPE and SPECIMEN data. The high correlation between first and second nearest neighbour values is demonstrated by the linearity of the cluster. Since the correlation itself is of less importance than the scatter of points, a different statistic has been used to compute dispersion of values. The set of lines passing through the origin of the plot and intersecting the scatter of points represent the lines of different ratios of second to first nearest neighbour values. Thus when the values are identical, that is an OTU is linked equally to two other OTUs, the 45° line (marked 1) applies. Where second nearest neighbour values are .9 of the first, the line marked .9 applies and so on. If the frequency of points falling in the field delimited by the ratio lines are plotted, the histograms included on the main plots are produced. For figure 72, this histogram clearly shows strong clustering of nearest neighbour values in the region of equality. On the plot, black dots and black columns in the histogram represent SPECIMENS, while open bands and triangles are for TYPES.

Figure 73 is similar to the previous one and has been plotted for the same data, but is based on distance values. The visual

FIGURE 72. Relationships between 1st and 2nd nearest neighbours for Saurauia species scored as RANVARs and using I_D . The inset histogram illustrates the frequency of ratios of 1st to 2nd neighbour values. Triangles and open bars in histogram represent the species scored for type description data. Black dots and bars represent species scored on leaf data.

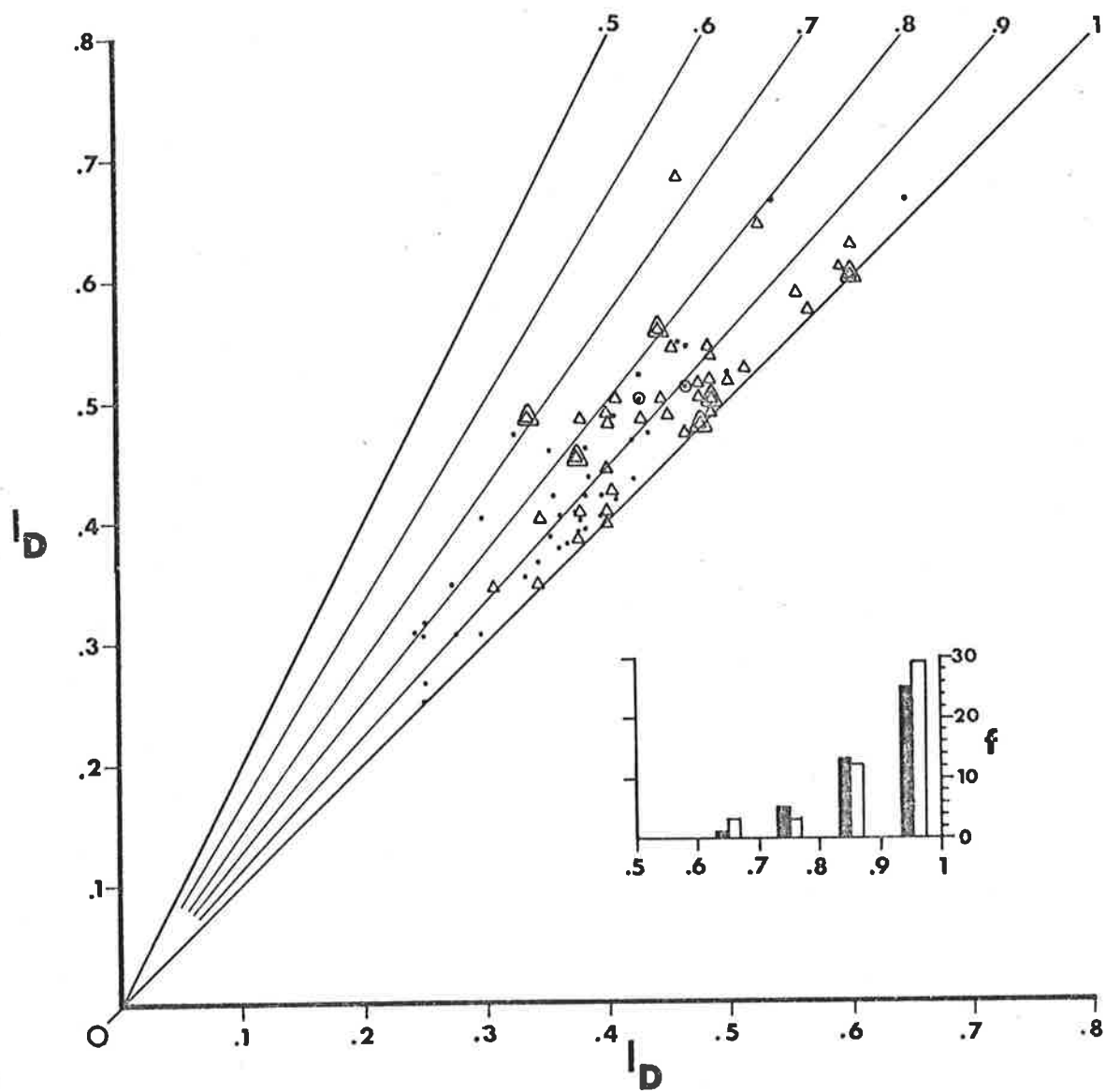
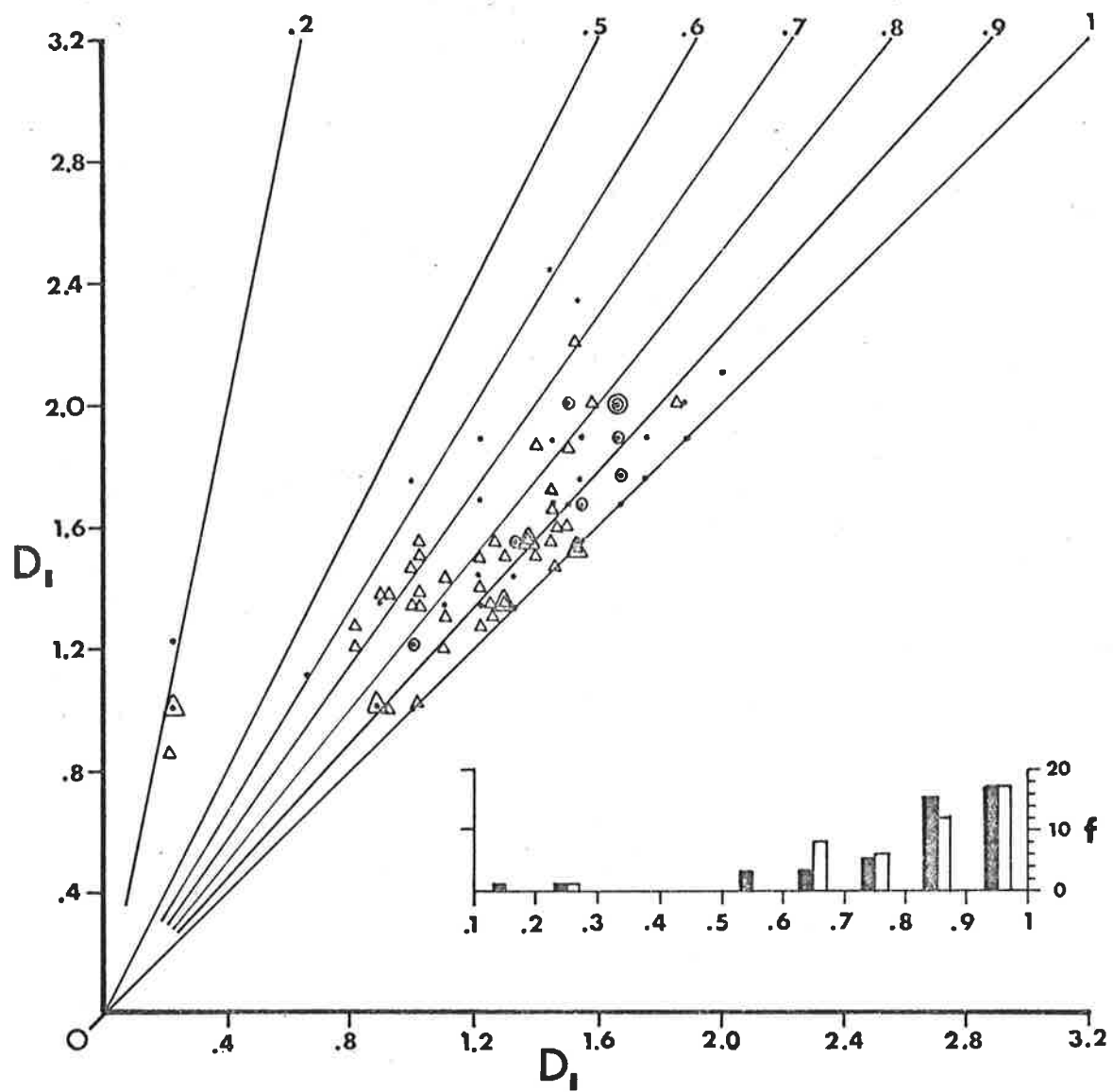


FIGURE 73. Plot of 1st and 2nd nearest neighbour relationships for Saurauia species using RANVARS and D_I . Symbol convention as for figure 72.



impression one gains from this second plot is that the dispersion of points is much higher than for figure 72. This can also be seen in the histogram accompanying the plot. A measure of dispersion for each histogram can be calculated by the following means. On the assumption that each histogram represents half of an approximately normal distribution, variances can be calculated on the whole distribution reflected about the line where the ratio of first and second neighbour values is equal to one. If this is done for each distribution, the following table of F values can be computed for a pairwise comparison of variances between alternate plots (Table 9).

SPEC I _D				
SPEC D _I	4.923			
TYPE I _D	1.060	4.648		
TYPE D _I	3.451	1.428	4.391	
	SPEC I _D	SPEC D _I	TYPE I _D	TYPE D _I

TABLE 9. Variance (F) ratios for pairwise comparisons of histograms of nearest neighbour ratios.

For significance at 1%, the F-ratio computed over 90 d.f. must exceed 1.6. At this critical value only the differences between variances of I_D and D_I plots are significant. The variances for SPECIMENS against TYPES within either an I_D analysis or a D_I analysis are not significantly different. These results can be interpreted in two ways. Firstly they say something about the homogeneity of the set of OTUs based alternatively on I_D and D_I values. Using dissimilarity coefficient values it could be expected that a poorer discrimination of OTUs would result, when compared to that produced by the distance measure. Referring back to figures 56, 57 and 58 this can be seen to

be the case. The separation of clusters of OTUs is much clearer on the last mentioned plot. The second interpretation of these results relates to the accuracy of first nearest neighbour plots in expressing the full multi-dimensional relationships between OTUs. From the discussion of this topic in previous pages it is easy to see that the nearest neighbour plot using D_I will be the most reliable. This can be checked by plotting various nearest neighbour plots, including at least first and second neighbour relationships and inspecting them for isotropy expressed as a tendency to equilateral relationships.

Constructing such plots is very much more complex than the simple one order neighbour trees. The method is essentially the same but involves more critical decisions regarding the direction of branching in particular. The angular relationships between branches are defined not simply by a criterion of maximisation, but by the actual values of similarity to the second nearest neighbour OTUs. When only second nearest neighbour relationships are to be added, the plots are relatively simple. The graphical method involves using the nearest neighbour plot to judge the first dimension of relationships and then swinging branches about their branching points. The branching points behave like joints in a pin-jointed structure in an engineering sense. That is, they are free to swing in the two dimensional plane of the paper. The actual relationship of a branch to the pre-existing structure is determined by the next shortest line that links it to an OTU besides its nearest neighbour. The simplest method of doing this is to find the nearest neighbour by reference to the matrix of similarities and then using that OTU as a centre, construct a circle of radius equal to the scaled similarity value. In the simplest case where a third OTU is being located with respect to a pre-existing cluster of two, the third OTU must lie on the circumference of this

circle. By then finding the second nearest neighbour and constructing a suitable circle using it as a centre the intersections of the circles so produced will define the possible positions of the new OTU. In the three OTU case it can be seen that there are two intersections of the circles. For three OTUs the choice of which of these positions will be used to represent the new OTU is purely arbitrary. When a fourth OTU is to be added, an extra dimension of control is given to the position of the new OTU. In such a case the third nearest neighbour relationships can be used to locate the new point. If a single point can be found where the circles constructed about the first, second and third nearest neighbours are intersecting, this point will define the location of the fourth member of the cluster. In general such a point will not be found. The practice I have followed in these cases has been to generate the second nearest neighbour defined points with no distortion, then construct the third nearest neighbour circle and choose that second neighbour point that lies closest to the circle. By this method the distortion on the third neighbour location can be minimised. The graphical constructions required to plot four test OTUs are shown in figure 61. The distance matrix for these is given below.

Bifid				
Occul	1.000			
Schum	1.333	1.778		
Tafa	1.222	1.000	1.222	
	Bifid	Occul	Schum	Tafa

This is a distance matrix for SPECIMENS, RANVARS, MEANS for four Saurauia species.

Figure 74 will illustrate the definition of some terms associated with nearest neighbour topology.

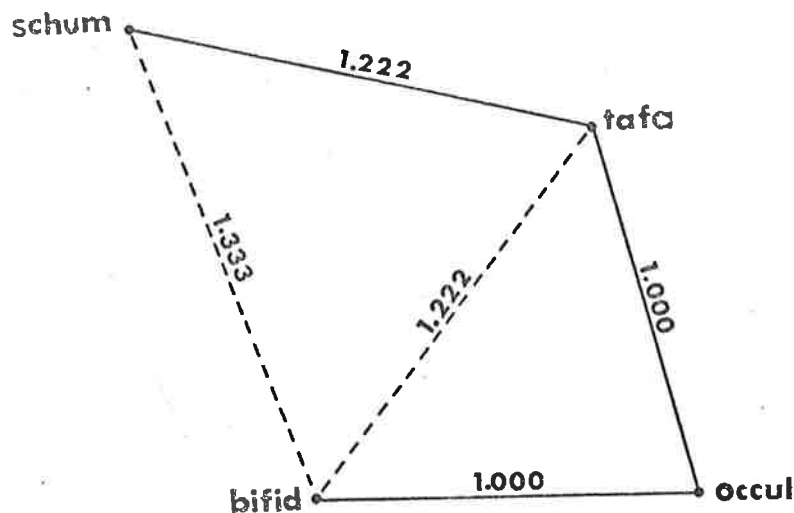


FIGURE 74. Four Saurauia species plotted using first and second nearest neighbour similarity values. See text for further details.

On the figure nearest neighbours are connected by solid lines. However, it is clear that these express different relationships between OTUs. The line joining BIFID and OCCUL represents a mutual nearest neighbour connective. That is both these OTUs have the other as their nearest neighbour. The link is therefore bidirectional. Note that Tafa and OCCUL are also mutual nearest neighbours because the distance between them is the same as it is between BIFID and OCCUL. OCCUL is termed a "local centre" because it is strongly linked to two OTUs. As this figure illustrates, it is possible to find OTUs which are simultaneously mutual nearest neighbours to more than one OTU. The angle between the two first formed branches of the figure is now defined by the second nearest neighbour relationship between unconnected members of the cluster, i.e. BIFID and Tafa. Notice that if the previous two branches had been of unequal length (for example Tafa had been further from OCCUL than BIFID was), then the link between Tafa and OCCUL would have been unidirectional. That is Tafa

would have OCCUL as its nearest neighbour but OCCUL would have TAFA as its second nearest neighbour. It would be possible in plotting such relationships to use a coding of links to indicate the difference between bidirectional and unidirectional nearest neighbour links. However, I have decided that such relationships can be judged from the equality of the subtended angles and so to avoid confusion I have used solid linking lines wherever such a link is at least unidirectional nearest neighbour. The angle subtended by the branches arising from OCCUL can now be defined by the distance between BIFID and TAFA. This is a second nearest neighbour link and in the example is bidirectional, that is BIFID and TAFA are mutual second nearest neighbours. The lines for second neighbour connectives are represented by dashes.

The link to the fourth member of the cluster can now be defined. Note that the distance to SCHUM from TAFA is the same as the distance from BIFID to TAFA, yet the lines are coded differently. This is because the SCHUM-TAFA connective represents a nearest neighbour relationship for SCHUM, but only a second nearest neighbour relationship for TAFA. This is in fact a case of simultaneous but non-mutual second nearest neighbours, because although BIFID and TAFA are mutual, TAFA and SCHUM at an equal distance are not. SCHUM can be seen to be first linked via TAFA. The direction of the branch to SCHUM is defined by its second neighbour, BIFID. SCHUM is a third nearest neighbour to BIFID, but has BIFID as its second neighbour. For this reason the broken line joining them is a second neighbour connective. Without incorporating the final remaining relationship it can be seen that there are two alternative positions for SCHUM. The second position is shown by a pair of intersecting arcs, near OCCUL. If the distance from OCCUL to SCHUM is plotted as a circle, it can be seen that the final position of SCHUM on the plot is less distorting of

the relationship between it and OCCUL. The actual distortion of the SCHUM-OCCUL distance is only +8%. If this link had been undistorted, it could have been included as a third neighbour link, represented by a dotted line.

The graph of figure 74 has an extremely high connectivity, that is five of the six possible connecting lines can be shown without distortion. In fact the distortion of the sixth is minimal. The cophenetic correlation coefficient of this clustering would be very high. In the three OTU case it would in fact be 1.0. The converse, a value approaching zero would be obtained when all connectives were of the same length for a large number of OTUs such that only first and second neighbour relationships could be shown for adjacent points. If all the values are the same then the diameter of the hypersphere is equal to $1/\sqrt{3}$ times the distance value. In the three case example, this is represented by an equilateral triangle. For the four case example it becomes a tetrahedron, for five it is a pentahedron (a non-existent structure in three dimensions) and so on. The OTUs in such a space are arranged on the circumference of the hypersphere. The degree of elongation of the hypersphere and thus the heterogeneity of the OTU set and also the degree of distortion of the cluster in a two dimensional representation, can therefore be estimated from the angles between OTU connectives.

The angles between OTUs must firstly be measured for each OTU and its first and second nearest neighbours. These do not actually have to be physically measured on a plot since they are determined by the distances between the OTUs. A measure of heterogeneity of a similarity or distance matrix can be expressed in terms of the variance of the values. An alternate approach is to measure the heterogeneity of a set of OTUs based on nearest neighbour relationships. Let us start with the three OTU case. The heterogeneity of the OTUs

is reflected in the variance of distances between them and thence to the difference between the two dimensional configuration and the equilateral situation. As the heterogeneity increases, the relative difference between largest and smallest values of distance increases. A measure of heterogeneity can incorporate this effect. Heterogeneity cannot be measured on less than three OTUs. A cluster of two OTUs has no reference with regard to the relative degree of separation of the points. By summing all the ratios of sides of the triangle formed by linking three OTUs according to nearest neighbour criteria, and then dividing this sum by six (the number of ratios), a measure of average relative length of the sides is produced. For the example in figure 74 the heterogeneity of the triplet (TAFA-BIFID-OCCUL) would be $(1.222/1) + (1.222/1) + (1/1.222) + (1/1.222) + (1/1) + (1/1)$, with the whole sum then divided by six. The value of heterogeneity in this case is 1.0134. The value for an equilateral triangle is 1.0. If values of heterogeneity are computed for isosceles triangles where the ratio of longest to shortest sides is progressively increased, the values of heterogeneity are found to be almost linearly related to that ratio. Only at low values of heterogeneity does the slope change significantly. The values for a number of examples are given below in Table 10.

Distances between OTUs	Heterogeneity
1,1,1	1
1,2,2	1.166
1,4,4	1.75
1,5,5	2.066
1,10,10	3.7
1,20,20	7.0166
1,50,50	16.84
1,100,100	33.503

TABLE 10. Values of heterogeneity for OTUs arranged at the vertices of a triangle where the lengths of the sides are equal to a value of dissimilarity.

If standard deviations for each set of three lengths are computed, it is found that the heterogeneity is also almost linearly related to standard deviation. The overall heterogeneity of OTUs in a nearest neighbour plot could therefore be evaluated. The expression for heterogeneity of three OTUs i , j and k can be written

$$H_{ijk} = \frac{1}{6} \sum_1^6 \frac{d_{ij}}{d_{ik}}$$

where d_{ij} and d_{ik} are the distances between pairs of OTUs. To compute overall heterogeneity the mean value over the matrix for all the three way comparisons can be found.

The computational loads for such a procedure can be quite high, especially when coupled with the prior need to carry out the clustering of OTUs. Therefore, because a suitable computer program was not available for the present study, only subjective appraisal of homogeneity has been attempted. An estimate of the scale of the co-phenetic correlation coefficients for various nearest neighbour trees and phenograms can be made from figures 72 and 73. The correlation coefficient for plots produced from RANVARS, MEANS, D_I will be the most satisfactory. These plots include figures 65 and 70 which best illustrate the relationship between species.

CHAPTER 10

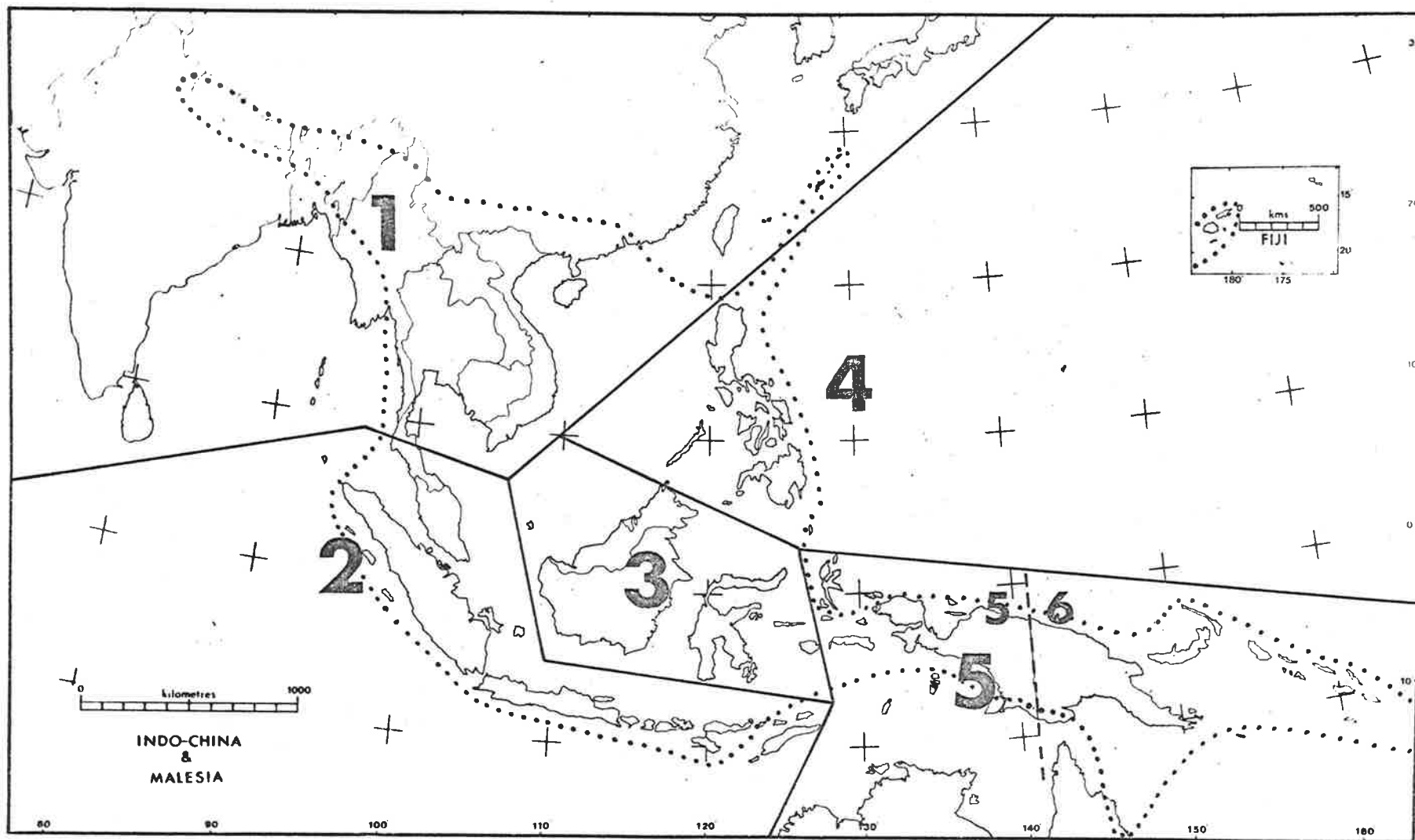
BIOGEOGRAPHY AND EVOLUTION OF THE
OLD WORLD SPECIES OF SAURAUIA

The genus Saurauia is widespread in the tropics of both the New and the Old World. Approximately 70 species occur in Central and South America, from Mexico to Bolivia. The Old World species number close to 200 and are found from Nepal through Indo-China and Malesia to Fiji (fig. 75).

Such disjunct distributions are known for many genera of flowering plants (van Steenis and van Balgooy, 1966). It has been postulated that this affords evidence for a continuous land bridge between Alaska and Siberia in the past (Martin, 1977). Geophysical evidence suggests that the Alaskan and Asian continental margins fused in the Late Cretaceous (Smith and Briden, 1977). It is likely that eustatic sea level changes have allowed this bridge to be periodically exposed through the Tertiary. The existence of a land bridge as early as the Late Cretaceous is further supported by the similarity of Late Cretaceous faunas of Asia and North America (Kielan-Jaworowska, 1974).

The palaeobotanical evidence for Saurauia in Asia and America is questionable. There are no records of Saurauia from the Cretaceous. However, impression remains ascribed to Saurauia have been reported from Tertiary floras of Alaska (Hollick, 1936), Wyoming (Leopold and MacGinitie, 1972), Croatia (von Ettingshausen, 1870) and France (Langeron, 1900). There is considerable doubt as to the identity of Hollick's specimen, however, based as it is on a single fragmentary specimen, lacking any cuticular remains. The leaf lacks the typical acuminate apex of Sauruaia and possesses no features which distinguish it from numerous other genera. Hunter (1966) compared the fossil to

FIGURE 75. Biogeographic regions defined for the Old World species of Sauruaia. The documented distribution of the genus in the Old World lies within the dotted line. Straight lines mark regional boundaries.



several Central American species but concluded that "it seems very unlikely that the specimen in question represents a Saurauia." Leopold and MacGinitie do not figure their fossil Saurauia so that I can say nothing about its identity. Likewise I have not seen the specimens from Croatia (von Ettingshausen, 1870) or France (Langeron, 1900), but Hunter considered that this latter leaf resembles Saurauia tristyla DC. As most identifications of this extant species are doubtfully correct, this suggested affinity can be considered of little meaning.

Fossil seeds referred to Actinidia have been described from the Eocene of Britain (Chandler, 1962). The seeds of Actinidia closely resemble those of Saurauia, and appear to be distinct from those of any other genus. Whichever genus is actually represented, this at least appears to be a reliable record of the family from the Tertiary of Europe.

Until the fossil record is better elucidated, the evolutionary history of the genus must be deduced from the living species. This has been approached, in the present study, by a consideration of the morphological relationships within the genus. The present biogeographic range of Saurauia species can be divided into a number of geographically definable regions. The expectation is that the morphological similarities within and between those regions are dependent on the evolutionary history of the genus. A detailed study of these relationships should allow the pathways of diversification to be traced.

The bases for the classification of the Old World species are discussed elsewhere, but it is useful at this point to redefine the major divisions of the genus. Diels (1922) proposed a number of sections for the Indo-Malayan-Australasian species of Saurauia. These sections were defined in terms of the position and form of the

inflorescence and the nature of the indument on floral and vegetative parts. For the purposes of this study these sections have been retained (Chapter 6). A separate division is made, however, on the basis of style number, for reasons which have already been considered (Chapters 8 and 9). In this chapter it is proposed to define a number of geographic regions which are based on intuitively determined geographical isolation and on documented endemism.

Biogeographical divisions

Using the criterion that a meaningful regional boundary should reflect a real barrier to the mixing of adjacent floras, five major regions are proposed (fig. 75). The boundaries mark seaways or lowland areas and enclose more or less continuous land masses or island chains. Region 1 includes India, Burma, China and South-East Asia (excluding peninsular Malaysia), Hainan and the Ryukyu Islands. This region is separated from adjacent regions 2, 3 and 4 by the lowland region of peninsular Thailand and the South China Sea. Region 2 encloses Malaya, Sumatra, Java and the Bali-Flores island chain and is separated from region 3 by the Java Sea. Region 3 contains Borneo and Celebes and is delimited from regions 4 and 5 by the Sulu, Celebes, Molucca and Banda seas. Region 4 encloses the Philippines and Banguay Island which lies off the northern coast of Sabah. Banguay Island has been included in this region rather than in region 3 because of a single species of the Philippines which occurs there. Region 5 extends from the Moluccas and New Guinea to Queensland, the Bismarck and Solomon Islands and Fiji. Of all the regional boundaries, all but one are fairly broad seaways. The boundary between regions 1 and 2 runs across the narrow part of the Malay peninsula near Surat Thani in Thailand. This is a lowland region and appears to be an effective barrier to the dispersal of Saurauia species between Malaya

and Indo-China. Most of these species are found in uplands.

Endemism in the genus

Of 212 species surveyed from type description and published locality data, 96% are endemic to one or other of these five regions (fig. 76). Of those species that transgress regional boundaries, only one (S.tristyla DC) is recorded from all regions. This species is also doubtfully a single species across its whole range. Hoogland (1977) states that the records of S.tristyla from Malaya and Indo-China are most probably referable to the species S.fragrans Hoogl., S.pentapetala (Jack) Hoogl., and S.roxburghii Wall. S.tristyla appears to be restricted in reality to the Moluccas and Celebes.

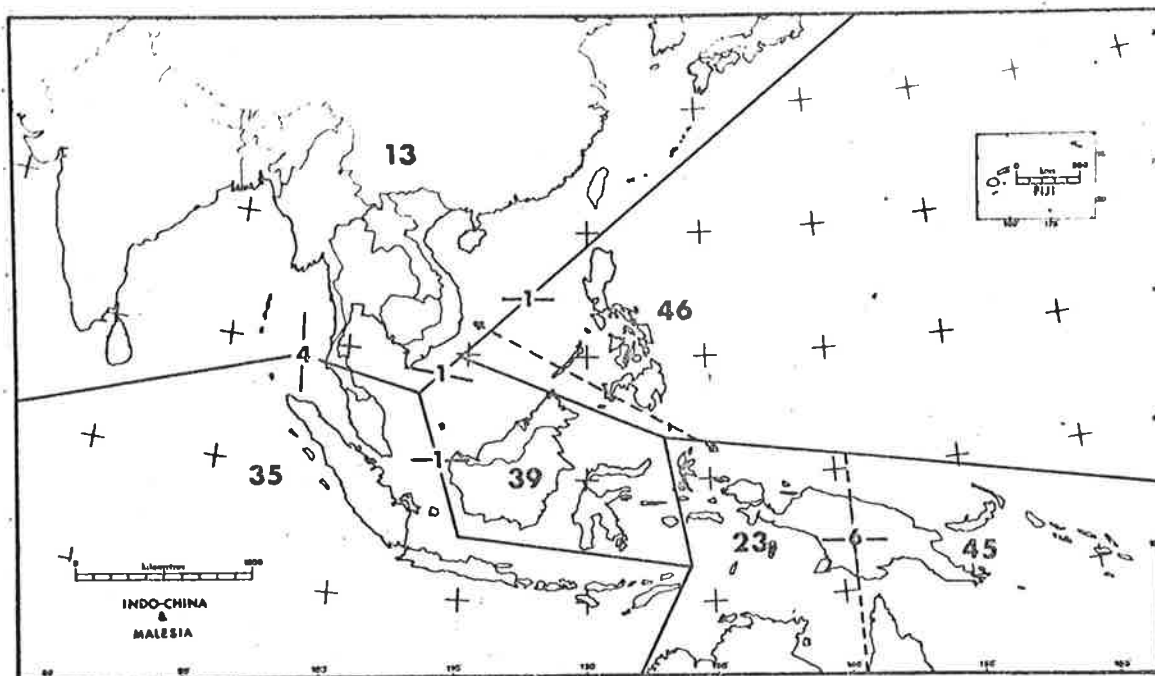
Of the remaining three species in common between regions 1 and 2, S.roxburghii and S.napaulensis DC are widespread through South-East Asia, while the other species, S.leprosa Korth., barely extends into Thailand.

The distribution of species in region 5 indicates an unexpected degree of local endemism within the region. Of the 74 species, only six are in common across the border between east and west New Guinea. Four of these species are known from the adjacent regions of west and northern Papua-New Guinea and east-central Irian Jaya. Of the two remaining transgressive species, one (S.conferta Warb.) is widely distributed throughout New Guinea and the Solomon Islands, while the other (S.novo-guineensis Scheff.) is disjunctly distributed in Irian Jaya and the Solomons. On the basis of this endemism, region 5 has been subdivided into regions 5 and 6, with the boundary lying on the border between Irian Jaya and Papua-New Guinea (fig. 75).

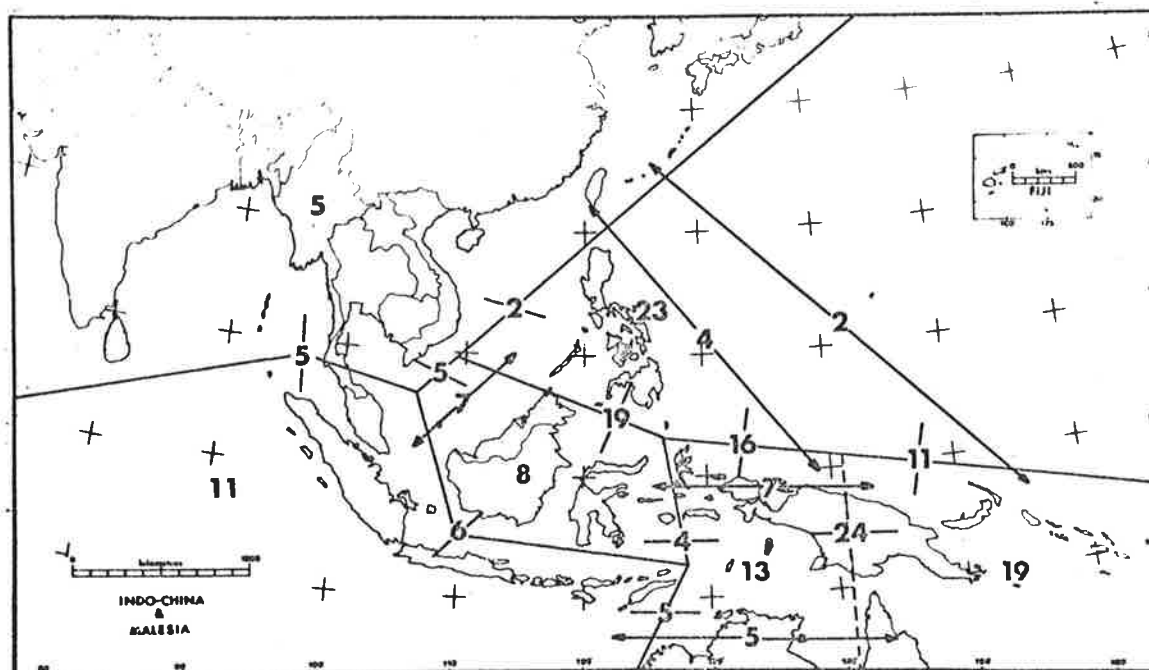
The endemism of region 6 cannot be explained in terms of geographical isolation. Most of the New Guinea species are found along the central mountain chain which is continuous between regions 5 and 6. Some other factor must therefore account for the phenomenon.

FIGURE 76. Endemism of Saurauia species. The numbers in regions indicate the species endemic to each region. Numbers lying on boundaries of regions or on arrows connecting regions are species found in more than one region.

FIGURE 77. Locations of nearest neighbours for 192 Saurauia species. Numbers within regions indicate the number of species that have their nearest neighbour species based on type description data, in the same region. Numbers transgressing boundaries indicate species with nearest neighbours in regions other than their own.



76



77

Throughout the history of botanical collecting in Malesia, east and west New Guinea have been separated by a political boundary. For the most part collectors of east New Guinea plants did not collect in west New Guinea and vice versa, and their collections were lodged in different herbaria. It seems reasonable therefore to suggest that the high degree of endemism is an artifact which does not have a real taxonomic determinant. It is likely that taxonomic revisions of these species will demonstrate considerable synonymy. In the case of endemism in other geographical regions, it is certain that political boundaries have had a similar effect. For this study, however, it was proposed a priori to take the species as they presently stand.

Two approaches can be taken to analysing the interrelationships between the species of different regions. Firstly, the similarities of species within regions can be compared to similarities between regions, either by pairwise comparisons of all species, or by considering only the pairs of most similar species. Secondly, consideration of the geographic distribution of the sections in the genus should yield information on where endemism is likely to be significant and aid in the documentation of the evolutionary history of the genus.

Similarity between species

The first approach involves the pairwise computation of similarity between all species. The methods and data for this procedure are given elsewhere and involve the use of I_D and the computer program TAXON. Data from 192 type descriptions were used to compute a similarity matrix of 18,336 comparisons. To reduce computational loads, only those 1,021 values of similarity greater than .5 were selected for further analysis. This has no significant effect on ordinations produced from the similarity matrix. A useful starting point for analysing the matrix is to select for each species the highest simi-

larity value to another species. The species pairs produced are described as nearest neighbours. Detailed analysis of nearest neighbour values has already been carried out with regard to dendrogram generation. Taken in the context of a biogeographical study, however, these values contribute important information with regard to the taxonomic homogeneity of a regional group of species.

Nearest neighbours

In a large genus such as Saurauia where there is considerable local diversification and an apparent high degree of endemism it is reasonable to expect that nearest neighbours will mostly be in the same geographical region. Figure 77 illustrates the location of nearest neighbours for 192 species. Numbers within regions are the number of species that have their nearest neighbour in the same region as themselves. Numbers crossing regional boundaries are the number of species that have their nearest neighbour in another region than their own. Only 41% of species have their nearest neighbour in the same region as themselves. This nearest neighbour endemism is strongly at variance with the distributional endemism derived from type descriptions and collection data.

Clearly this supports the idea of similarity between groups of species in adjacent provinces. The degree to which species have nearest neighbours in different regions is however so great that it is necessary to accept the possibility that there are also a significant number of species transgressing the boundaries.

Accepting that the local endemism of the genus is to an unknown degree an artifact of collections, it should be possible to set up a model which will give some idea of the actual number of real species of Saurauia in the Old World.

Number of species

Given the numbers of endemic species and the numbers of species with nearest neighbours in the same province, adjacent provinces and non-adjacent provinces, the following assumptions can be made on which to base a model. (1) There is some degree of synonymy in the genus such that the actual number of species is less than the number that currently stand in the literature. (2) If a species transgresses a regional boundary, it is most likely to be described as a separate species in each region. (3) When further specimens of a previously described species are collected from the same region as the type, they are more likely to be assigned to that species than to a new one. If this is the case for all species, then the model for the number of actual species can be described in terms of the contribution of nearest neighbours within regions, between adjacent regions and between non-adjacent regions. An example will serve to illustrate this reasoning.

Two specimens of what is in reality the same species (A) are collected, one from each of two geographical regions (a and b). The specimen from a is described as species A. The specimen from b is described as species B. Further collections of A in a are assigned to A, while collections of A in b are always assigned to B. If this occurs for all species transgressing the borders of a and b, then the actual number of transregional species is $(n_a + n_b)/2$, where n_a and n_b are the numbers of species in each region and 2 is the number of regions. For the case of r regions this expression can be generalised to $\frac{\sum_{i=1}^r n_i}{r}$. It is probable that only those species that have their

nearest neighbours in adjacent regions are the transgressive species. It is less likely that species either with their nearest neighbour in

non-adjacent regions or in the same region are synonymous with their nearest neighbour (n.n.). In the first case the distances, environmental differences and number of real migrational barriers encountered by a species transgressing the whole geographical range of the genus are likely to mean that the species is unable to maintain its biological entity. In the second case the assumption that further specimens of a previously described species are likely to be assigned to the pre-existing species would reduce the likelihood of within region synonymy as compared to between region synonymy. Thus for the multi-region case the actual number of species is related to (no. of species with n.n. in non-adjacent regions) + $\frac{1}{2}$ (no. of species with n.n. in adjacent regions) + (no. of species with n.n. in the same region). For the data on nearest neighbour location for the Old World species of Saurauia, the contribution of nearest neighbours in adjacent regions can be taken as an error term in the assignment of nearest neighbours. This results from the following. Because the values of nearest neighbour location are computed on a limited species set and because of the possibility that real nearest neighbours for some species were not included in the analysis it is likely that some nearest neighbour pairs are not real nearest neighbours. This is especially true when the n.n. are in non-adjacent regions. This problem is also inherent in the choice of nearest neighbour values as compared to higher order neighbour values, especially where there is little numerical difference between them. The strong correlation between 1st and 2nd nearest neighbour values has already been considered in chapter 9. In response to this the use of total linkage of regions, based on all similarity values will be considered later, in an attempt to overcome the problems of the arbitrary choice of only nearest neighbour values.

To use the model to compute the number of species the following terms are defined.

Ns = no. of species with n.n. in the same region.

Nad = no. of species with n.n. in an adjacent region.

Nna = no. of species with n.n. in a non-adjacent region.

The error proportion over all terms is $(Nna/(Ns+Nad+Nna))$. The proportion of species which will be ultimately resolved into synonymy is p.

The expression for computing the number of species given any degree of within or between region synonymy is therefore:

$$[(Ns(1-p)+1)+(Nad(1-p)+1) + (Nna(1-p)+1)] \div \left(\frac{Nna}{Ns+Nad+Nna} \times 100 \right) \%$$

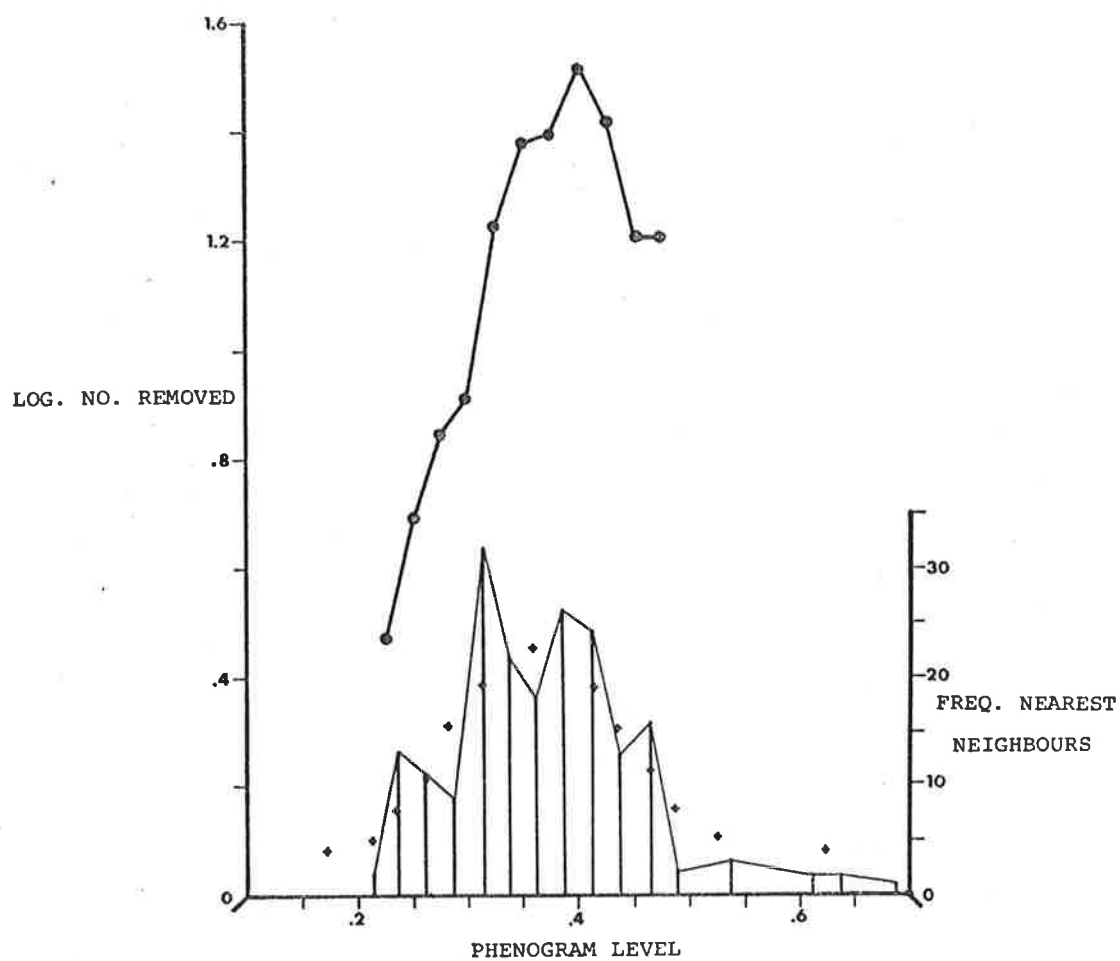
Each term is increased by 1, to prevent the possibility of zero terms in the case of 100% synonymy. In a case of 100% synonymy, there will in reality be a single species.

For the model applied to Saurauia, Ns = 79, Nad = 97, and Nna = 25. Individual values for within and between region nearest neighbours are illustrated in figure 78. Within the limits (0 - 50%) between region synonymy and (0 - 30%) within region synonymy, the computed values for the number of species are given in Table 11. Note that the sum of nearest neighbour species is 201 because some species have more than one nearest neighbour. The table has been standardised to 192 species to make it easier to interpret.

		Between region synonymy (%)					
		0	10	20	30	40	50
Within region synonymy (%)	0	192 [±] 24	182 [±] 23	174 [±] 22	164 [±] 20	155 [±] 19	146 [±] 18
	10	184 [±] 23	175 [±] 22	166 [±] 21	157 [±] 19	147 [±] 18	139 [±] 17
	20	177 [±] 22	167 [±] 21	159 [±] 20	149 [±] 18	139 [±] 17	131 [±] 16
	30	169 [±] 21	160 [±] 20	151 [±] 19	141 [±] 17	132 [±] 16	123 [±] 15

TABLE 11. Number of species in the genus modelled from the degree of synonymy.

FIGURE 78. Plot of the number of species removed by successively lower phenon lines in the phenogram of all species (fig.69) The histogram is plotted for all nearest neighbour values. These are the values used to construct the phenogram. Crosses indicate a fitted normal curve to accentuate the peaks of the histogram.



It is possible to check this range of values against a phenogram derived for all species (fig. 69). If one chooses progressively lower phenon lines, the rate of change of the number of species removed from the phenogram can be used as an estimator of the most reasonable sample based phenon, below which species can be considered distinct. The number of species distinguishable, the number removed at each step and the phenogram levels are summarised in Table 12.

Phenogram level	No. of species remaining	No. removed in previous step	Log. no. removed
0	192	0	-
.100	192	0	-
.150	192	0	-
.200	191	1	0
.225	188	3	.477
.250	183	5	.699
.275	176	7	.845
.300	168	8	.903
.325	151	17	1.230
.350	127	24	1.380
.375	102	25	1.398
.400	69	33	1.519
.425	43	26	1.415
.450	27	16	1.204
.475	11	16	1.204
.500	5	6	.778
.550	4	1	0
.600	3	1	0
.700	1	2	.301

TABLE 12. Number of species distinguishable on a phenogram as a function of the level of a phenon line. Values determined from figure 69.

If log. of the number of species removed at a level is plotted against the level, the region of maximum positive slope of the curve gives the maximum value of $d(dn)$ where dn is the rate of change of the

number of species (fig. 78). The taxonomic significance of this value is that it marks the optimal level at which synonymous species may be distinguished from non-synonymous, nearest neighbour species. Where there is synonymy in a group of species, the distribution of nearest neighbour similarity values tends to be bimodal, or at least skew towards greater similarity. That is, there are more closely similar species than one would expect in a normal population. This shift in frequencies is expressed in the curve of figure 78 as a convex section over the interval .225 to .30. If a histogram of nearest neighbour similarity values plotted on figure 78 is inspected, the reason for this convexity is obvious. A peak containing a high proportion of probably synonymous species lies in the interval .15 to .275. A second peak containing species which are members of complexes of closely related species lies in the interval .275 to .35. From .35 to .70 the species are essentially distinct from each other and not members of close-knit complexes.

Inspection of the log. plot reveals that the steepest section lies near the .3 level, which also marks the cut-off between the first two peaks in the histogram. The number of species distinguishable at .30 is 168. This value is the same as that estimated in chapter 8 from a subjective assessment of species relationships. This value also corresponds to the outlined portion of Table 11. If .325 is chosen as the cutoff, the number of species becomes 151. If one chooses successively lower cutoff levels and plots the geographical distribution of the combined species as they resolve into synonymy at each level, it could be expected that the resulting map would yield evolutionarily significant biogeographical information. That is, as the ranges of species become overlapping, one might hope to find local centres of related species being produced. This, however, does not

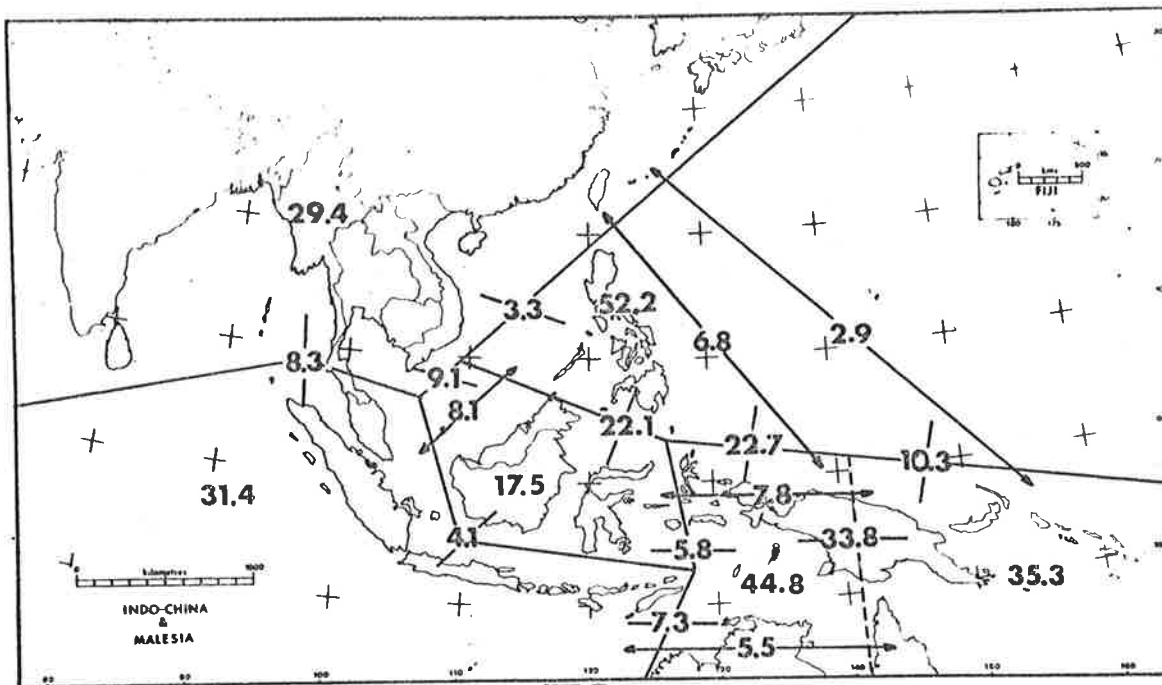
occur. The choice of .3 as the species cutoff is sufficiently arbitrary and the degree of uncertainty in any one nearest neighbour pair is such that the effect is masked by randomly generated overlaps, which obscure any clustering present. Choosing even .325 as the cutoff, drastically changes the distributions so formed. It is, therefore, necessary in studying the overall relationships between species within and between regions, to find a measure that is insensitive to individual value uncertainty and the effect of arbitrary significance levels.

Two approaches to this problem involve either the computation of within and between region nearest neighbour scalars, or scalars which give a measure of overall similarity of species within and between regions, regardless of nearest neighbour considerations.

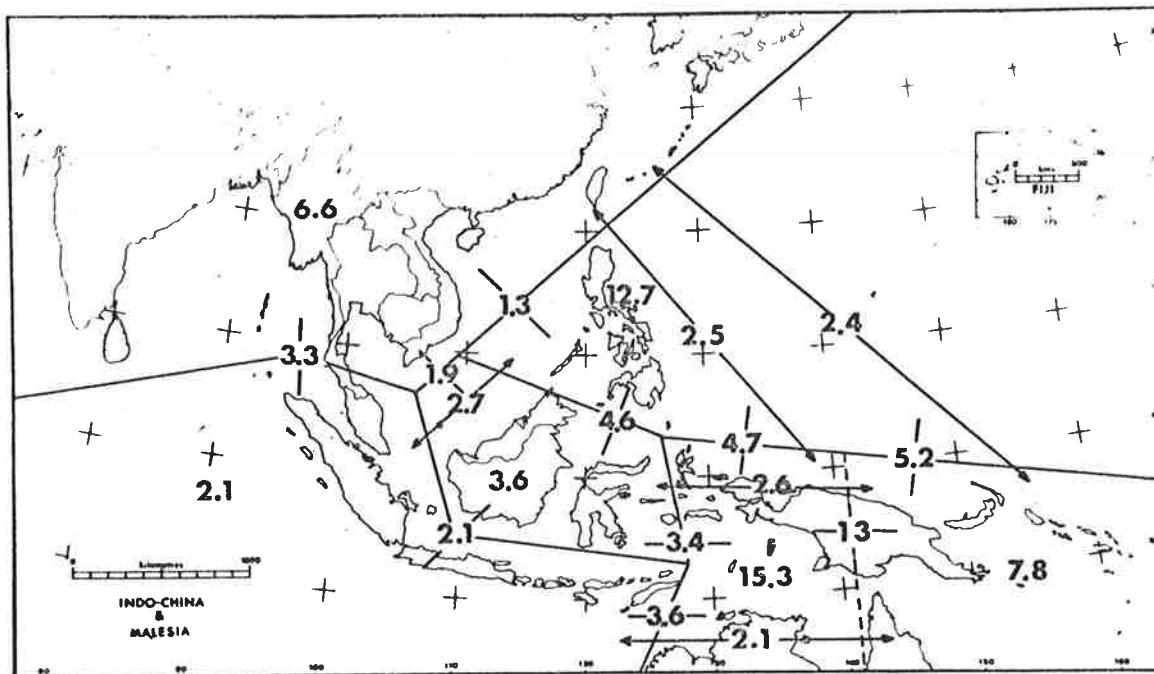
A nearest neighbour scalar is a measure of the overall similarity of region defined species groups, computed from the number of nearest neighbour values within regions and between regions. The number and location of nearest neighbours are illustrated in figure 79. The actual number of species in each region ranges from 13 to 46. It is therefore necessary in comparing within and between region similarities to standardise the number of nearest neighbour values to the number of possible n.n. values. The number of possible n.n. values within a region equals the number of species. The number of possible n.n. values between two regions equals the sum of the number of species in the two regions. For example, the scalar of within region similarity for region 6 equals $19/51$, where 19 is the number of nearest neighbours in the region and 51 is the number of species in the region (figs. 76 and 77). The scalar for similarity between regions 5 and 6 equals $24/(23+6+45)$, where the divisor is the sum of all species in the two regions. The values of within and between region nearest neighbours are given in figure 79.

FIGURE 79. Nearest neighbour scalars within and between regions for Saurauia. For explanation see text. Scalar values are multiplied by 100 to assist interpretation.

FIGURE 80. Scalars for within region and between region similarities, computed over all similarities less than 0.5. Scalars multiplied by 100. Further details in text.



79



80

Based on n.n. scalars the relationships of geographically defined species groups are easily seen. Region 4 contains a group of species which form a morphologically homogeneous complex. The species responsible for this high degree of nearest neighbour similarity form the first distinct complex that segregates in the phenogram for all species (fig.69). This same species group also forms the well defined chain in the upper left of the nearest neighbour tree (fig.70). The species of region 5 form a group with somewhat less internal similarity. This is an effect of the high degree of nearest neighbour linking to the species of region 6. Together these two regions form a group of species that is strongly linked internally. In both the phenogram (fig.69) and the nearest neighbour tree (fig.70) they form clearly distinguishable species groups. The significant linking of the species of regions 5 and 6 to those of region 4 is indicated in figure 79. Region 4 is further strongly linked to region 3. It is interesting to note that the species of region 3 form a unique group, in that based on nearest neighbour values they appear to have greater similarity to the species of region 4 than they do to themselves. This suggests that there were insignificant barriers to the migration of species between Borneo or the Philippines. Regions 1 and 2 both have high degrees of internal similarity. Region 2 is also significantly linked to Indo-China, the Philippines and Irian Jaya. The species of the Indo-China region (region 1) lastly form a group that is about as internally linked as those of Java and Sumatra. The most significant transgressive nearest neighbour linkings of the species of region 1 are with the adjacent regions 2 and 3. There is also a surprising degree of similarity with some species of Irian Jaya. This anomaly will be discussed when the biogeography of the generic groups is considered later.

The pattern of nearest neighbour linking is readily distinguished not only in figure 79, but in the phenogram (fig. 69) and the nearest neighbour tree (fig. 70). This is not surprising in that these last two plots were constructed from nearest neighbour values. The degree of distortion inherent in choosing nearest neighbour values to represent overall species relationships is discussed in chapter 9. Some measure of this distortion may be assessed if one compares the values for nearest neighbour linking illustrated in figure 79 with the values for overall linking shown in figure 80.

The degree to which species form interrelated complexes can be judged from the skewness of a frequency distribution of similarity values for all pairwise comparisons of species. A distribution skewed towards greater dissimilarity (1-similarity) reflects a situation in which there is clear speciation. A symmetrical distribution, or one that is skew towards high similarity values indicates that the species are indistinct from one another. One expects to find species complexes in this latter situation.

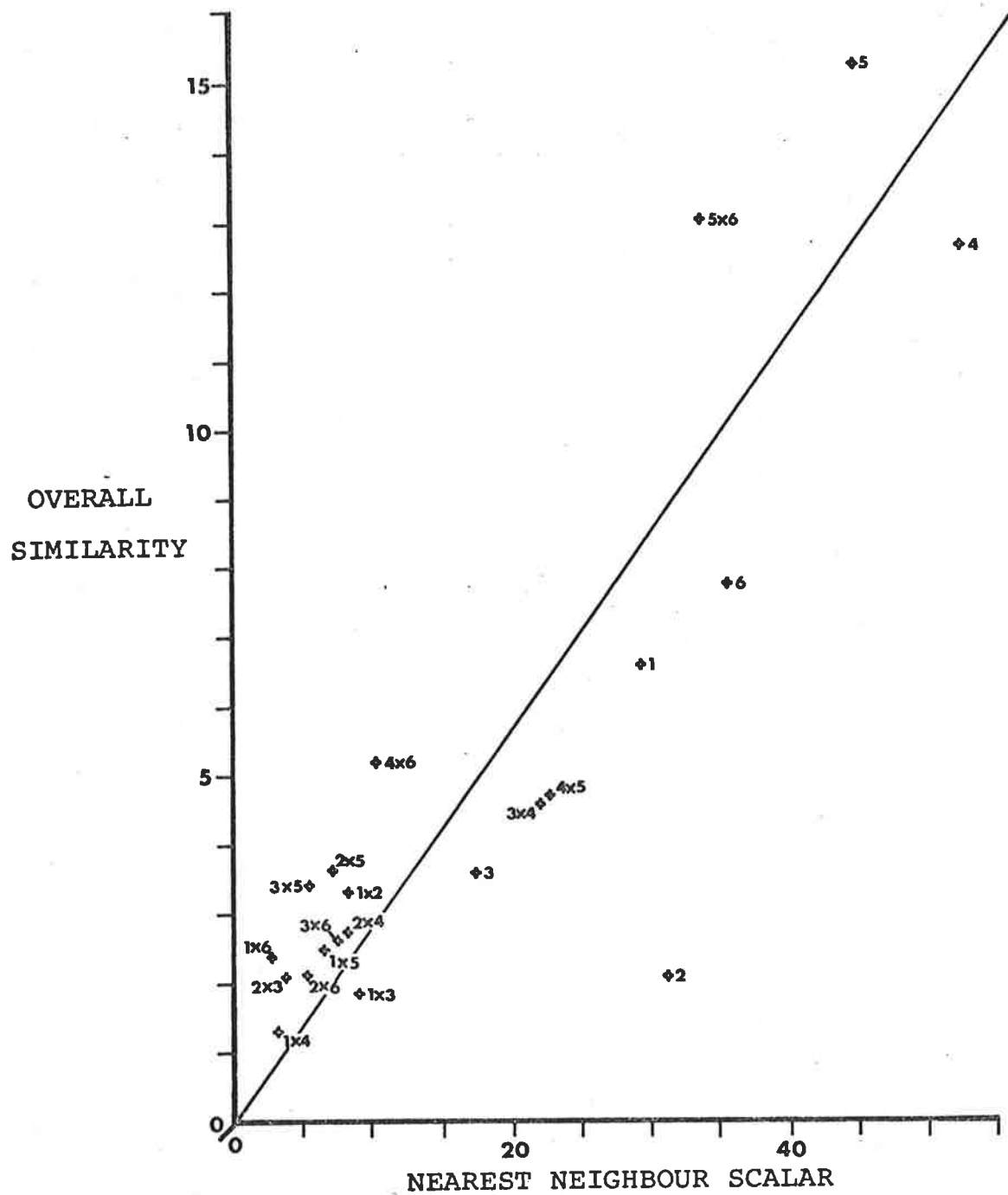
The number of pairwise comparisons between species increases in proportion to the square of the number of species. A frequency plot of similarity values for 192 species contains 18,336 values. Even when computerised, handling such a number of comparisons becomes prohibitive. Based on a subjective assessment of the morphological range of the genus and an estimate of skewness of the distribution of similarity values an arbitrary cutoff level for comparison was chosen. The possible values of similarity range from 0 to 1 where the greater degree of similarity is indicated by lower values.

Overall similarities for figure 80 were computed using only those values less than or equal to .5. This produced only 1021 comparisons. Because overall similarities were calculated, the number of possible comparisons is related to the product of the number of species, rather

than the sum as in the case of nearest neighbour linkage. For within region comparisons, the overall similarity for each region^{is} calculated from $(\text{no. of values l.t. or eq. to .5}) / \frac{1}{2}(\text{no. of species})^2$. The matrix of within region comparisons is symmetrical around the self match diagonal. Therefore the divisor is multiplied by $\frac{1}{2}$. For between region comparisons there is no self match diagonal. The overall between region dissimilarity for each pair of regions is therefore calculated from $(\text{no. of values l.t. or eq. to .5}) / (\text{product of no. of species in each region})$.

The overall similarity values illustrated in figure 80 yield essentially the same story as the nearest neighbour values of figure 79. Once again the magnitude of values can be interpreted as a measure of the skewness of the distribution of pairwise similarity values. Large values indicate that a greater proportion of similarities are significant. The ordination of regions in figure 80 differs very little from that of figure 79. This is an indication of the close relationship between nearest neighbour and overall similarity in a large sample of species. This observation has a strong theoretical basis which is related to the fact that a set of species in a genus is not a random morphological assemblage, but one which is internally related through evolution. The relative ordinations produced by these two approaches to measuring overall regional similarity are summarised in figure 81. This is a plot of nearest neighbour scalars against overall similarity scalars for all within and between region comparisons. The correlation coefficient computed for these values is 0.8514 for 19 degrees of freedom. This value is significant at very much less than 0.1%. The dispersion of individual points around the fitted regression line is a measure of the relative numbers of nearest neighbour and overall similarity values less than .5. Thus region 5 which lies above the line has comparatively more values less than .5 than region 4, which

FIGURE 81. Nearest neighbour and overall similarity scalars plotted for pairs of geographical regions. For explanation see text.



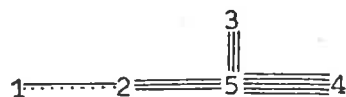
lies below the line. This indicates that the species of region 5 have overall greater similarity between themselves than do the species of region 4. The only region that diverges markedly from the general trend is region 2. This group of species has a high ratio of nearest neighbours to the number of values less than .5. This indicates that although the species of region 2 are most similar to themselves, they are essentially distinct species. Whereas in regions close to the fitted line there is a high degree of within region similarity, this is not the case for region 2. This suggests that there is little chance of synonymy amongst the species of this region. In fact relative to species in other regions, those of Java, Sumatra and peninsular Malaysia, are morphologically heterogeneous. There are three possible explanations for this. Firstly, the species of region 2, especially those of Malaya, have been the subject of recent taxonomic revision (Hoogland, 1977). Because collections studied by modern taxonomists are generally more complete and more readily available than ever before, it could be expected that the problems of synonymy have a greater chance of being resolved. The effect of this is to reduce the degree of similarity between species by removing intermediate forms previously recognised as separate species. A second possibility for the heterogeneity of region 2 species, is that not all of the species have yet been collected. Most of the species described from Java and Sumatra have been collected from four general localities; Padang and Gunong Kerintji in central Sumatra, Gunong Besagi in south-eastern Sumatra and the region of Gunong Gedeh, Buitenzorg and Bogor in Java. It is possible that with collections from such restricted localities many other species have been missed. Unfortunately with the extensive clearing of forests on these islands and the high population densities there may never be improvement in these collections. The final

explanation for the heterogeneity of the region 2 species is that they may in fact be biologically distinct. This is possible when one considers that the degree of overall similarity of the species of region 2 is less for a within region comparison than it is for between regions. The species of region 2 are most similar to species from regions 1 and 5, somewhat less to 4 and equally similar to species from region 3 as they are to themselves (fig. 80). If the degree of similarity to species of regions 1 and 5 reflects evolutionary relationships it is possible that the species of region 2 are heterogeneous because of their diverse provenances.

The consideration of regional relationships at the species level leads into a discussion of the evolutionary history of the group as a whole. On the assumption that overall similarities based on a set of reproductive and vegetative characters may reflect evolutionary relationships it is possible to trace the most likely migrational pathways of the genus through the Old World. For the sake of this discussion regions 5 and 6 will be considered as a single geographic region (region 5).

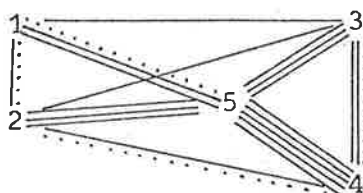
Migration routes

A model of migrational pathways should maximise the similarity between consecutive, regional species groups along the pathway. Similarities between groups are derived from overall similarities given in figure 80. Using only maximum between region similarities the following path has been constructed:-



Each solid linking line equals 2 units of overall similarity (fig. 80). Dotted lines equal one unit. The number of lines linking groups is a measure of their overall similarity. There are geo-

graphical reasons for suspecting that this model does not adequately illustrate the interrelationships of the groups, e.g. regions 1 and 3 are adjacent. Therefore a second model using all between region similarities has been constructed.



It is clear from both models that region 5 has the greatest similarity to all other regions. This does not imply that the genus originated in region 5, but it does suggest that it has played an important part in the evolution of the extant species of Saurauia. Because it is likely that the migration of the genus has not been as simple as dispersal into new regions by the introduction of single species, there are a large number of possible pathways, depending on the starting point and direction of migrations. Fortunately, useful information can be gained from distribution data for the generic sections and from the history of movements of the Australian and Asian continental blocks. Further discussion will now consider regions 5 and 6 separately.

Biogeography of the generic sections

Near the beginning of this chapter it was mentioned that there are two approaches to the analysing of the interrelationships between species of different regions. The first of these, the consideration of species similarities has already been discussed. The second approach, the analysis of the geographical distribution of the sections in the genus, will now be considered.

It is generally accepted that the genus Saurauia can be divided into a number of sections on the basis of reproductive and vegetative

morphology. Although the ten sections proposed by Diels (1922) have been considered artificial and difficult to apply (Burt, 1936), evidence is now offered for the retention, at least in part, of Diels' scheme. The sections are defined and discussed in detail in chapter 6.

If one plots the geographical distributions of the generic sections an interesting fact is revealed. Figure 82 illustrates these distributions. It is clear that there are a number of sections whose geographical range is truncated by the boundary between region 4 and region 5. The actual number of species which are members of those sections not traversing this boundary is 72, or 37% of the Old World species. One of these sections, the Ramiflorae, is numerically the most important section of regions 1 to 4. Of the four sections found in regions 5 and 6 that are endemic, three have specialised, calyptrate inflorescences. Of the species of regions 1 to 4 that have large floral bracts, none have specialised to the extent of having more or less sessile flowers in a head that these endemic New Guinea sections possess.

Clearly the boundary between the New Guinea regions and the regions to the west marks a major disjunction in the genus. If one refers back to the discussion of endemism and synonymy in the genus, it is clear that the lack of species in common between New Guinea and the rest of Malesia is due in part to the endemism of generic sections. It is also clear that synonymy between these two areas is to be found only in those sections that transgress the boundary between New Guinea and regions to the west. The sections that are found throughout the Old World range of Saurauia can also be seen in figure 82. Difficulties encountered in trying to assign a species to a section are most often encountered when the species belongs to one of the four widespread sections, Uniflorae, Rufae, Setosae and Squamulosae.

FIGURE 82. Frequencies of Saurauia species grouped by generic section (columns) and geographical regions (rows). The thick line separating regions 4 and 5 in this and the three figures following indicates a major distributional disjunction within the genus.

FIGURE 83. Frequencies of 3-styled species grouped by generic section and geography as in figure 82.

FIGURE 84. 4-styled species arranged as in figure 83.

FIGURE 85. 5-styled species arranged as in figures 83 and 84.

	RAMIFLORAE	UNIFLORAE	OBTECTAE	RUFÆ	SETOSAE	SQUAMULOSAE	ARMATÆ	BIBRACTEATÆ	OBVALLATÆ	CALYPTRATÆ
					PLEIANTHÆ					
6		5		3	17	10	1	4	3	
5		5		2	8	8	5	1	6	4
4	13	9	3	1	10	5				
3	10	3		3	9	4				
2	14	2		2	6	2				
1	5		4	1	4					

82

3 STYLES

					2		1		
				1	2		1		
7	9		1	9	1				
8	1			9					
10	3		1	1					
2									

83

4 STYLES

					2				
3	4	1		3	1				
1				1					
1	1			2					
2									

84

5 STYLES

	7		3	14	7	2	3	5	
	3			7	4	4		3	4
5		1		2	1				
2	1			2	1				
1	1			4	1				
5		3							

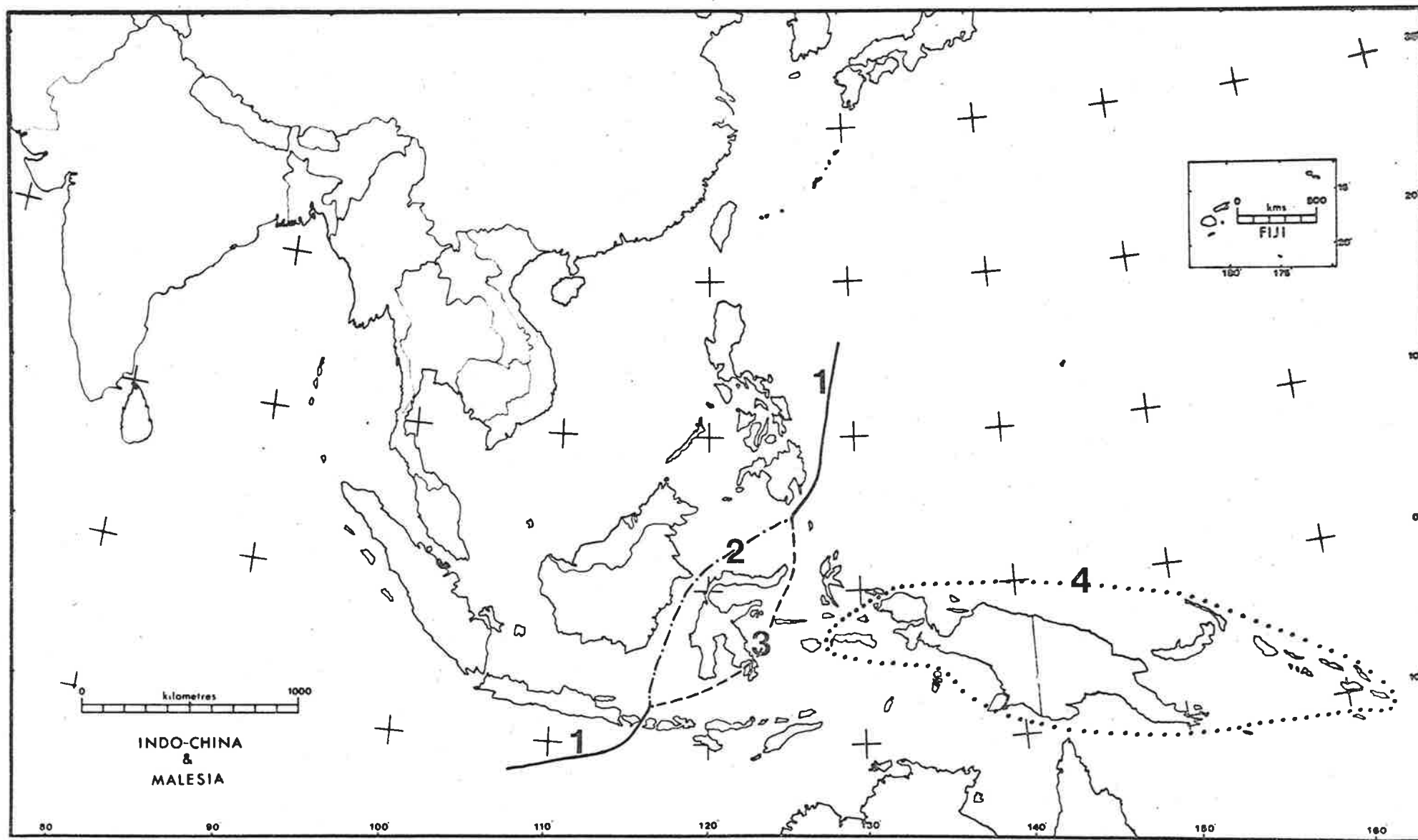
85

To try and find a reason for this disjunction in distributions, it was decided to investigate other characters to see if they also exhibited disjunctions. The character of number of styles was found to reflect the same phenomenon. The number of styles found in the genus ranges from 2 to 7, with 3, 4 or 5 being most common. It has been noted that style number is not necessarily constant within a species (Hoogland, 1977). However, it appears that although species with 3 styles may occasionally possess 2 or 4, and species with 5 may exhibit 6 or 7, it is relatively uncommon to find a species with 3 to 5 styles. Figures 83-85 illustrate the disjunct distribution of the different style number classes. Five styled species are common throughout the range of the genus (fig. 85). Four styles are known from regions 5 and 6 in only two species, both of which are predominantly five styled (fig. 84). Only a small proportion (9.6%) of species with three styles are found in regions 5 and 6 (fig. 83).

Since the sections of the genus are defined on floral and vegetative characters, and the disjunct distributions of the sections are so obvious, the reason for the disjunction must be fundamentally important in the evolution of the genus. If rather than defining the disjunction by arbitrary regional boundaries, one plots the actual geographical distributions of the sections, an even more interesting phenomenon appears. Figure 86 illustrates the limits of the distributions of the sections of the genus, plus the style number classes. It is obvious that the disjunction of species groups within the genus Saurauia reflects a disjunction that is apparent in many genera and families of both plants and animals. The phenomenon of Wallace's line is well documented in a wide range of terrestrial organisms (Keast, 1972; Schuster, 1972). The reasons are now well understood in terms of the drift through geological time of the Australian continental

FIGURE 86. Limits of generic sections and style number classes within the Old World Saurauia's. The limits are:-

1. Wallace's Line.
2. Eastern limit of Ramiflorae, section 2, (ramiflorous and cauliflorous spp.) and Obtectae, 4D (leaves with a close appressed tomentum).
3. Eastern limit of species with 4 styles. Note that this limit differs from the previous one only because of one species out of 25 which occurs on Celebes.
4. Limit of Obvallatae, 4G and Bibracteatae, 4F (multiple flowered inflorescence enclosed in leafy bracts), Armatae, 4C (leaves and stems beset with prickles) and Calyptratae, 3 (single flowers enclosed in bud by bracts).



block relative to Asia. Wallace's line, first observed as a biogeographic phenomenon, marks the active boundary between these two plates. Botanically, the most important thing about Wallace's line is that it separates the flora of what was Gondwanaland from the flora of Laurasia. Since these two floras have evolved in isolation probably since the middle Cretaceous the botanical discontinuities across Wallace's line are generally expressed at the family level (van Steenis and van Balgooy, 1966; Schuster, 1972). A priori the reasons for major floristic differences across Wallace's line would hardly be expected to be manifested at the generic level. This is especially true for Saurauia, when the total geographic range of the genus is considered.

A large number of species of the genus are known from Mexico, and South America. If Saurauia originated in Gondwanaland, it must have migrated into North America via South America some time after the mid-Cretaceous and is now dispersing through Asia after introduction from the Australian block. Unfortunately this does not explain the presence of fossils of Actinidiaceae in Europe and the noticeable lack of extant Australian species. The presence of the family in Europe either implies that it is a very ancient Gondwanalandic family or that it is in fact Laurasian in origin. Only one species of Saurauia is known in Australia. Saurauia andreana F.v.M is found in coastal northern Queensland. Approximately 70 species are found in New Guinea, just across the Tasman Sea. If Saurauia evolved in Gondwanaland it is difficult to explain why only a single species was able to survive in Australia, while becoming extremely successful in New Guinea. There is little doubt that Saurauia originated in Laurasia and has recently dispersed southwards onto the Australian plate as it drifted north and into contact with Asia. The early

history of the genus in Asia and America can be inferred only from present distributions. The inadequacy of the fossil record has already been mentioned.

The South American species of Saurauia belong to the section Pleianthae. The Old World species of this section are divided into seven sub-sections. The New World species are separated into four other groups. It is possible however to treat them as if they belonged to the Old World species and allocate them to one or other of Diels' sections. The only groups of Diels' scheme which are known from the New World are the Setosae and Squamulosae. Of 21 species revised by Hunter (1966), 9 belong to the former section and 12 to the latter. Of these species, 17 have only 5 styles, 2 have 4 to 5, 1 possesses 3 to 5 and one is strictly 4 styled. The Setosae and Squamulosae both have axillary, paniculate inflorescences with small unspecialised bracts and simple, sparsely to moderately dense, obtuse scales or long setae on the flowers and leaves. By comparison with the specialisation of inflorescences in the Ramiflorae, Bibracteatae, Obvallatae and Calyptratae and the indument of the Obtectae, Rufae and Armatae, the Setosae and Squamulosae appear relatively primitive. The probably evolutionary relationships between the sections of Saurauia are illustrated in figure 87.

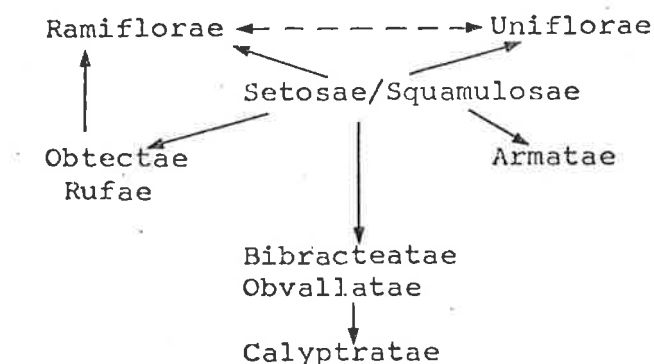


FIGURE 87. Phylogenetic relationships
between generic sections of
Saurauia.

This diagram is based on a subjective assessment of the most likely pathways of morphological specialisation, given the Setosae/Squamulosae as the starting point. The origin of the Ramiflorae is probably the most uncertain. Some species of Ramiflorae have solitary flowers on the leafless branches suggesting an affinity with the Uniflorae. It is possible however that this is a secondary reduction. The dense foliar indument of many species of Ramiflorae suggests a more likely relationship with the Obtectae or Rufae. Lastly, the fact that many species have a setose or squamate indument infers possible evolution directly from the Setosae. It seems possible that the Ramiflorae have evolved from species of more than one section. This evolutionary scheme is further supported by the disposition of species of the different sections relative to each other on the nearest neighbour tree (fig.70). Thus one finds that the Ramiflorae are most closely associated with Setosae, Obtectae and Rufae. Details of other relationships between sections shown in figure 87, are also supported by the same diagram.

The history of evolution and dispersal of Saurauia can be derived from the relationships between geography, generic section and style number. These relationships are demonstrated in figures 82-85. Unfortunately, due to inadequate type descriptions, the Setosae of region 1 have not been scored for style number. Based on probable evolutionary relationships between the generic sections and their present distribution it is proposed that the five styled Setosae/Squamulosae are the most primitive group of species in the genus. It is possible that when the genus crossed the Bering Strait land bridge after the Late Cretaceous these were the only group that had evolved. There may have been a number of three styled species, but they were probably in the minority. The dispersal and diversification

of the genus in Asia and Malesia was dependent on the relative positions of the Australian and Asian blocks at several critical periods in the history of the genus. This is suggested by figures 79 and 80, which indicate the close relationship between the species of New Guinea and other regions. It is also supported by figure 82 which demonstrates that the ancestral group of species is now the largest group and is to be found mainly in New Guinea.

The history of the continental drift of Australia relative to Asia is well documented (Smith, 1972; McGowran, 1973; Smith and Briden, 19). Following the breakup of Australia and Antarctica in the Eocene, the northward drift of the Australian-New Guinea block first brought it into close conjunction with the Indonesian island chain in the mid-Miocene (Smith and Briden, 19). Since then, continued drift has brought New Guinea much closer to the Philippines where it now lies. Australia therefore has had a long and probably discontinuous contact with Asia. This situation is important in understanding the history of Saurauia. During, and previous to Miocene time, the Setosae/Squamulosae were probably the only group existing in Asia. This inference is possible because of both their present biogeography and their relatively unspecialised nature. It is proposed that during the Miocene close approach of Australia to the Java/Sumatra island chain 5-styled members of these two sections were able to cross into New Guinea. The presence of almost exclusively 5-styled species in New Guinea at the present time supports this. Since the introduction of Saurauia to these new regions, the genus has undergone a marked degree of diversification. The four sections endemic to New Guinea are evidence of this. At some time later, it seems likely that the Setosae/Squamulose once again transgressed a regional boundary and were able to reach the Philippines or

Borneo. The great degree of similarity between species of these sections in the New Guinea and the Philippines has already been noted and is mostly responsible for the high degree of between region similarity indicated by figures 79 and 80. It is possible that by the time a second transgression was able to occur, the Uniflorae and Rufae had already evolved, as both these groups also transgress Wallace's line. From the Philippines the genus appears to have made a major ingress into Borneo (figs. 79 and 80). The greater degree of similarity of region 2 species to the Philippines, than to adjacent Borneo, suggest that this circuitous migration may have been the case, rather than a more direct route from region 2 into region 3. Thus the region 2 species are related to those of regions 4 and 3, via their connection through New Guinea.

The evolution of the Ramiflorae remains in doubt. This section is taxonomically heterogeneous and appears to be related to several other sections on morphological grounds (fig. 87). The fact that it strongly supports a Wallace's line disjunction suggests that the Ramiflorae evolved in post-Miocene time, after the introduction of Saurauia to New Guinea. The section is numerically important throughout regions 2, 3 and 4. The presence of species in region 2 which are most closely linked to regions 5 and 6, implies an independent introduction of these species from elsewhere than Borneo or the Philippines.

The species of region 1 exhibit some unexpected relationships. Figures 79 and 80 indicate a significant degree of both nearest neighbour and overall similarity between regions 1 and 5/6. A large amount of this effect may be due to background noise in the methods of measuring similarity. To improve the accuracy of the evolutionary model, it is possible to once again use an overall similarity approach but at the level of single species rather than generic sections. The

overall similarity for each species to all others has been calculated. To reduce computational loads only those values of dissimilarity less than or equal to .5 have been considered. For each species an overall similarity scalar has been computed. This scalar has been divided into two components. These are (1) overall linkage of species to those of regions 1 to 4, (2) overall linkage of species to those of regions 5 and 6. The two components of the scalar have then been compared to determine whether the provenance of species affects their overall linkage. The hypothesis is that species will be most closely linked to the group of species on their side of Wallace's line. This has been tested on both type description data for 192 species and on leaf data for 44 species. For the leaf data, the cutoff for dissimilarity was set to .6 to improve the reliability of estimates of overall similarity from the smaller sample. The two scalars were then plotted for species grouped by generic sections. Figures 88 to 94 illustrate the results of this procedure. Shading of groups is to delimit style classes. Numbers on the plots refer to the geographic region from which the species were recorded. The straight line passing through the origin on each plot delimits the regions in which the species are expected to fall. This line has been computed from the numbers of species in each of the two regional groups for each set of data so that the area on each side of the line is proportional to the relative number of species in each group.

Given that the distribution of similarity values is truncated by the cutoff of .5 or .6, the value of the scalar for a species is related to the number of comparisons above that cutoff. The scalars are computed as the sum of $(1-I_D)$ for all matches where I_D is less than or equal to .5 or .6. Therefore species that lie close to the origin on the plots are essentially distinct at the cutoff level since

FIGURE 88. Species of the Obvallatae, Bibracteatae, Calyptratae and Armatae scored for type description data and plotted on the two components of overall similarity scalars. Shaded area encloses 3-styled species. For further explanation see text.

FIGURE 89. Species of the Ramiflorae plotted as for figure 88. Shading encloses 5-styled species.

FIGURE 90. Species of the Squamulosae plotted as for the previous two figures. Shading encloses 3-styled species.

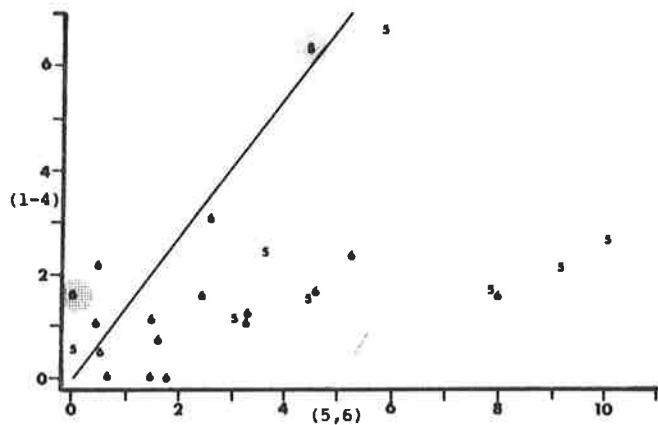
FIGURE 91. Species of the Obtectae and Rufae plotted as for the preceding figures. Shading encloses 3-styled species.

FIGURE 92. Species of the Setosae plotted as the preceding figures. Shading encloses 5-styled species.

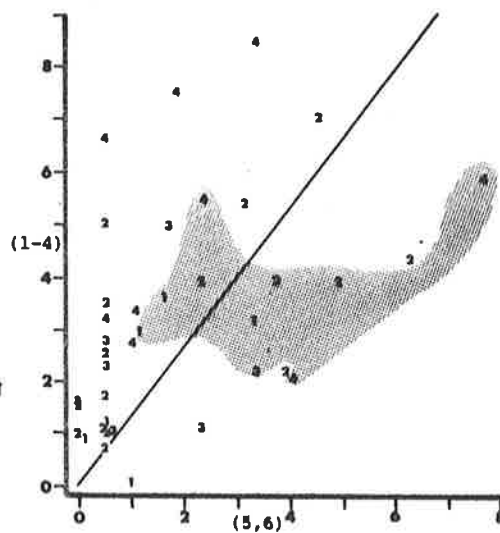
FIGURE 93. Species of the Uniflorae plotted as for the preceding figures. Shading encloses 5-styled species.

In all figures the region 1 to 4 component is plotted on the vertical axis and the region 5 and 6 component is plotted on the horizontal axis.

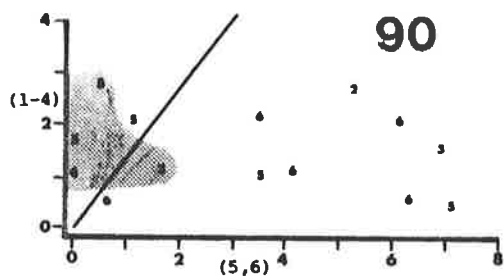
88



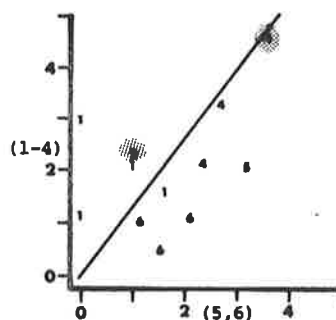
89



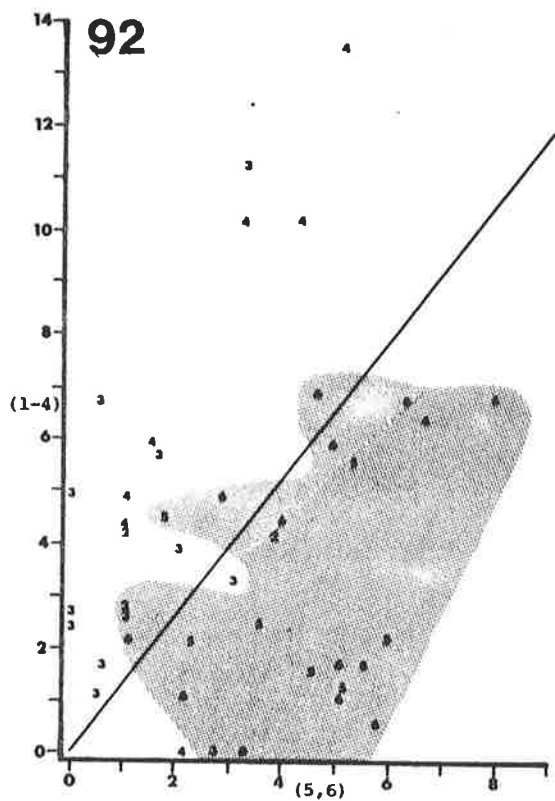
90



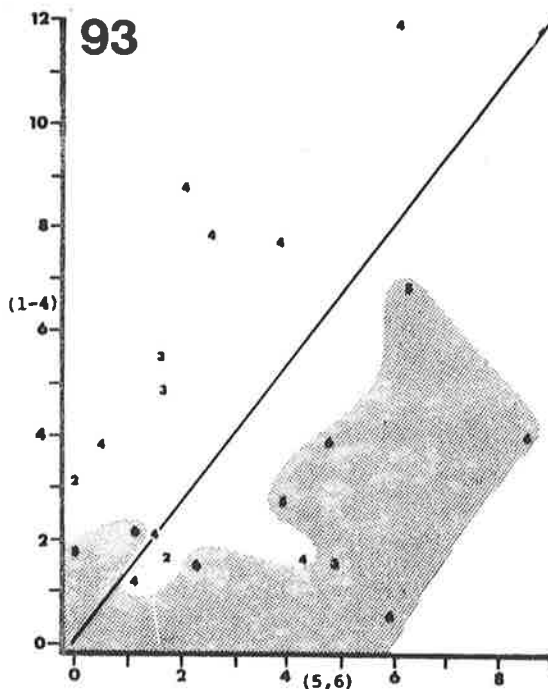
91



92



93



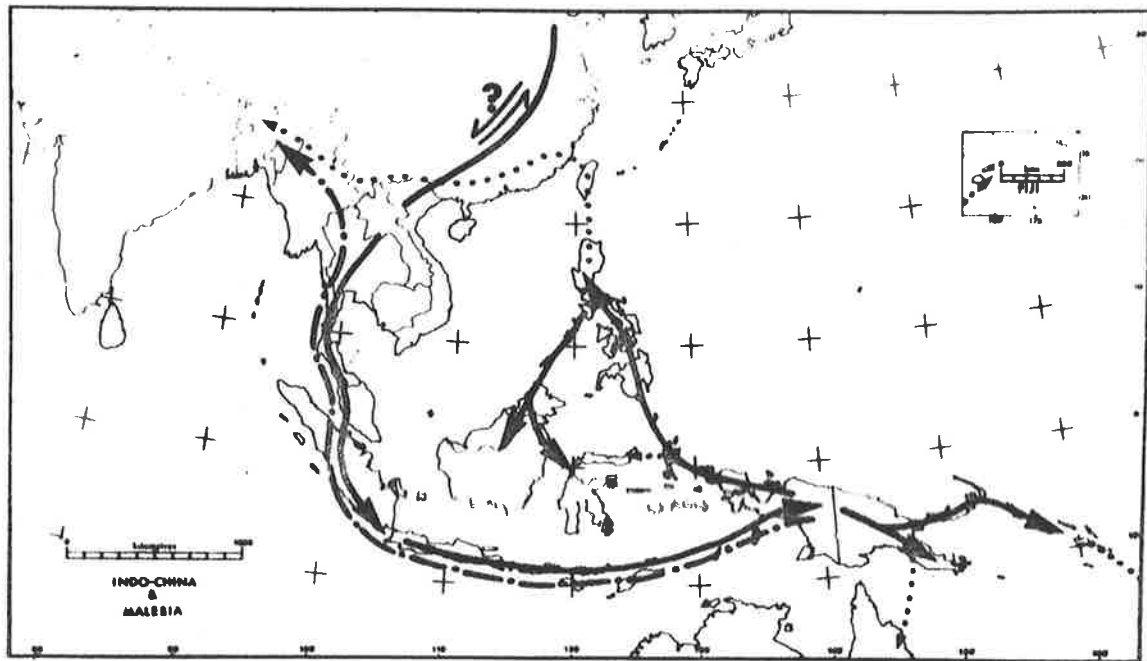
they have few significant comparisons. Species distant from the origin have more numerous comparisons with other species. The reliability of scalar values, in general, decreases towards the origin since the actual number of values contributing tends to decrease. The variance of similarity values is quite low, so that confidence limits for each species have little effect on the ordination of species. This is especially true when the sum of the two scalar components is larger than 3.

If a species falls into the region of the plot that corresponds to its actual geographic location, it is considered to be correctly assigned to a geographical group. For species scored on type description data, 76% are correctly assigned. Species whose similarities are computed from leaf data are correctly assigned in 79% of cases.

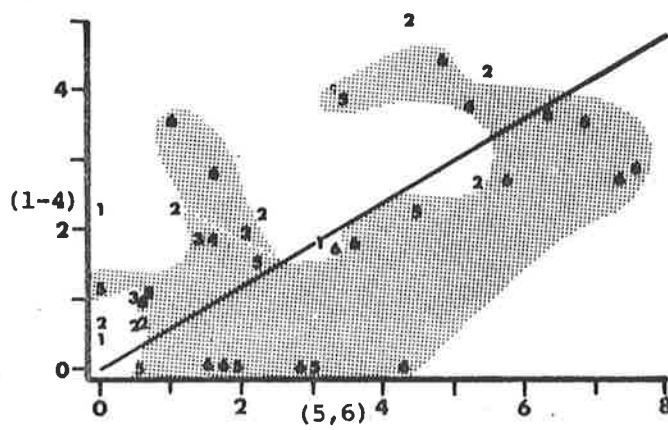
The overall similarity of a species relative to one or other group, equals the ratio of the scalars for the two region groups. This can be measured from the graphs by computing the slope of a line passing through the species point and the origin. The regional dividing line has been computed from the relative numbers of species in each regional group. Therefore in assessing the degree of correct assignment to a group, it is important to take into account not only whether a species is placed on the correct side of the line, but how far it is from the dividing line. On this basis it can be seen that the Ramiflorae have the greatest degree of incorrect assignment. The distribution of many species far into the New Guinea side of the plot suggests that the Ramiflorae of regions 1 to 4 may have arisen from more than one introduction of ancestral species. The species of Ramiflorae which are currently found in regions 1 and 2 are most strongly linked to regions 5 and 6. There are two region 4 species which lie well into the region 5 and 6 side of figure 94, but for

FIGURE 94. Species of all generic sections scored for leaf data and plotted on the two regional components of the overall similarity scalar for I_D values less than or equal to .6. Shading encloses 5-styled species.

FIGURE 95. Proposed evolutionary pathways for Saurauia in the Old World. Line conventions are explained in the text. The origin of the genus itself is uncertain and is indicated by the bidirectional arrow at the top of the figure.



95



94

the most part, species of region 4 form a distinct cluster above the line. The strong linkage of region 1 and 2 Ramiflorae to New Guinea species rather than to those of regions 3 and 4, suggests that there have been two independent evolutionary pathways leading to the development of the Ramiflorae. The low overall similarities between regions 2 and 3 can now be explained in terms of two relatively recent migrations out of New Guinea. One of these was into Borneo via the Philippines, while the other was into Java, Sumatra and Indo-China. The low similarity over all species between regions 2 and 3 also seems to be a result of insignificant migrations across the Java Sea.

Conclusions

The existence of a major taxonomic disjunction which reflects Wallace's line has been observed within the genus Saurauia. This effect has been demonstrated on a sample of about 190 species. The question of endemism and synonymy has been investigated in an attempt to ascertain the reliability of the status of species within this sample. The actual number of species is shown to lie most probably in the range 150 - 180. This makes the interpretation of biogeographic disjunctions as a real biological phenomenon on the basis of 192 species, only slightly less reliable. The reasons for the disjunction appear to be explainable in terms of the most primitive sections and their probable migrational history. The evolutionary model proposed for the Old World species of Saurauia is shown to be consistent with a continental drift determinant.

In conclusion, the migrational pathways of Saurauia through Indo-China and Malesia can be summarised as follows. 5-styled Setosae/Squamulosae were probably in existence in South-East Asia and the Java-Sumatra island chain in the Miocene. As the Australian

block came into close contact with the Indonesian chain, species from this group were able to migrate into New Guinea. In New Guinea the group underwent an important diversification, giving rise to four endemic sections. The Uniflorae and Rufae probably also arose at this time in New Guinea. At some later period as the Australian block moved further north, the genus was once again able to effect a major migration. At this time, the sections Setosae, Squamulosae, Rufae and Uniflorae were introduced into the Philippines. The four generic sections endemic to New Guinea may have arisen after this second migration, but it is not possible to determine this from the currently available information.

The timing of the origin of the Ramiflorae and Obtectae is obscure. It is possible that the Obtectae which hardly differ from the Rufae originated in the Philippines and reached South-East Asia directly from there. The homogeneity of this group in South-East Asia supports the idea that it may have arisen from the introduction of a single species across the Luzon Strait. The Ramiflorae form two groups. One of these is centred in regions 3 and 4, and the other is in regions 1 and 2. The species of this section in Borneo and the Philippines appear to have evolved following the introduction of the genus from New Guinea. The Ramiflorae of Java, Sumatra and the Asian continent probably arose following a reintroduction of species from New Guinea. Figure 95 illustrates the proposed migrational pathways of the genus throughout the history of its evolution in the Old World, leading to the specialisations within generic sections that can now be observed. On the figure, thick black lines indicate pathways that are major in terms of the number of species that have arisen from the migration. Dotted lines are minor pathways for only one or two species e.g. S.andreana is the single species found in Australia and

S.rubicunda is the single species endemic to Fiji. The broken line leading from New Guinea via Java/Sumatra to Indo-China traces the reintroduction of species leading to the evolution of the Asian Ramiflorae.

CHAPTER 11

DISCUSSION

This thesis describes a leaf architectural study of Saurauia (Actinidiaceae). The primary aim has been to use Saurauia as a test case for evaluating a new set of numerical leaf architectural characters. The methodology and philosophy of the research program is antithetic to classical taxonomy by intent. This approach has been maintained to demonstrate that alternate classifications of dicots leading to an autonomous parataxonomy can exist. Evaluation for a large plant group of a numerically based leaf architectural system by purely objective means has not been attempted before. Existing leaf architectural classifications are dependent on a terminology such as that of Hickey (1973). Therefore any testing of the value of leaf architectural classifications of dicotyledonous leaves has used Hickey's system. Critical appraisal of such classifications has shown that they are inadequate (Dolph, 1976 b). Reasons for their inadequacy are firstly that as they are non-numerical, the characters violate the requirements of many statistical procedures. For example, it is technically invalid to apply measures of taxonomic distance to two-state characters or to non-ordered multistate characters. Secondly, statistical correlations between characters are poorly understood, such that the redundant component of the character set is unknown. Thirdly, being non-numerical, there will be errors of assignment of OTUs to character states, simply because intermediate forms cannot be adequately accommodated by a discrete terminology. By using a strictly numerical character set, these weaknesses can be overcome, or at least their influence on the classifications produced, can be understood.

Leaf architectural methods are of particular value in palaeobotany. Therefore the present study has been geared to the requirements of that discipline. A discussion of the information available to a general palaeobotanical study is in order.

Megafossil floras are dominated by leaves. In any assemblage of leaf fossils there will be some degree of missing data. From experience I can recognise the minimal information content one should expect before commencing a study of a fossil flora which has no existing comparative taxonomic framework. This is particularly pertinent to Australia, where Tertiary floras stand in taxonomic isolation. When one seeks to classify fossils only within the context of a single flora, it is possible to define most taxa on low order characters, that is physically coarse characters such as shape, primary and secondary venation and margin type. Once a flora is to be compared with other floras, the requirement of higher resolution becomes important. Thus for an autonomous, monofloral parataxonomy, a subset of leaf architectural features will probably suffice. For a multifloral parataxonomy, especially for floras spatially and temporally widely separated, even the whole set of architectural features may be insufficient. A floristic study which compares Tertiary dicots with extant plants, should not be generally considered unless cuticular characters can be included with the architectural features. For example, the floristic comparison of fossil leaves from the Tertiary fossil flora of Maslin Bay, South Australia, has been attempted by me, on only five leaf taxa with unambiguous architectural features and with good cuticular preservation. For general palaeobotanical work, especially the primary description of a new flora, a set of architectural features of no higher order than tertiary venation should suffice. This restriction may be imposed by using only primary characters in initial analyses

of plant assemblages. Primary characters are those that are common to all OTUs in the assemblage. Where OTUs are not discriminated by the initial character set, they can be further studied by whatever characters are available in each subset of OTUs.

The limitation in this project to a set of primary leaf architectural characters, also limits the number of characters available. It is generally held by numerical taxonomists that a large character set, where characters are distributed over a range of organs of the OTUs, is desirable to achieve stability in a numerical classification (Sokal and Sneath, 1973). The criterion of widely sampled character types is mostly irrelevant in this study, which is based on a single organ of the plants. The requirement of a large character set is still worthy of discussion. The unquestioning acceptance of the philosophy that a large character set is a good one, cannot go unchallenged. It can be easily shown that a small set of metric characters with low correlation and each contributing significantly to overall species variance, can be as good as a large, highly correlated set containing many binary or non-ordered characters and in which characters are often relatively invariant. The latter character set contains all of the undesirable features discussed in chapter 2. These include non-equality of class frequencies, non-additivity of distances between classes, non-unidirectionality of characters, character correlation and differing numbers of class intervals. The influence of these properties is that characters and states are weighted in a prior analysis. This practice violates the philosophy of strict objectivity in numerical taxonomy.

If the character set of Hickey (1973) is used without critical assessment or deletion, the resultant classification will be an artifact of the interrelated characters and weighting of characters and states. Such a classification could be expected to reveal little of

unambiguous biological value. This has been empirically tested and confirmed by Dolph (1976 b). Before such a character set is allowed to become the backbone of a leaf architectural classification of dicotyledonous leaves, it must be subjected to careful empirical analysis. Any attempt to use it without such analysis is premature. This applies to the initiation of a leaf data bank using Hickey's system (Dolph, 1978).

For Saurauia a leaf character set as large as possible was initially chosen. This is important to ensure that greater freedom is possible in selecting the best characters. Of the original set of 25 metric architectural characters, only 10 were used in the final classification. These ten characters are the most efficient discriminators of Saurauia species. As I expect most of the ten to be applicable to general leaf architectural classification, they are listed and discussed below.

Leaf width (maximum). This character contributes significantly to overall variance of the character set. It also has a much more significant F-ratio than leaf length, to which it is highly correlated. In a palaeobotanical study, maximum leaf width is also more likely to be measurable than length.

Length/width ratio. In general palaeobotanical studies this character may often be impossible to measure on single leaves, though sampling should overcome this. In the case of incomplete leaves it may be possible to substitute leaf length with a value such as the distance between the axes of 20% maximum width.

Base angle 2. This character has been chosen rather than apex angle because the leaf base is more likely to be preserved than the apex, and this latter structure has been found to be much more variable within species. Base angle 2 is defined as the

angle subtended by the intersection of the basal axis of 20% maximum leaf width with the margin and the point on the midrib defined as the leaf base.

Position of maximum width. This character is important in expressing leaf shape. Its dependence on leaf length means that it may be difficult to measure on some fossil. Again a measure of length based on the axes of 20% maximum width may be chosen.

Percent acuminate. As this character expresses the percentage of total leaf length encompassed by an acuminate apex, its use in palaeobotanical studies would be limited. It has proven to be of use in Saurauia.

Number of secondary veins. This character was the most efficient discriminator for Saurauia species. It is possible to define an interval over which veins are counted in cases where leaf terminations are missing. This may also be desirable where the secondary veins near leaf terminations are difficult to differentiate from other vein orders. The distinction between vein orders has been discussed in chapter 3 and should be consulted before any characters of venation are applied.

Secondary vein intercostal shape. Although this character is related by geometrical constraints to both leaf width and number of secondary veins, it contains enough unique information to be generally useful. It is particularly applicable to fossils as only a small number of vein intercostals need be measured.

Secondary vein angle. This character is easily measured on both living and fossil leaves and may be scored for only those veins around the region of maximum width if the leaf base or apex are lacking.

Tertiary vein angle. This character was measured in Saurauia for only one pair of veins per leaf. The vein angle measured at 50% maximum width and 50% leaf length was not significantly different from the mean of veins measured at 18 locations on the leaf. If it is possible to generalise this for other taxa, it will be particularly useful for measuring fossils.

Dentition frequency. Although entire margined leaves will receive a zero value for this character, it serves to distinguish the states entire/non-entire. Unlike a terminology it also does not differentiate these two types as being fundamentally different. For example, it is easily observed that some leaves of a species with sparse dentition may also be entire. The very satisfactory F-ratio of this character made it the second best discriminator of Saurauia species.

The final character set used in this study is oriented towards the requirements of Saurauia. For other leaf architectural studies the whole set of characters described in chapter 3 should be consulted. I recognise many areas where the present character set is deficient for general purposes. For example there are no characters relating to lobed or compound leaves, except that the leaf shape index can accomodate these. There are no characters of higher vein orders. This should be possible to rectify. However, caution is advised when differentiating higher vein orders as the distinctions are often very subjective. Characters which I expect to be especially useful in leaf architectural studies are the leaf shape index, secondary vein intercostal shape and secondary vein straightness. Selection of numerical and especially geometrically oriented characters is recommended for leaf architectural studies. Such characters may be tedious to measure (a small penalty for their efficiency) and so practical aids for

scoring them have been developed. These include leaf rubbings, cleared leaves, cellulose acetate peels and measuring grids for leaf shape and hair density.

The first aim of the present study, to define a numerical leaf architectural character set of no higher order than tertiary veins has therefore been achieved. The second aim, to generate a leaf based classification and compare it with one based on data from type descriptions will now be considered.

The intent of the second aim was to demonstrate that leaf characters were able to define taxonomic groups within Saurauia. To achieve this a statistical analysis which would accomodate the data set of numerical characters, was needed. A similarity analysis was chosen as most suitable, though other analyses, in particular principal components analysis, were applied. The value of using similarity is that it is a simple concept to understand, especially with regard to interpreting results. A number of measures of similarity are available and these are discussed in chapter 2. However, no existing procedure applicable to the data was able to handle multiple character state scores. This is an important requirement where no prior statement can be made about the relative variability of the species being tested. It is better to assume that some species are more variable than others than that all are equally variable. If this assumption is made, the resultant ordination is more likely to reflect actual morphological relationships between species, even if they prove to be unequally variable. If the assumption that all species are equally variable is made however, the resultant ordination may incorrectly classify variable taxa. Because of this a similarity measure which is able to accomodate measures of dispersion rather than single values such as means will be more generally satisfactory. As

no such method was available, I have developed a similarity/distance metric which is a generalisation of the class of expressions known as the Minkowski metrics. A discussion of the derivation and relationships of similarity measures is given in chapter 2. The method developed here is able to accomodate multiple character state scores. Measures of dispersion such as ranges, studentised ranges and standard deviations will allow the use of multiple values. Rather than scoring OTUs for single mean values therefore, all values encompassed by the measure of dispersion are included. This is especially important in palaeobotany where OTUs may be single leaves. The use of values such as a range of secondary vein angles, will maximise retrievable information and not encounter the problem of sample size affecting measures of dispersion.

The multistate similarity/distance metric can be seen as two related forms. As a strict dissimilarity measure it is given the symbol I_D (Index of dissociation). This form computes similarity from positive associations between character states of OTUs. That is, OTUs are considered similar on a character only when they have at least one character state in common. For continuous metric characters this occurs when their ranges overlap. Similarity values range from 0 to 1. In its present form the metric computes dissimilarity. Similarity is the complement of dissimilarity, i.e. $1-I_D$. OTUs identical across all characters are given a dissimilarity of 0. OTUs which have no states in common are given a dissimilarity of 1. The limitation of I_D in a multivariate sense is obvious. Any degree of mismatch is treated identically. To overcome this a Euclidean distance metric D_I has been derived. It is able to compute the distance between OTUs, averaged over all characters. The value of D_I ranges from 0 to a value which is equal to one less than the average range of characters.

It is important to standardise characters or to use a fixed number of class intervals to avoid character weighting. If this latter method is used, the range of D_I is from 0 to (1-no. of character states).

Similarity and distance values were used to generate phenograms and nearest neighbour trees for 44 species using both leaf data and type description data. Although 93 species were scored for leaf characters, only 44 were scored for three or more specimens and also on type data. This subset of species was chosen firstly to make the classification as reliable as possible and to allow comparison between leaf and type data alternatives. Although the data included in type descriptions is that traditionally used for diagnosing plants, the leaf based classifications were shown to be more efficient at separating species. This is discussed in chapter 8. The reason for better separation of species on leaf data is probably related to the fact that flowers were often lacking from type specimens and terminology (especially of induments) was often ambiguous. Despite this, ordinations on leaf and type data show some consistency. For example, species of the generic section *Setosae* were usually grouped together. These species include *Saurauia bifida*, *S. dielsiana*, *S. holotricha*, *S. tafana*, *S. rubrisquamata* and *S. schumanniana*. The species *S. poolei*, *S. purgans*, and *S. occulta* were also generally clustered close to the *Setosae* group. *S. bracteosa*, *S. malayana*, *S. roxburghii* and *S. vulcani* were also consistently clustered. This ordination was maintained even when further species were added to the analyses and is particularly well illustrated in the nearest neighbour tree for 193 species scored for type data (fig. 70, p.241). The classification schemes of *Saurauia* species on both leaf and type data demonstrate the existence of species complexes within the genus. These include the species mentioned above and other complexes which appear in the final classification of figure 70. The taxonomic status of the members of these

complexes will be resolved only by a revision of the species involved. However, some degree of synonymy is suggested by the fact that a complex of species from the Philippines including S.bakeri, S.elmeri, S.clementis and S.gracilipes were all described by Merrill. In this complex the species are probably artificial and the result of overly fine splitting on the part of the taxonomist.

The graphical method chosen to illustrate the classifications of Saurauia was a nearest neighbour tree (figs. 65 and 70). Higher order neighbour trees were plotted, but it was found that the graphical complexity of these masked the overall ordination. To make the simple nearest neighbour trees more useful in representing the dimensions of the similarity matrix, species were defined by the geographical regions from which they were collected. With these regions imposed on the graphs, a consistent ordination of species became obvious. This ordination was easiest to see in plots on leaf data (fig. 65). Clustering of species into geographically defined groups is very apparent in the nearest neighbour tree for all species on type data (fig. 70). In general regionally defined groups are consistent with the morphologically defined groups.

Geographical groups of species were defined in terms of six geographical regions. Regional boundaries were chosen on the basis of two criteria. Firstly they should maximise regional endemism assuming a priori that all species were distinct and herbarium labels were correct. Secondly they should mark intuitively determined migrational barriers. Such barriers for Saurauia include seaways and lowland areas which were assumed to retard dispersal. This latter assumption was made because lowlands are known to be effective barriers to the migration of South American species (Hunter, 1966). Most Old World species are found at altitudes above 1000 metres. Ordination

of species into geographical regions by similarity was tested for 193 species scored for type description data. Similarities were used in two ways to measure the degree of linkage of species within regions and between regions. Firstly only nearest neighbour values were used. These values were used to compute scalars whose magnitude was a measure of nearest neighbour similarity of regional groups of species. As nearest neighbour values do not represent a simple proportion of the distribution of similarity values, a second approach was taken. This involved using all $(1-I_D)$ values less than or equal to 0.5 to compute overall linkage within and between regions. Only half of the range of I_D was used because the whole similarity matrix is very large (18000 entries). Overall linkage scalars were computed. Again the magnitude of these is related both to contributing species number and the average value of similarity. By this means it has been possible to show how geographical regions are related in terms of the morphological similarities of the species they contain. This supports the idea that New Guinea and particularly Irian Jaya contain species which are highly similar to those of surrounding regions. By joining regions which are most strongly linked it has been possible to trace a migrational pathway for the Old World species. During this phase of the analysis it also became obvious that there were some strong biogeographic disjunctions within the genus. This can be illustrated by referring to figure 82-5 which show the geographic distribution of generic sections and groups of species defined by style number. The disjunction can be best seen if one inspects the distribution of the Ramiflorae and of 3-styled species. The boundary between region 5, Irian Jaya and islands to the west, Philippines, Borneo, Celebes, Java and Sumatra marks a strong disjunction in these groups. Likewise a disjunction can be seen in the generic sections Calyptratae, Obval-

latae and Bibracteatae which are endemic to regions 5 and 6 (New Guinea, Solomon Islands and Bismarck Archipelago). If these disjunctions are plotted on a map of the Old World (fig.86) it can be seen that the boundary between morphologically defined areas lies along Wallace's line. Wallace's line is recognised as the boundary between the Asian and Australian continental blocks and therefore the front on which floras and faunas of the two regions are mixing. The two adjacent biotas had been separated by a seaway until the initial contact of the two continental blocks in the mid Miocene. Therefore there is a great degree of endemism at the family level in the two regions. The Australian flora is recognised as being basically derived from Gondwanic stock, while that of Asia is Laurasian in origin. If the disjunction in Saurauia was taken as a normal Wallace's line effect the present distribution of the genus would be difficult to explain without giving it great antiquity. Therefore an isolation mechanism involving initial contact between continental blocks, a later separation by a seaway and then a more recent second contact is proposed. By this means it is possible to explain the evolution of endemic generic sections in New Guinea after the initial contact and then the distribution of the primitive sections Setosae and Squamulosae into the Philippines and Borneo at a later date. This migrational pathway is illustrated in figure 95.

The discovery of a morphological disjunction related to Wallace's line at the generic level was unexpected. There is no question that there is a direct relationship between the present distributions of the generic sections and this line. This raises some questions regarding disjunctions within other plant groups. Where families are known to presently transgress the line, there is a possibility of finding such disjunctions. Many families including Santalaceae,

Loranthaceae, Stackhousiaceae, Sapotaceae, Dilleniaceae, Celastraceae and Pittosporaceae are now distributed throughout Australia and Malesia (van Steenis and van Balgooy, 1966). The Celastraceae and especially Celastrus has a distribution almost identical to that of Saurauia. If Wallace's line disjunctions are found in this and other families it should be possible to validate the migrational pathway of Saurauia. It may also be possible to modify our understanding of the origins of the families now common to Australia and Malesia.

It is unlikely that there was any significant migration between the two regions before the mid-Miocene as this marks the time of the first close approach of Australia to Asia (Smith and Briden, 1977). Any plant groups in Australia before that time can be considered to be Gondwanic. Following on from this, any plant groups appearing for the first time in the Australian or New Guinea fossil record around mid-Miocene time are most likely to be Laurasian. Therefore, palaeobotanical studies of Australian lower Tertiary floras should concentrate on floristic comparisons with Australian plants. With the architectural characters and statistical techniques derived within this study it should be possible to realistically consider the Tertiary megafossil floras of Australia.

BIBLIOGRAPHY

- ATCHLEY, W.R., GASKINS, C.T. and D. ANDERSON, 1976. Statistical properties of rations. I. Empirical results. *Syst. Zool.* 25(2): 137-148.
- BACKER, C.A. and R.C. BAKHUIZEN VAN DEN BRINK, 1963. *Flora of Java*. 3 Vols. 2050 pp. Noordhoff, Groningen.
- BAKER, E.G., 1918. The Sumatran species of Saurauia. *Journ. Bot. Brit. For.* 56:161.
- BECKER, H.F., 1961. Oligocene plants from the Upper Ruby River Basin in southwestern Montana. *Geol. Soc. Amer., Mem.* 82:127 pp.
- BERRY, E.W., 1916. The lower Eocene floras of southeastern North America. *U.S. Geol. Surv. Prof. Pap.* 91.
- BERRY, F.W., 1922. Additions to the flora of the Wilcox Group. *U.S.G.S. Prof. Paper* 131:1-20.
- 1924. The middle and upper Eocene floras of southeastern North America. *U.S.G.S. Prof. Paper* 92:206 pp.
- 1930. Revision of the lower Eocene Wilcox flora of the southeastern states, with descriptions of new species, chiefly from Tennessee and Kentucky. *U.S.G.S. Prof. Paper* 156:1-144.
- 1933. The cuticle of an Eocene Combretum. *Jour. Wash. Acad. Sci.* 23:505-508.
- 1941. Additions to the Wilcox flora from Kentucky and Texas. *U.S.G.S. Prof. Paper* 193E:83-99.
- BLACKBURN, D.T. and D.C. CHRISTOPHEL, 1976. A method of permanently mounting biological tissue cleared in Herr's four-and-a-half clearing fluid. *Stain Tech.* 51(2):125-130.
- BLACKITH, R.E. and R.A. REYMENT, 1971. *Multivariate Morphometrics*. 412 pp. Academic Press, London and New York.
- BURTT, B.L., 1936. Saurauia purgans. *Hook. Icones Plantarum* 34 Ser.5. vol.4:1-3, t3316.
- BUSCALIONI, L., and G. MUSCATELLO, 1912 - 1927. *Studio monographico sulle specie Americane del gen. Saurauia Willd.* *Malphigia* 24: 381-412; 25:1-16, 105-118, 187-250, 389-436; 26:1-32, 97-144, 281-312, 389-420; 27:1-32, 131-158, 293-324, 487-502; 28:1-48, 107-138, 223-238, 315-330, 371-402, 473-448; 29:1-32, 97-112, 230-246, 319-366, 411-458; 30:35-444.
- CAIN, A.J. and G.A. HARRISON, 1958. An analysis of the taxonomist's judgement of affinity. *Proc. Zool. Soc. Lond.*, 131:85-98.
- CHANDLER, M.E.J., 1962. *The Lower Tertiary floras of southern England II : Flora of the pipe-clay series of Dorset (Lower Bagshot)*. Trustees of the British Museum, London.

CHRISTOPHEL, D.C. and D.T. BLACKBURN, 1975. A new procedure for mounting cleared leaves using polyester resin. Trans. R. Soc. S.Aust. 99(2):55-58.

——— 1978. Tertiary megafossil flora of Maslin Bay, South Australia : a preliminary report. In press.

COTTHEM, W.R.J. VAN, 1970. A classification of stomatal types. Bot.J. Linn. Soc., 63:235-246.

——— 1971. Vergleichende morphologische Studien über Stomata und eine neue Klassifikation ihrer Typen. Ber. Dtsch. Bot. Ges. Bd. 84(3/4):141-168.

CROVELLO, T.J., 1976 Numerical approaches to the species problem. Plant Syst. Evol. 125:179-187.

DANCIK, B.P. and B.V. BARNES, 1974. Leaf diversity in yellow birch (Betula alleghaniensis). Can. J. Bot. 52:2407-2414.

DEANE, H., 1900. Observations on the Tertiary flora of Australia with special reference to Ettingshausen's theory of the Tertiary cosmopolitan flora. Proc. Linn. Soc. N.S.W. 25(4):581-590.

DIELS, L., 1922. Die Dilleniaceen von Papuasien. In C.Lauterbach, Beiträge zur Flora von Papuasien VIII. Bot. Jahrb.f.Syst. Pflanzengesch. Pflanzengeog. 57:440-459.

DILCHER, D.L., 1971. A revision of the Eocene flora of southeastern North America. The Paleobotanist 20:7-18.

——— 1974. Approaches to the identification of angiosperm leaf remains. Bot. Rev. 40(1):1-157.

DOLPH, G.E., 1976. Taxometric partitioning of leaf collections. Palaeontographica, Abt. B. 156:65-86.

——— 1976 b. Interrelationships among the gross morphological features of angiosperm leaves. Bull. Torrey Bot. Club 103(1): 29-34.

——— 1978. A proposal for data banking information. Cour. Forsch. Inst. Senckenburg 30:159-164.

DORMAN, F.H., 1964. Australian Tertiary palaeotemperatures. Jour. Geology 74:49-61.

DOUTCH, H.F., 1972. The palaeogeography of northern Australia and New Guinea and its relevance to the Torres Strait area. In D. Walker (ed.) Bridge and barrier: the natural and cultural history of the Torres Strait. Publ. BG/3. Res. School Pacific Stud. Aust. Nat'l. Univ., Canberra.

DYER, W.T.T., 1875. Ternstroemiaceae in Hooker, J.D. Flora of British India, Vol.I, 279 pp.

- EADES, D.C., 1965. The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance. *Systematic Zool.*, 14:98-100.
- ELLIOTT, G.F.S., 1891. The effect of exposure on the relative length and breadth of leaves. *J. Linn. Soc. Botany* 28:375.
- EMILIANI, C., 1954. Temperatures of Pacific bottom waters and polar superficial waters during the Tertiary. *Science* 119:853-855.
- 1960. The temperature decrease of surface sea water in high latitudes and of abyssal-hadal water in open oceanic basins during the past 75 million years. *Deep-Sea Research* 8:144-147.
- ETTINGSHAUSEN, C. VON, 1852. Fossile Proteaceen. *Kgl. Akad. Wiss. Wien Sitzungsber* 9:820 pls. 1 and 2.
- 1861. Die blatt-skelete der Dikotyledonen. 380 pp. 95 pl and 276 figs. Druck and Verlag (Wien).
- 1870. Beiträge zur Kenntniss der Tertiärflora Steiermark. *Kgl. Akad. Wiss. Wien Sitzungsber.* 60:17-100. pls 1-6.
- 1871. Die fossile flora von Sagor in Krain. *Loc. cit.* 63: 406-413.
- 1878. Report on the phyto-paleontological investigations generally and on those relating to the Eocene flora of Great Britain in particular. *Roy. Soc. (London) Proc.* 28:221-227.
- 1883. Beiträge zur Kenntniss der Tertiärflora australiens. *Kgl. Akad. Wiss. Wien Denkschr.* 47:101-148, pls 1-7.
- 1886. Beiträge zur Kenntniss der Tertiärflora australiens. *Kgl. Akad. Wiss. Wien Denkschr.* 53:1-142, pls 8-15.
- 1891. Contributions to the knowledge of the fossil flora of New Zealand. *Trans. Proc. N.Zeal. Inst.* 22.
- FEATHERLY, H.I., 1965. Taxonomic terminology of the higher plants. Hafner, New York. 166 pp.
- FERGUSON, D.K., 1971. The Miocene Flora of Kreuzau, Western Germany. Pt. I : the leaf remains. North-Holland. Publ. Co., Amsterdam, London, 348 pp.
- FINET, A. and F. GAGNEPAIN, 1905. Contribution à l'étude de la flore de l'Asie Orientale. *Mem. Soc. Bot. Fr.* 4:1-170.
- FOSTER, A.S., 1952. Foliar venation in angiosperms from an ontogenetic viewpoint. *Amer. J. Bot.* 39:752-766.
- GARDINER, A.S. and J.N.R. JEFFERS, 1962. Analysis of the collective species Betula alba L. on the basis of leaf measurements. *Silvae Genet.* 11:156-163.
- GILG, F. and E. WERDERMANN, 1925. Actinidiaceae. In A. Engler, *Nat. Pflanzenfam.* ed. 2. 21:36-47.

- GILMARTIN, A.J. and M.J. HARVEY, 1976. Numerical Phenetics in routine taxonomic work. *Systematic Botany* 1:35-45.
- GOOD, R., 1966. The botanical aspects of continental drift. *Sci. Prog., Oxf.* 54:315-324.
- GOODALL, D.W., 1966. A new similarity index based on probability. *Biometrics* 22:882-907.
- GOWER, J.C., 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27:857-874.
- HAMILTON, A.H., 1916. The instability of leaf morphology in its relation to taxonomic Botany. *Proc. Linn. Soc. N.S.W.* 41:152-174.
- HARRISON, F.W. and M.M. ANDERSON, 1974. The application of Herr's 4½ clearing technique to zoological tissues. *Hydrobiologia* 46: 177-180.
- HERR, J.M. Jr., 1971. A new clearing-squash technique for the study of ovule development in angiosperms. *Amer. J. Bot.* 58:785-790.
- 1972. Applications of a new clearing technique for the investigation of vascular plant morphology. *J. Mitchell Sc. Soc.* 88(3):137.
- 1973. The use of Nomarski interference microscopy for the study of structural features in cleared ovules. *Acta Bot. Indica* 1:35-40.
- HICKEY, L.J., 1973. Classification of the architecture of dicotyledonous leaves. *Amer. J. Bot.* 60(1):17-33.
- HICKEY, L.J. and J.A. WOLFE, 1975. The bases of angiosperm phylogeny: Vegetative morphology. *Ann. Missouri Bot. Gard.* 62:538-589.
- HILLS, M.O., 1969. On looking at large correlation matrices. *Biometrika* 56(2):249-253.
- HOLLICK, A., 1936. The Tertiary floras of Alaska. *U.S.G.S. Prof. Paper* 182 p.155, pl.93, fig.4.
- HOOKE, J.D., 1875. *Flora of British India* 1. Reeve, Ashford (Kent).
- HOOGLAND, R.D., 1977. *Saurauiae Gerontogae* 1. Notes on Malayan species. *Gard. Bull. Sing.*, 30:111-122.
- HUNTER, G.E., 1966. Revision of Mexican and Central American Saurauia (Dilleniaceae). *Ann. Missouri Bot. Gard.* 53:47-87.
- JARDINE, N. and R. SIBSON, 1971. *Mathematical Taxonomy*. Wiley, London. 268 pp.
- KEAST, A., 1972. Australian mammals : zoogeography and evolution. In A. Keast, F.C. Erk and B. Glass (eds.) *Evolution, Mammals and Southern Continents*. State University of New York Press, Albany, pp.195-246.

- KIELAN-JAWOROWSKA, Z., 1974. Migrations of the multituberculata and the late Cretaceous connections between Asia and North America. *Ann. S.Afr. Mus.* 64:231-43.
- KOCH, B.E., FRIEDRICH, W.L., CHRISTENSEN, E.F. and E.M. FRIIS, 1973. Den Miocaene brunkuls flora og dens geologiske miljø i Søby-fasterholt området sydøst for herning. *Dansk. ged. Foren., Årsskrift for 1972*: 1-57.
- KOORDERS, S.H., 1923. *Flora von Tjibodas* 7. 246 pp. Buitenzorg, Archipel Drukkenij.
- KRAUSEL, R., 1950. *Die paläobotanischen Untersuchungs methoden.* Gustav Fisher Verlag, Jena 98 pp.
- KRUSKAL, J.B., 1956. On the shortest spanning subtree of a graph and the travelling salesman problem. *Proc. Amer. Math. Soc.*, 7, 48-50.
- KUMMEL, B. and D. RAUP, 1965. *Handbook of paleontological techniques.* W.M. Freeman and Co., San Francisco and London. 852 pp.
- LANCE, G.N. and W.T. WILLIAMS, 1967. Mixed data classificatory programs. Agglomerative systems. *Aust. Computer. J.* 1:15-20.
- LANGERON, M., 1900. Contribution a l'étude de la flore fossile de Sezanne. *Soc. Hist. Not. Autun. Bull.* 13:333-370.
- LEE, A.T., 1948. The genus *Swainsonia*. *Contrib. New South Wales Henb.* 1:131-271.
- LEGENDRE, P. and P. VAILLANCOURT, 1969. A mathematical model for the entities species and genus. *Taxon* 18:245-252.
- LEOPOLD, E.B. and H.D. MACGINITIE, 1972. Development and affinities of Tertiary floras in the Rocky Mountains. In A. Graham (ed.) *Floristics and Palaeofloristics of Asia and eastern North America*: 147-200. Elsevier Publ. Co., Amsterdam.
- LEWIS, M.C., 1972. The physiological significance of variation in leaf structure. *Sci. Prog.*, 60:25-51.
- MACBRIDE, J.F., 1956. *Flora of Peru (Actinidiaceae).* *Field Mus. Publ. Bot.* 13 (Part 3A): 677-686.
- MCGOWRAN, B., 1973. Rifting and drifting of Australia and the migration of mammals. *Science* 180:759-761.
- MÄDLER, K. and A. STRAUS, 1971. Ein System der Blattformen mit spezieller Anwendung für die Bestimmung neogener Blattreste (Miözan und Pliozän). *Bot. Jb.* 90 (4):562-574.
- MAHALONOBIS, P.C., 1936. On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India*, 2:49-55.
- MARTIN, H.A., 1977. The Tertiary palynological record and evolution of the Australian flora. *Abstr. 2nd Aust. Geol. Convention*, Melbourne.

- MARTIN, P.G., 1977. Marsupial biogeography and plate tectonics. In B. Stonehouse and D. Gilmore (eds.). The Biology of Marsupials. Macmillan Press Ltd., London and Basingstoke.
- MAYR, E., 1966. The proper spelling of taxonomy. Systematic Zoology 15:18.
- MELCHIOR, M., 1964. Actinidiaceae. In M. Melchior ed. A. Engler's Syllabus der Pflanzen familien. Twelfth ed. Vol.2. Gebrüder Born traeger, Berlin.
- MELVILLE, R., 1960. A metrical study of leaf shape in hybrids, II. Some hybrid oaks and their bearing on Turing's Theory of Morphogenesis. Kew Bull. (Roy. Bot. Gardens), 14 (2): 161-177.
- 1976. The terminology of leaf architecture. Taxon 25 (5/6): 549-561.
- MILDENHALL, D.C., 1973. A guide to morphological characteristics for use in the description of extant and fossil leaves. Tuatara 20 (2): 75-87.
- MINKOFF, E.C., 1965. The effects on classification of slight alterations in numerical technique. Systematic Zool., 14: 196-213.
- MOUTON, J.A., 1966. Sur la systématique foliaire en paleobotanique. Bull. Soc. Bot. France 113 (9):492-502.
- NIE, N.H., HULL, C.H., JENKINS, J.G., STEINBRENNER, K. and D.H. BENT, 1975. SPSS, Statistical package for the social sciences. McGraw-Hill, New York. 675 pp.
- ODELL, M.F., 1932. The determination of fossil angiosperms by the characteristics of their vegetative organs. Ann. Bot. 46:941-963.
- PEARSON, K., 1926. On the coefficient of racial likeness. Biometrika, 18:105-117.
- PERRY, L.J., HARBISON, R.M. and R.M. LUMB, 1975. The use of Herr four-and-a-half clearing fluid for the rapid microscopic examinations of thick sections of normal and neoplastic tissues. Stain Technol. 50:47-50.
- RAO, C.R., 1948. The utilization of multiple measurements in problems of biological classification. J. Roy. Statist. Soc., Ser.B, 10:159-193.
- REID, E.M. and M.E.J. CHANDLER, 1933. The London Clay Flora. British Museum (Natural History), 560 pp.
- RIDLEY, H.N., 1922. Flora of the Malay Peninsula. Reeve and Co., London.
- ROGERS, D.J. and S.G. APPAN, 1969. Taximetric methods for delimiting biological species. Taxon 18:609-624.

- ROHLF, F.J., 1962. A numerical taxonomic study of the genus Aedes (Diptera: Culcidae) with emphasis on the congruence of larval and adult classifications. Ph.D. thesis, University of Kansas. 98 pp.
- 1970. Adaptive hierarchical clustering schemes. *Systematic Zool.*, 19:58-82.
- 1973. Algorithm 76. Hierarchical clustering using the minimum spanning tree. *Computer J.* 16:93-95.
- ROHLF, F.J. and R.R. SOKAL, 1969. *Statistical Tables*. W.H. Freeman and Co., San Francisco. 253 pp.
- ROHLF, F.J., KISHPAUGH, J., and D. KIRK, 1971. NT-SYS. Numerical Taxonomy System of Multivariate Statistical Programs. Tech. Rep. State University of New York at Stony Brook. New York.
- SCHUSTER, R.M., 1972. Continental movements, "Wallace's line," and Indo Malayan-Australasian dispersal of land plants : some eclectic concepts. *Bot. Rev.* 38 (1):3-85.
- SMALL, E., YUI, P.Y., and L.P. LEFKOVITCH, 1976. A numerical taxonomic analysis of Cannabis with special reference to species delimitation. *Systematic Botany* 1:67-84.
- SMIRNOV, E.S., 1968. On exact methods in systematics. *Systematic Zool.* 17:1-13.
- SMITH, A.C., 1941. Studies of Papuasian Plants IV. *Jour. Arnold Arb. Harvard Univ.* 22 (4):497-528.
- SMITH, A.G. and J.D. BRIDEN, 1977. *Mesozoic and Cenozoic Palaeocontinental Maps*. Cambridge University Press, London. 63 pp.
- SNEATH, P.H.A., 1976. Phenetic taxonomy at the species level and above. *Taxon* 25 (4):437-450.
- SNEATH, H.A. and R.R. SOKAL, 1973. *Numerical Taxonomy*. Freeman, San Francisco. 573 pp.
- SOEJARTO, D.D., 1968. Studies of South American Saurauia (Actinidiaceae). Ph.D. thesis. Harvard University.
- SOKAL, R.R., 1961. Distance as a measure of taxonomic similarity. *Systematic Zool.*, 10:70-79.
- SOKAL, R.R. and C.D. MICHENER, 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* 38:1409-1438.
- STACE, C.A., 1965. Cuticular studies as an aid to plant taxonomy. *Bull. Brit. Mus. Bot.* 4:1-78.
- STEARNS, W.T., 1956. In P. Synge (ed.) *Supplement to the Dictionary of Gardening*. Oxford: 318-322.
- 1973. *Botanical Latin*. 2nd ed. David and Charles Ltd. Newton Abbott.

- STEBBINS, G.L., 1950. Variation and evolution in plants. Columbia University Press, New York, 643 pp.
- STEENIS, C.G.G.J. VAN, and M.M.J. VAN BALGOOY, 1966. Pacific Plant Areas 2. Blumea Supplement 5, pp.312.
- STURGES, H.A., 1926. The choice of a class interval. J. Am. Statistical Assoc. 21:65-66.
- SYSTEMATICS ASSOCIATION COMMITTEE FOR DESCRIPTIVE TERMINOLOGY (SADT). 1962. Terminology of simple symmetrical plane shapes (Chart 1). Taxon 11:145-156; 245-247.
- VOIGHT, E., 1936. Die Lackfilmmethode, ihre bedeutung und Anwendung in der Paläontologie, Sedimentpetrographie, und Bodenkunde. Z.dtsch, geol. Ges. 88:272-292.
- WATSON, L., WILLIAMS, W.T. and G.N. LANCE, 1966. Angiosperm taxonomy: a comparative study of some novel numerical techniques. J. Linn. Soc. Bot., 59:491-501.
- 1967. A mixed data numerical approach to angiosperm taxonomy: the classification of Ericales. Proc. Linn. Soc. Lond., 178: 25-35.
- WEBB, L.J., 1959. A physiognomic classification of Australian rain forests. Jour. Ecology 47:551-570.
- WHITNEY, V.K.M., 1972. Algorithm 422. Minimal spanning tree. Commun. Ass. Comput. Mach., 15:273-274.
- WILLDENOW, H., 1801. Ges. Naturf. Fr. (Berlin) Neue Schr. 3: 407, 1801.
- WILLIS, J.C., 1908. A Manual and Dictionary of the flowering plants and ferns. 3rd ed. Cambridge University Press.
- WOLFE, J.A., 1970. Tertiary climatic fluctuations and methods of analysis of Tertiary floras. Palaeogeog. Palaeoclimatol. Palaeoecol., 9:27-57.
- 1972. An interpretation of Alaskan Tertiary floras. In A. Graham (ed.), Floristics and Palaeofloristics of Asia and eastern North America:201-133. Elsevier Publ. Co., Amsterdam.
- 1973. Fossil forms of Amentiferae. Brittonia 25:334-355.
- WOLFE, J.A. 1959 and D.M. HOPKINS, 1967. Climatic changes recorded by Tertiary land floras in northwestern North America. In K. Hatai (ed.), Tertiary Correlations and Climatic Changes in the Pacific - Symp. Pacific Sci. Congr., 11th, Tokyo, Aug.-Sept., 1966, 25:67-76.